Pós-Graduação em Ciência da Computação

Vicente Vieira Filho

Diretrizes para construção de modelos preditivos de abandono de usuário em jogos móveis

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2017

Vicente Vieira Filho

Diretrizes para construção de modelos preditivos de abandono de usuário em jogos móveis

Este trabalho foi apresentado à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR(A): **Prof. Geber Lisboa Ramalho**
CO-ORIENTADOR(A): **Prof. Paulo Jorge Leitão Adeodato**

RECIFE
2017

Catalogação na fonte
Bibliotecária Elaine Cristina de Freitas CRB4-1790

V658d Vieira Filho, Vicente
Diretrizes para construção de modelos preditivos de abandono de usuário em jogos móveis / Vicente Vieira Filho. – 2017.
204 f.: fig. tab.

Orientador:Geber Lisboa Ramalho.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, Recife, 2017.
Inclui referências e apêndices.

1. Inteligência Artificial. 2. Jogos Móveis . 3. Previsão de Abandono. 4. Mineração de dados I. Ramalho, Geber Lisboa. (Orientador) II. Titulo.

006.3 CDD (22. ed.) UFPE-MEI 2018- 64

Vicente Vieira Filho

Diretrizes para construção de modelos preditivos de abandono de usuário em jogos móveis

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovado em: 23/02/2017.

_____

Orientador: Prof. Dr. Geber Lisboa Ramalho

BANCA EXAMINADORA

_____

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio
Centro de Informática / UFPE

_____

Prof. Dr. Germano Crispim Vasconcelos
Centro de Informática / UFPE

_____

Profa. Dra. Marley Maria Bernardes Rebuzzi Vellasco
Departamento de Engenharia Elétrica/PUC-RJ

_____

Prof. Dr. André Menezes Marques das Neves
Departamento de Design/UFPE

_____

Prof. Dr. Bruno Feijó

Departamento de Informática/ PUC-RJ

Agradecimentos

Aos meus pais, Etiene Lopes e Vicente Vieira, pelo amor e apoio incondicional durante toda essa jornada. Deles nunca me faltaram palavras de conforto ou suporte emocional.

Aos meus avós, por serem minha constante fonte de inspiração pelas suas trajetórias repletas de desafios superados com muita dedicação ao estudo e trabalho.

À minha namorada, Tatiana Soares, por todo apoio, compreensão e carinho especialmente nos momentos mais difíceis e atribulados dessa longa caminhada.

Aos meus amigos do grupo de pesquisa, Tulio Caraciolo, Leonardo Vieira, Ícaro Malta e Átila Malta, por todo o suporte e contribuição na realização deste trabalho.

Aos meus orientadores e amigos, Geber Lisboa e Paulo Adeodato, pela constante motivação, orientação e instrução durante toda a elaboração, construção e discussão do projeto de pesquisa.

Às empresas BigHut Games e Oktagon Games pela cessão das informações dos jogos móveis para permitir a realização dessa pesquisa.

A todos familiares, amigos e colegas que de alguma forma me ajudaram nessa conquista.

Resumo

A previsão de abandono é uma atividade essencial para antecipar a intenção do consumidor de descontinuar um serviço, permitindo ao fornecedor do serviço a aplicação de ações proativas de retenção e fidelização. Como a previsão de abandono é bastante dependente do domínio de aplicação, tem-se usado técnicas de *Domain Driven data Mining* (D3M) e *Behavior Scoring* (BS) com sucesso em indústrias consolidadas como telecomunicações, crédito e varejo, por exemplo. Em indústrias mais recentes, como a de jogos para dispositivos móveis (jogos móveis), a aplicação de mineração de dados para previsão do abandono de jogadores ainda é incipiente. Os trabalhos identificados na revisão de literatura aplicam metodologias genéricas, baseadas na mineração de dados tradicional orientada a dados, e frequentemente toma decisões *ad hoc*. Não há uma discussão profunda sobre as especificidades do domínio de jogos e seus possíveis impactos no modelo preditivo, assim como não há ainda diretrizes claras que possam ajudar na construção de tal modelo. Para avançar no estado da arte em jogos móveis, começamos por tentar responder às seguintes questões: Quais as principais especificidades dessa indústria relevantes para a modelagem do problema? Como essas especificidades podem ser tratadas no processo de construção do modelo preditivo? Qual o peso dessas características, e seus possíveis tratamentos, não desempenho da previsão? A solução proposta é generalizável para outros jogos dentro do mesmo domínio? As respostas para essas questões são fundamentais porque fornecem diretrizes para construção de modelos preditivos para o abandono de jogadores seguindo as abordagens de D3M e BS. Para responder tais questões, além dos estudos na literatura, nós realizamos o planejamento, a execução e a avaliação de um projeto experimental de previsão de abandono em jogos para dispositivos móveis em 3 bases de dados reais com 201.146 jogadores. Os experimentos levaram em consideração possíveis tratamentos para desafios identificados e seus efeitos no desempenho

do modelo preditivo. Nosso estudo empreendeu uma análise crítica da construção de modelos preditivos que ajudarão os desenvolvedores e pesquisadores a criar melhores soluções para o abandono de jogadores em jogos móveis.

Palavras-chave: Jogos Móveis. Previsão de Abandono. Análise Crítica. Mineração de Dados. Behavior Scoring. Domain-Driven Data Mining.

Abstract

The churn prediction is an essential business activity for the anticipation of a consumer intention to discontinue a service, allowing service providers to implement retention and loyalty actions. Since the churn prediction activity is quite dependent on the domain of application, the use of Domain Driven Data Mining (D3M) and Behavior Scoring has been successfully applied in well-stablished industries such as telecommunications, credit and retail. In more recent industries, such as mobile games, the application of data mining for predicting players' churn is still incipient. The studies identified in the literature review usually apply generic methodologies based on traditional data-driven data mining and they usually make *ad hoc* decision. There is no in-depth discussion about the specificities of the domain of mobile games and their possible impacts on the performance of the predictive models as well as there are no clear guidelines to assist in the construction of a churn prediction model. In order to advance in the state of the art on the mobile game domain, we firstly tried to answer the following relevant questions: What are the main industry specificities relevant to modeling the problem? How can these specificities be addressed in the process of constructing the predictive model? What is the impact of these specificities, and their possible treatments, in the churn prediction performance? Is the proposed solutions and treatments generalizable to other games within the same domain? The answers to these questions are critical because they provide guidelines for building predictive models to anticipate players churn following the D3M and BS approaches. To answer such questions, we performed a literature review in churn prediction for mobile games but also in other well-established domains. We planned, executed and evaluated an experimental project for churn prediction of game players based on 3 datasets from real mobile games with 201,406 players. The experiments considered possible treatments for identified challenges and their effects on predictive model performance. Our study performed a critical analysis for building

predictive models that will help developers and researchers to create better solutions to avoid players' churn in mobile games.

Keywords: Mobile Games. Churn Prediction. Critical Analysis. Data Mining. Behavior Scoring. Domain-Driven Data Mining.

Lista de Tabelas

Figura 1-1: As etapas do projeto de pesquisa. 27
Figura 2-1: Alternativas de desenvolvimento comercial em mercados maduros. Esquerda) a figura demonstra os principais eixos relativos ao valor comercial de uma organização: nos eixos x e y, a redução do abandono e aumento do consumo representam aspectos relacionados à gestão do valor do cliente, e no eixo z, o desenvolvimento de estratégias de aquisição de consumidores de alto valor. Direita) a figura apresenta as políticas direcionadas ao consumidor com relação a cada um dos eixos estratégicos. 30
Figura 2-2: Etapas do ciclo de vida do cliente. A jornada do consumidor é representada através de gráfico que relaciona o período de relacionamento do usuário com a empresa (eixo X) com o engajamento apresentado pelo consumidor ao longo do ciclo de vida (eixo Y) em cada uma das etapas descritas sobre esse eixo. 32
Figura 2-3: Representação conceitual do modelo representativo. A construção do modelo representativo do relacionamento da companhia com o cliente objetiva a previsão de uma variável de saída resultado a partir de um conjunto de dados de entrada. 37
Figura 2-4: Ilustração do efeito do desenvolvimento de políticas e procedimentos de excelência no atendimento e satisfação do cliente para o desenvolvimento de barreiras de saída (abandono). 39
Figura 2-5: Particionamento dos dados na metodologia de Behavior Scoring adaptada de (Kennedy et al. 2013/3). 40
Figura 2-6: Exemplificação do particionamento de dados em função do tempo em uma base de dados relacional. 41
Figura 2-7: Representação da modelagem do problema de previsão de abandono. O eixo X representa o aspecto temporal, em dias, tendo o ponto 0 como a data limite, ou ponto de observação. As linhas representam o histórico de comportamento de clientes com a empresa e

os pontos indicam a existência de uma ação relativa ao relacionamento como uma compra, pagamento ou ligação para o SAC. A indicação (+) representa a variável de saída do cliente e a sua indicação como churner. O sinal (-) indica a classificação do cliente como non-churner. 42
Figura 2-8: Tipos de disposição de janelas de performance. A) Janela Superposta. B) Janela Disjunta e C) Janela Única. 43
Figura 2-9: Metodologia CRISP-DM para projetos de mineração de dados.....46
Figura 2-10: Distribuição dos artigos segundo a técnica de mineração adotada (Ngai et al. 2009/3).....51
Figura 2-11: Relação semântica entre área de conhecimento, processo e diretrizes.....55
Figura 3-1: Faturamento da indústria mundial de jogos para os anos de 2015 e 2016 (real) e previsão de faturamento para os anos de 2017, 2018 e 2019 (Newzoo 2016). O faturamento apresentado está subdividido conforme s plataformas de distribuição e consumo dos jogos digitais – de cima para baixo estão PC/MMO, Casual Webgames, TV/Console, Handheld, Tablet e Smartphone.....61
Figura 3-2: Menu de jogos no Nokia 1100 com espaço para somente 2 títulos (Snake e Space Impact).....63
Figura 3-3: Número de aplicativos disponíveis nas principais lojas de aplicativos ao final de 2016 (Costello 2016; AppBrain 2017) 65
Figura 3-4: Modelo de negócios free-to-play supera o pay-to-play em junho de 2011 (Valadares 2011).....67
Figura 3-5: Jogos para dispositivos móveis de sucesso. a) Imagem da esquerda referente ao jogo Candy Crush, um dos jogos casuais para dispositivos móveis mais populares da história com mais de 1 bilhão de downloads e com faturamento diário acima de 3 milhões de dólares. b) Imagem da direita referente ao jogo Game of War, um dos jogos hardcore para dispositivos móveis mais populares com faturamento diário acima de 2 milhões de dólares.....69
Figura 3-6: Receita da indústria de jogos para dispositivos móveis por modelo de negócios nos 4 principais países europeus (Reino Unido, Alemanha, França e Itália). O eixo X apresenta a linha do tempo com a receita realizada até 2016 e projetada até 2020. O eixo Y apresenta a receita em bilhões de dólares americanos. Fonte: (CyberAgent 2016). 70
Figura 3-7: Funil ARM.....71
Figura 3-8: Ciclo de vida do usuário em jogos free-to-play para dispositivos móveis. O eixo X indica o tempo de relacionamento enquanto o eixo Y representa o engajamento do jogador. 71

Figura 3-9: Demonstração da aplicação do modelo de negócios em jogos reais. Superior-Esquerda) Tela de jogo com oferta de itens virtuais para jogadores iniciantes com foco na conversão do usuário não-pagante para usuário pagante. Superior-Direta) Tela de jogo com ação de retenção para incentivar o usuário entrar diariamente no jogo para obter benefícios (itens virtuais). Inferior-Esquerda) Tela de jogo com demonstração do cross-selling. Inferior-Direita) Tela de jogo com banner na parte superior para geração de receita através de publicidade. 73
Figura 3-10: Relação entre a taxa de retenção e o LTV. O eixo da esquerda apresenta a taxa de aumento no LTV provocado pelo respectivo aumento no eixo da direita referente à taxa de retenção (**r**). A simulação mantém constante os valores das demais variáveis (m=10; d=10; n=10)..... 77
Figura 3-11: Relação entre taxa de retenção e a taxa de aumento no LTV. A simulação da taxa de aumento no LTV (eixo y) utiliza os valores de retenção (eixo x) e mantém constante os valores das demais variáveis (m=10; d=10; n=10). A taxa de aumento no LTV usa como base a retenção de 15%..... 77
Figura 3-12: Distribuição dos artigos pesquisados em relação à fonte de dados e ao domínio. 81
Figura 3-13: Distribuição dos artigos com relação à data de publicação (cima), plataforma (esquerda) e modelo de negócio (direta) dos jogos avaliados nos trabalhos..... 83
Figura 3-14: Probabilidade de abandono nos meses subsequentes (esquerda) e Duração média da sessão e usuários churners e non-churners (Kawale et al. 2009)..... 85
Figura 3-15:Avaliação dos modelos a partir da curva Lift (Borbora et al. 2011). 86
Figura 3-16: Análise dos métodos através da curva ROC aplicada aos três jogos sob análise (Castro & Tsuzuki 2015)..... 90
Figura 3-17: Histograma de dias de inatividade do jogo Diamond Dash (Runge et al. 2014). 95
Figura 3-18: Centróides dos clusters produzidos a partir das amostras de dados (Borbora & Srivastava 2012). Esquerda) Churners. Direita) Non-churners. O eixo x apresenta a janela de desempenho com tamanho de 13 enquanto o eixo y mostra o número de sessões por semana. 96
Figura 4-1: Distribuição da quantidade de sessões sobre o período de 16 dias para dois usuários hipotéticos. O eixo X representa a linha do tempo da janela de performance e o eixo Y indica o número de sessões realizadas em cada um dos dias. A) Usuário fictício nº 1. B) Usuário fictício nº 2..... 106
Figura 4-2: Taxa de decaimento da retenção com relação ao tempo..... 112

Figura 5-1: Esquema representativo de um experimento. 117
Figura 5-2: Gráficos de efeitos principais com apresentação das médias de performance ajustadas. 123
Figura 5-3: Gráficos de efeitos das e interação com apresentação das médias de performance ajustadas 124
Figura 5-4: Curva ROC e área sob a Curva ROC (AUC). 130
Figura 6-1: Jogo 7 Seas desenvolvido pela empresa Big Hut Games. Neste jogo casual do gênero quebra-cabeça (Match-3), o usuário é convidado a explorar os sete mares e encontrar os tesouros escondidos em centenas de missões. 136
Figura 6-2: Imagens ilustrativas do jogo Dino Jump: The Best Adventure. 138
Figura 6-3: Imagem ilustrativa do jogo Armies and Ants. 140
Figura 6-4: Distribuição das sessões e instalações do jogo durante o período de observação (julho de 2015 a maio de 2016). Esquerda) Distribuição das sessões de jogo. Direita) Distribuição das instalações do jogo. 142
Figura 6-5: Distribuição dos jogadores por tempo e frequência de jogo durante o período de observação (julho de 2015 a maio de 2016). Esquerda) Distribuição dos jogadores pela frequência relativa e tempo total de jogo. Direita) Distribuição dos jogadores por tempo de jogo. 143
Figura 6-6: Representação gráfica do processo de condução de experimento relativo ao tratamento. 149
Figura 6-7: Gráfico de efeitos principais. 155
Figura 6-8: Gráfico de interação entre níveis de fatores. 156
Figura 6-9: Padrões de interação entre os fatores com maior contribuição na alocação da variação. 156
Figura 6-10: Gráficos de resíduos. A) no canto superior-esquerdo, o gráfico de Probabilidade Normal. B) no superior-direito, o gráfico de Resíduos Versus Valores Ajustados. C) no inferior-esquerdo, o Histograma de Resíduos. D) no inferior-direito, o gráfico de Resíduos Versus a Ordem de Coleta dos Dados. 160
Figura 6-11: Gráfico de boxplot da distribuição para identificação de outliers. 161
Figura 6-12: Comparação da performance dos experimentos para as três bases de dados. ... 162
Figura 6-13: Comparação de Efeitos principais para os três principais fatores do experimento com contribuição significativa (>1%). 163
Figura 7-1: Gráfico de efeitos principais (Técnica de Modelagem). 168

Figura 7-2: Ilustração do particionamento do espaço para as técnicas de Árvore de Decisão (esquerda) e Regressão Logística (direita) para dois exemplos distintos (cima e baixo). 170
Figura 7-3: Gráfico de interação entre Técnica de Modelagem e Disposição da Janela. 171
Figura 7-4: Gráfico de efeitos principais (Tamanho da Janela). 173
Figura 7-5: Gráfico de efeitos principais para o experimento com Redes Neurais como único classificador. 174
Figura 7-6: Gráfico de interação para o experimento com Redes Neurais como único classificador. 175
Figura 7-8: Gráfico de efeitos principais para o experimento com Redes Neurais (somente). 179
Figura 7-9: Avaliação da taxa de valores ausentes nos atributos (variáveis independentes) para cada uma das disposições. 180

Lista de Tabelas

Tabela 1-1: Análise comparativa entre os domínios de aplicação de previsão de abandono. ..25
Tabela 2-1: Análise das especificidades e diretrizes na aplicação de previsão de abandono no domínio de Telecomunicações. 57
Tabela 3-1: Classificação da revisão bibliográfica. 78
Tabela 3-2: Termo de busca. 79
Tabela 3-3: Distribuição dos artigos de acordo com o ano e domínio de pesquisa, onde MMOG = Massive Multiplayer Online Games e MOBA = Multiplayer Online Battle Arena.....82
Tabela 3-4: Lista final de atributos utilizados para modelagem do classificador. 89
Tabela 3-5: Comparativo das técnicas via precisão, recall e F1-score. 100
Tabela 3-6: Comparativo das técnicas via Lift e AUC. 101
Tabela 4-2: As especificidades, seus impactos em decisões-chaves e escolhas identificadas na literatura para o domínio de jogos móveis com relação à taxa de abandono e ciclo de vida. 104
Tabela 4-3: As especificidades, seus impactos em decisões-chaves e escolhas identificadas na literatura para o domínio de jogos móveis com relação aos dados dependentes da aplicação e à taxa de conversão.....108
Tabela 5-1: Performance para os tratamentos em um experimento Fatorial Completo (exemplo ilustrativo).....121
Tabela 5-2: Resultado da Análise de Variância (exemplo ilustrativo). 122
Tabela 5-3: Resultado da Análise de Regressão para o exemplo ilustrativo. 125
Tabela 5-4: Apresentação dos níveis do fator Tamanho da Janela de Performance. 128
Tabela 5-5: Apresentação dos níveis do fator Disposição da Janela de Performance. 128
Tabela 5-6: Apresentação dos níveis do fator Tipos de Dados. 128
Tabela 5-7: Projeto experimental e seus respectivos fatores e níveis.....131

Tabela 5-8: Exemplo ilustrativo da tabela com todos os tratamentos e os respectivos resultados. 132
Tabela 5-9: Exemplo ilustrativo dos resultados dos experimentos para as três bases de dados. 133
Tabela 5-10: Interpretação dos resultados da Correlação de Pearson, segundo (Cohen 1988). 133
Tabela 6-1: Formato da mensagem JSON enviada do jogo móvel para o servidor para armazenamento das principais ações realizadas pelos jogadores dentro do jogo 7 Seas. 136
Tabela 6-2: Lista com todos os parâmetros relativos às ações realizadas pelo jogador. 137
Tabela 6-3: Estatísticas básicas sobre a base de dados do jogo 7 Seas. 138
Tabela 6-4: Estatísticas básicas sobre a base de dados do jogo Dino Jump. 139
Tabela 6-5: Estatísticas básicas sobre a base de dados do jogo Armies and Ants. 141
Tabela 6-6: Campos dos registros de ações com maior prevalência de missing data. 143
Tabela 6-7: Fatores com impacto na construção dos dados. 146
Tabela 6-8: Fatores com impacto na construção dos dados. 149
Tabela 6-9: Exemplo de tratamento. 150
Tabela 6-10: Exemplo ilustrativo da tabela com todos os tratamentos e os respectivos resultados dos modelos preditivos. 152
Tabela 6-11: Resultado da transformação dos dados do jogo Dino Jump 152
Tabela 6-12: Resultado da transformação dos dados do jogo Armies and Ants. 153
Tabela 6-13: Resultado da Análise de Variância. 154
Tabela 6-14: Sumário do modelo de regressão e seus coeficientes. 158
Tabela 6-15: Resultado das correlações de Pearson (esquerda) e Rô de Spearman (direita). 164
Tabela 7-1: Avaliação do resultado da ANOVA para os fatores do projeto experimental. ... 167
Tabela 7-2: Avaliação do resultado da ANOVA para as interações entre fatores relevantes (contribuição > 1%). 167
Tabela 7-3: Resultado da ANOVA para o projeto experimental com o nível Rede Neural mantido estático (sem variação). 174
Tabela 7-4: Avaliação dos coeficientes da regressão logística. 182
Tabela 7-5: Apresentação dos coeficientes de maior magnitude da regressão logística. 183
Tabela 7-6: Resultados da correlação entre os resultados experimentais para as 3 bases de dados. 184
Tabela 7-7: Análise comparativa da performance do classificador gerado a partir das diretrizes com o estado da arte. 188

Sumário

1 Introdução 23

1.1 Contexto e Motivação 23

1.2 Objetivos 26

1.3 Abordagem 26

1.4 Estrutura da Tese 28

2 Fundamentos da Previsão de Abandono 29

2.1 O Abandono sob a Ótica de Negócios 29

2.1.1 Relacionamento com o Cliente 31

2.1.2 Ciclo de Vida do Cliente 32

2.1.3 Importância da Retenção de Clientes 34

2.2 A Previsão de Abandono 35

2.2.1 Mineração de Dados 36

2.2.2 Mineração de Dados Orientada ao Domínio 38

2.2.3 Behavior Scoring 38

2.2.3.1 Tamanho da Janela 42

2.2.3.2 Disposição da Janela 42

2.3 A Construção de Modelos Preditivos e a Previsão de Abandono 45

2.3.1 Etapa 1: Entendimento do Negócio 46

2.3.2 Etapa 2: Entendimento dos Dados 47

2.3.3 Etapa 3: Preparação dos Dados *48*
2.3.4 Etapa 4: Modelagem *50*
2.3.5 Etapa 5: Avaliação *51*
2.3.6 Etapa 6: Implantação *54*
2.4 As Diretrizes em Domínios Estabelecidos 54
2.4.1 Definição do Conceito de Diretriz *55*
2.4.2 As Especificidades e Diretrizes para Previsão de Abandono em Telecomunicações *57*
2.5 Conclusão e Observações 58
3 Previsão de Abandono em Jogos Móveis 60
3.1 A Indústria de Jogos Móveis 60
3.2 Evolução dos Modelos de Negócios em Jogos Móveis 61
3.2.1 Modelos de Distribuição *62*
3.2.2 Modelos de Negócios *66*
3.3 O Modelo de Negócios Free-to-Play 67
3.4 A Gestão do Relacionamento com o Jogador 70
3.4.1 Ciclo de Vida do Jogador *70*
3.4.2 Importância da Retenção *74*
3.5 A Construção de Modelos Preditivos em Jogos Móveis 78
3.5.1 Etapa 1: Entendimento do Negócio *83*
3.5.2 Etapa 2: Entendimento dos Dados *84*
3.5.3 Etapa 3: Preparação dos Dados *91*
3.5.3.1 Janelas de Observação e Resultado 91
3.5.3.2 Rótulo do Usuário 93
3.5.4 Etapa 4: Modelagem *96*
3.6 Conclusão e Observações 102
4 Avaliação das Especificidades de Jogos Móveis 103
4.1 As Especificidades e seus Impactos 103

4.1.1 Taxa de Abandono e Ciclo de Vida 104
4.1.1.1 Tamanho da Janela 104
4.1.1.2 Disposição da Janela 105
4.1.2 Dados Dependentes da Aplicação e Taxa de Conversão 107
4.1.2.1 Dados Cadastrais 108
4.1.2.2 Dados Financeiros 108
4.1.2.3 Dados Comportamentais 109
4.1.3 Modelo Freemium e Cancelamento Voluntário Implícito 110
4.2 As Escolhas a Serem Avaliadas Experimentalmente 111
4.2.1 Configuração da Janela de Performance 111
4.2.2 Tipos de Dados 113
4.2.3 Configuração da Janela de Resultado 114
4.3 Conclusão e Observações 115
5 Planejamento dos Experimentos 116
5.1 Fundamentos de Projetos Experimentais 116
5.1.1 Princípios dos Experimentos 118
5.1.2 Tipos de Projetos Experimentais 119
5.1.3 Análise de Experimentos 121
5.1.3.1 Análise de Variância (ANOVA) 121
5.1.3.2 Análise de Regressão 124
5.2 Plano Experimental 126
5.2.1 Caracterização do Problema 127
5.2.2 Escolha dos Fatores de Influência e Níveis 127
5.2.2.1 Tamanho da Janela 127
5.2.2.2 Disposição da Janela 128
5.2.2.3 Tipos de Dados 128
5.2.2.4 Técnica de Modelagem 128
5.2.3 Seleção da Variável de Resposta 129
5.2.4 Determinação do Modelo de Planejamento de Experimento 131
5.2.5 Comparação dos Resultados 132

5.3 Conclusão e Observações 134

6 Execução dos Experimentos e Apresentação dos Resultados 135

6.1 Apresentação das Bases de Dados 135

6.1.1 Jogo Móvel #1: 7 Seas 135

6.1.2 Jogo Móvel #2: Dino Jump 138

6.1.3 Jogo Móvel #3: Armies and Ants 140

6.2 Preparação dos Experimentos 141

6.2.1 Entendimento de Negócios e Dados 142

6.2.2 Preparação dos Dados: Seleção e Limpeza 144

6.2.2.1 Exclusão de registros desnecessários 144

6.2.2.2 Exclusão de campos desnecessários 145

6.2.2.3 Transformação do Grão 145

6.2.3 Preparação dos Dados: Construção 145

6.2.3.1 Configuração das Janelas de Performance 146

6.2.3.2 Tipos de Dados 146

6.2.4 Preparação dos Dados: Definição do Rótulo 147

6.2.5 Preparação dos Dados: Formatação 148

6.3 Condução dos Experimentos 149

6.3.1 Configuração do Tratamento 150

6.3.2 Leitura do Arquivo CSV 150

6.3.3 Seleção dos Atributos 150

6.3.4 Seleção da Técnica de Modelagem 150

6.3.5 Validação do Modelo 151

6.3.6 Avaliação da Performance 151

6.3.7 Resumo da Execução 152

6.4 Apresentação dos Resultados do Jogo 7 Seas 153

6.4.1 Análise de Variância 154

6.4.2 Análise de Regressão 157

6.4.3 Validação do Modelo 158

6.5 Apresentação dos Resultados dos Outros Jogos 161

6.6 Conclusão e Observações 164

7 Análise dos Resultados 166

7.1 Achados Relevantes 166

7.1.1 Técnica de Modelagem 167

7.1.1.1 Análise 168

7.1.1.2 Discussão 169

7.1.1.3 Conclusões 171

7.1.2 Tamanho da Janela de Performance 172

7.1.2.1 Análise 173

7.1.2.2 Discussão 175

7.1.2.3 Conclusões 177

7.1.3 Disposição da Janela de Performance 178

7.1.3.1 Análise 178

7.1.3.2 Discussão 179

7.1.3.3 Conclusões 180

7.1.4 Tipos de Dados 180

7.1.4.1 Análise 181

7.1.4.2 Discussão 182

7.1.4.3 Conclusões 183

7.1.5 Discussão sobre os Resultados para as Bases de Jogos 183

7.2 Diretrizes 184

7.2.1 Decisão-Chave: Técnica de Modelagem 185

7.2.2 Decisão-chave: Tamanho da Janela de Performance 185

7.2.3 Decisão-chave: Disposição dos Dados 185

7.2.4 Decisão-chave: Tipos de Dados 186

7.2.5 Outras Recomendações 186

7.3 Aplicação das Diretrizes 187

8 Conclusão e Trabalhos Futuros 190

8.1 Objetivos e Contribuições 190
8.2 Limitações 191
8.3 Trabalhos Futuros 191

Referências 193

1 Introdução

A área de pesquisa relativa à previsão de abandono de usuários é importante para o sucesso comercial de empresas em todas as indústrias e na área de jogos móveis é particularmente relevante devido às características desse mercado. Nesse capítulo, nós apresentamos o contexto, motivação e objetivos do presente projeto além da apresentação da estrutura organização do trabalho.

1.1 Contexto e Motivação

Os últimos anos presenciaram a ascensão dos jogos para dispositivos móveis, doravante chamados somente de jogos móveis, associada a adesão ao modelo de negócios Free-to-Play, em que o jogador não paga para ter acesso ao jogo, mas paga dentro do jogo para ter acesso a certos bens virtuais. Essas mudanças têm levantado novas questões sobre a forma como as pessoas consomem jogos, bem como provocado uma mudança fundamental na visão de jogos como produtos para jogos como serviços. Isso significa que o modelo de negócios com o consumidor evoluiu de uma transação (único pagamento adiantado) para um relacionamento de monetização (pagamentos constantes realizados ao longo de vários meses).

Neste novo panorama, a receita por usuário ativo substituiu a quantidade de unidades vendidas no varejo como medida de sucesso comercial. Além disso a conexão com os usuários finais é persistente. A monetização agora depende de engajamento, retenção e estímulo à realização de compras dentro do jogo. Isso significa que, para sobreviver, os desenvolvedores de jogos móveis devem colocar em prática estratégias de retenção para engajar jogadores por um período mais longo e assim criar mais oportunidades de convertê-los em usuários pagantes. A receita de jogos móveis *Free-to-Play* é diretamente dependente das taxas de engajamento

dado que a monetização é irrelevante sem retenção. Em outras palavras, um jogo que não retém usuários, não consegue gerar receita.

Nesse cenário, a previsão de abandono surge como uma atividade fundamental para antecipar a intenção de abandono do jogador e permitir à empresa desenvolvedora de jogos aplicar ações proativas de retenção para reverter essa tendência. Entretanto, a construção de modelos preditivos no domínio de jogos ainda é uma área de pesquisa incipiente com poucos trabalhos publicados. Dos 9 artigos identificados nesse domínio, somente 2 deles tratam do problema no domínio de jogos para dispositivos móveis, onde o abandono é mais crítico.

As pesquisas avaliadas na revisão de literatura aplicam metodologias genéricas, baseadas na mineração de dados tradicional orientada a dados, sem a devida discussão sobre as especificidades do domínio de jogos móveis e seus possíveis impactos na construção e na performance do modelo preditivo. Além disso, as decisões de projeto relativas ao processamento e tratamento dos dados, responsável por 50-80% do tempo consumido na atividade de construção do modelo preditivo, são *ad hoc* (Hall et al. 2011). As pesquisas aplicam de maneiras distintas as técnicas *Domain-Driven Data Mining* (D3M) e *Behavior Scoring* (BS), consagradas em outras áreas de aplicação da previsão de abandono, e há pouco ou nenhum reuso de conhecimento entre as pesquisas. Em outras palavras, ainda não há diretrizes claras para a construção de modelos preditivos para abandono em jogos móveis.

Contrariamente à área de jogos móveis, domínios maduros, como telecomunicações e varejo, contam com mais de 20 anos de pesquisa e centenas de trabalhos publicados na área de previsão de abandono (Ngai et al. 2009; Hashmi et al. 2013). Nesses domínios os desafios específicos relacionados ao domínio já são conhecidos, assim como as soluções adotadas para tratamento dessas especificidades do domínio. As pesquisas aplicam as técnicas D3M e BS de maneira similar, com reuso das diretrizes construídas e aprimoradas ao longo do tempo. Essas diretrizes são orientações e recomendações que servem para guiar pesquisadores e profissionais na construção de modelos preditivos, de forma a incentivar o reuso de boas práticas analisadas e aprovadas em problemas recorrentes no desenvolvimento de aplicações. Isso evita desperdício, duplicação de trabalho e ajuda a convergir mais rapidamente para um bom modelo preditivo.

A Tabela 1-1 compara as algumas características do domínio de jogos free-to-play móveis com a área de telecomunicações. As diferenças são significativas. Enquanto a área de telecomunicações enfrenta taxas entre 20 - 40% ao ano (Hung et al. 2006), algo em torno de 2,2% ao mês, a taxa de abandono média dos jogos para dispositivos móveis é de aproximadamente 85% ao mês (Lovell 2011). Nesse caso, a duração do relacionamento com o

usuário é extremamente curta assim como a janela de tempo para extração dos dados do usuário para análise.

	Telecomunicações	**Jogos *free-to-play* para dispositivos móveis**
Taxa de Abandono	20-40% ao ano	85% ao mês
Tempo de Observação	12-24 meses	2-3 meses
Categorias de Dados	Pessoais Comportamentais Monetários	- Comportamentais Monetários
Taxa de Conversão	100%	2,2%
Natureza dos Dados	Independente da aplicação	Dependente da aplicação
Definição do rótulo	Explícito (cancelamento da assinatura/contrato)	Implícito (suspensão do uso)

Tabela 1-1: Análise comparativa entre os domínios de aplicação de previsão de abandono.

Além da restrição no volume de dados, os tipos de dados disponíveis também apresentam limitações. É comum em outras indústrias (Hashmi et al. 2013)a utilização de dados pessoais (dados da conta e informações sociodemográficas), comportamentais (histórico do relacionamento com a instituição) e monetários (histórico de compras e pagamentos). Em dispositivos móveis, entretanto, é vedada a obtenção dos dados pessoais do usuário por questões de segurança. Os dados comportamentais até estão disponíveis, mas diferem bastante a depender da categoria do jogo. As ações de um jogador realizadas dentro do jogo *Angry Birds*, por exemplo, produzirão dados comportamentais completamente distintos daqueles extraídos do jogo *Age of Empires*. Os dados monetários também estão disponíveis, entretanto a distribuição entre usuários pagantes e não-pagantes é extremamente desequilibrada. A taxa de conversão média da indústria, de não-pagante para pagante, é de 2%.

A determinação da situação do usuário quanto ao abandono, se ainda ativo ou desistiu do jogo (*churner* ou *non-churner*), também é um problema nos jogos e essa informação é essencial para modelagem de classificadores binários através de aprendizagem supervisionada. Em outras indústrias, a definição desses status é realizada, em geral, com base em uma ação explícita do usuário, como o cancelamento do serviço de telefonia móvel. Nesse caso, a informação é clara e explícita. Em jogos free-to-play, por outro lado, o abandono do usuário acontece de forma implícita pela redução do engajamento do usuário, sem aviso prévio.

1.2 Objetivos

Todas essas diferenças mostram, seguindo uma lógica D3M, que seria muito importante entender melhor as peculiaridades da indústria de jogos móveis e suas consequências em termos de construção e desempenho de modelos preditivos de abandono de jogadores. Entre outras iniciativas, deveria ser possível responder questões tais como:

- Quais as principais especificidades dessa indústria relevantes para a modelagem do problema?
- Qual o peso dessas especificidades, e seus possíveis tratamentos, na performance da previsão?
- Quais diretrizes podem ser aplicadas para tratamento dessas especificidades no processo de construção do modelo preditivo?
- As diretrizes propostas são replicáveis para outros jogos dentro do mesmo domínio?

Neste contexto, o objetivo geral da nossa pesquisa é identificar e estudar os desafios específicos do domínio de jogos móveis de forma a propor diretrizes para construção de modelos preditivos de abandono. A expectativa é que as futuras pesquisas utilizem esse processo para orientar as pesquisas, aperfeiçoar a performance dos modelos preditivos e facilitar a comparação dos resultados alcançados.

1.3 Abordagem

A abordagem para elaboração das diretrizes (Figura 1-1) é iniciada com a revisão da literatura. Essa etapa inclui a atividade de revisão *ad-hoc* da literatura relativa ao tema de previsão de abandono de clientes em diferentes indústrias. O estudo do problema em diferentes domínios de aplicação revela a importância da aplicação de técnicas baseadas em D3M e BS para tratamento das especificidades relativas ao domínio. Nessa etapa é realizada ainda a revisão sistemática da literatura para análise das abordagens, métodos e técnicas aplicadas no estado da arte para construção de modelos preditivos de abandono em jogos móveis.

Na etapa de estudo das especificidades do domínio, a avaliação do contexto de aplicação e das características das indústrias são avaliadas. Essa pesquisa aborda as características de produção, distribuição e comercialização dos jogos móveis com impacto no modelo de

relacionamento entre os jogadores e as empresas desenvolvedoras de jogos. Essa análise examina as características únicas do relacionamento entre os usuários e os jogos móveis para compreensão das decisões-chave, e suas respectivas escolhas, no estado da arte da construção de modelos para previsão de abandono. Enfim, elencamos as escolhas que pretendemos avaliar experimentalmente a fim de obter possíveis diretrizes para área.

A avaliação das escolhas é realizada através do planejamento, execução e avaliação de projeto experimental. As decisões-chave e as respectivas propostas de escolhas são modeladas como fatores e níveis de fatores em um projeto experimental fatorial completo. A variável de resposta considerada, para determinação do efeito de cada uma das possíveis propostas, é a performance da previsão de abandono. Os experimentos são executados para 3 bases de dados pertencentes a jogos móveis distintos. Essas bases contam com mais de 200 mil jogadores e praticamente 5 milhões de sessões de jogos.

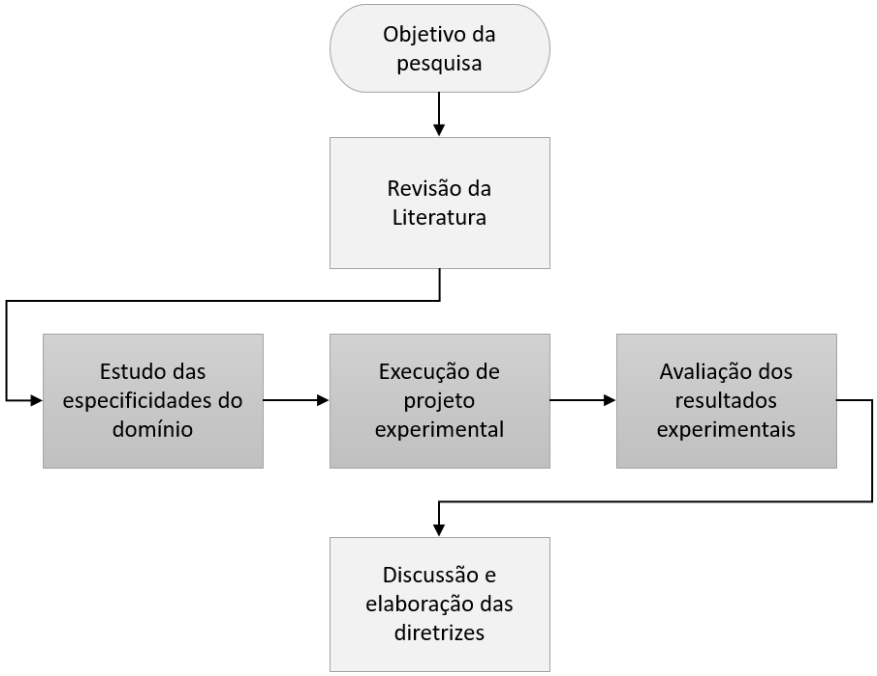

Figura 1-1: As etapas do projeto de pesquisa.

Os resultados são analisados para identificação do impacto na performance do modelo preditivo de cada uma das decisões-chaves e das escolhas propostas. Os resultados revelam uma série de descobertas utilizadas para elaboração das diretrizes para construção de modelos preditivos de abandono em jogos móveis. As diretrizes são ainda aplicadas em uma das bases de dados investigadas para construção dos classificadores e comparação com as abordagens adotadas na revisão da literatura.

1.4 Estrutura da Tese

Os capítulos restantes dessa tese estão organizados da seguinte forma. No Capítulo 2, nós apresentamos os fundamentos teóricos da disciplina de previsão de abandono de usuários. Em seguida, nós investigamos a aplicação da previsão de abandono em diferentes indústrias e concluímos que os desafios específicos relacionados a domínio estabelecidos já são conhecidos assim como as soluções adotadas para tratamento dessas especificidades do domínio.

O Capítulo 3 apresenta o contexto do problema de previsão de abandono em jogos móveis e também a revisão bibliográfica dentro do tema da pesquisa. O Capítulo 4 investiga as principais especificidades do domínio de jogos móveis para compreensão das decisões-chave, e suas respectivas escolhas, no processo de construção de modelos preditivos de abandono. Nós elencamos ainda as escolhas candidatas à diretrizes.

No Capítulo 5 nós apresentamos o plano do projeto experimental para avaliação das escolhas para os desafio-chaves do domínio e seus respectivos efeitos na performance do modelo preditivo. O Capítulo 6 apresenta as atividades realizadas na execução do projeto experimental para as bases de dados de jogos móveis reais. Em seguida os resultados dos experimentos são apresentados e brevemente discutidos.

No Capítulo 7 nós discutimos os principais achados realizados com a condução dos experimentos para elaboração das diretrizes para construção de modelos preditivos de abandono em jogos móveis. As diretrizes são avaliadas com relação às abordagens adotadas na revisão da literatura. Finalmente, no Capítulo 8, nós apresentamos a conclusão da tese com a descrição da contribuições, limitações e trabalhos futuros.

2 Fundamentos da Previsão de Abandono

O termo *previsão de abandono*, ou *churn prediction* do Inglês, é usado para descrever o processo de avaliação do risco de um determinado consumidor abandonar um serviço ao qual está vinculado. Esse abandono corresponde ao término do relacionamento entre o consumidor e a empresa prestadora do serviço e representa a etapa final do ciclo de vida do cliente. Nesse cenário, a previsão de abandono consiste em uma importante atividade na gestão do relacionamento com o usuário para antecipar o movimento de abandono e proativamente realizar ações promocionais para retenção e fidelização do cliente.

Neste capítulo nós apresentamos a fundamentação teórica e prática das principais estratégias de negócios utilizadas na indústria para criar e manter um relacionamento mutuamente benéfico e de longo prazo com seus clientes. Em seguida, apresentamos os conceitos básicos, os processos, as abordagens, as técnicas e as métricas da previsão de abandono e evolução dessa área de pesquisa com aplicação em diferentes indústrias. Por fim, mostramos que em domínios mais estabelecidos, como a indústria de Telecomunicações, os principais desafios da construção de modelos preditivos são conhecidos, assim como as suas diretrizes e melhores práticas.

2.1 O Abandono sob a Ótica de Negócios

Em um cenário de alta competitividade globalizada (Kim & Mauborgne 2005), as organizações disputam entre si a atenção do cliente para conseguir comercializar seus produtos e serviços, expandir suas carteiras e assim aumentar o lucro. O objetivo maior dessas empresas consiste em acrescentar valor às suas organizações através do prolongamento do tempo de relacionamento e aumento do consumo médio dos clientes, assim como da aquisição de novos clientes (ver Figura 2.1).

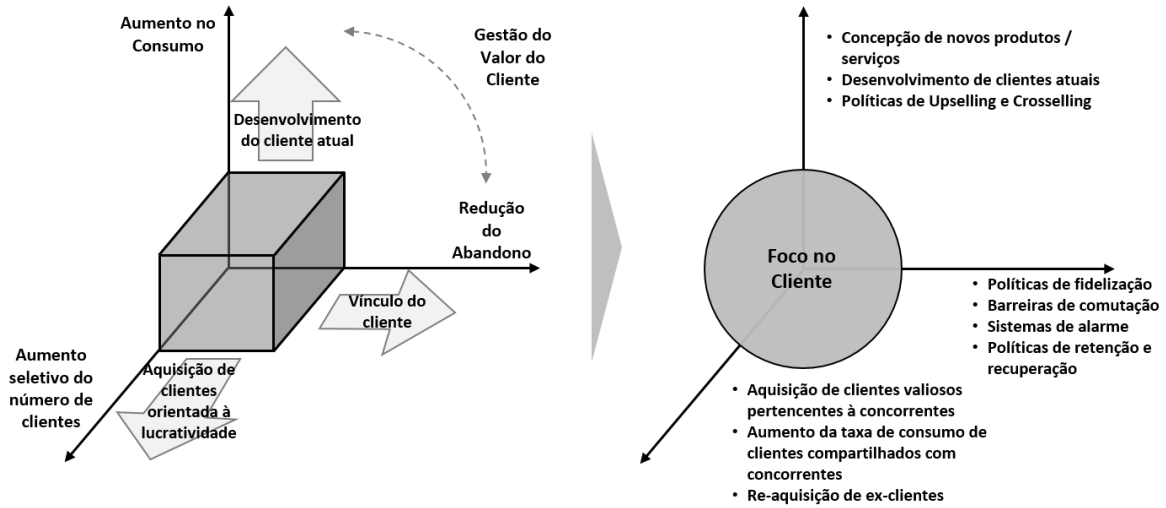

Figura 2-1: Alternativas de desenvolvimento comercial em mercados maduros. Esquerda) a figura demonstra os principais eixos relativos ao valor comercial de uma organização: nos eixos x e y, a redução do abandono e aumento do consumo representam aspectos relacionados à gestão do valor do cliente, e no eixo z, o desenvolvimento de estratégias de aquisição de consumidores de alto valor. Direita) a figura apresenta as políticas direcionadas ao consumidor com relação a cada um dos eixos estratégicos.

Portanto, a compreensão dos mecanismos de fidelização do usuário, antecipação da intenção de abandono do cliente e realização de ações orientadas à retenção dos consumidores são elementos importantes para promoção da vantagem competitiva entre as empresas atuantes de um mesmo mercado. Dessa maneira, a estratégia comercial defensiva orientada à retenção e criação de vínculo com o consumidor é muito mais efetiva, e menos custosa, do que a estratégia agressiva de expansão da carteira de clientes com a atração em massa de potenciais consumidores.

Segundo (Fox & Kotler 1994), conquistar clientes novos custa até 7 vezes mais caro do que manter os mesmos clientes que já possui. E mais, segundo pesquisas recentes (Stillwagon 2014), um incremento em 5% na retenção implica em 50% de aumento no lucro por cliente, em média, considerando múltiplas indústrias, e chegando a 90% de aumento em indústrias específicas, como a de Seguros. Como resultado, as principais empresas estão começando a modificar os seus paradigmas comerciais da captura massiva de novos consumidores para a conservação dos clientes já existentes.

A complexa missão de fidelização do consumidor, entretanto, colide com intensa e constante exposição dos clientes a novas ofertas realizadas pelos competidores. Por haver diversas opções, os consumidores têm o poder de escolher a oferta que mais lhes convêm, obrigando as empresas a trabalhar em sintonia de qualidade e preço com as demais empresas concorrentes.

O resultado prático é o aumento dos custos operacionais com efeito direto na redução das margens de lucros das empresas competidoras. Nesse ambiente, a importância de compreender os mecanismos envolvidos na construção de vínculos mais profundos para fidelização dos clientes torna-se extremamente importante para garantir a continuidade da empresa no mercado. Portanto, a gestão do relacionamento com o cliente para a criação e estabelecimento de relações comerciais estáveis tem se tornando o alvo principal dos esforços de marketing.

2.1.1 Relacionamento com o Cliente

A gestão do relacionamento com o cliente, do inglês *Customer Relationship Management* (CRM), consiste em uma visão estratégica e orientada a ação do relacionamento entre o consumidor e a companhia (Xu et al. 2002; Bauer et al. 2002). CRM provê metodologias, estratégias, processos e tecnologias para suportar a gestão desse relacionamento. O objetivo do CRM é aumentar de forma eficiente e efetiva a aquisição, crescimento e retenção de clientes rentáveis, iniciando, construindo e mantendo relações apropriadas com o público consumidor da empresa.

Até a década de 50, as empresas focavam no volume de produção para fabricação em grande quantidade de produtos e serviços pouco diferenciados. A orientação na época era explorar a economia de escala. Nessa década, frameworks como o 'marketing mix' foram desenvolvidos para explorar a demanda do mercado. A abreviatura dos "4Ps" de produto, preço, promoção e praça foi usada para descrever as alavancas que, se puxadas adequadamente, levariam ao aumento da demanda pela oferta da empresa. O objetivo desta abordagem "transacional" para o marketing era desenvolver estratégias para otimizar as despesas com o mix de marketing e assim maximizar as vendas.

Na década seguinte, a orientação das empresas estava voltada para a qualidade dos produtos. As empresas assumiam que enquanto o produto fosse de alta qualidade, as pessoas iriam querer comprar e consumir o produto. As empresas investiram na criação de departamentos de pesquisa e desenvolvimento para elaboração e testes de novos produtos dentro desse paradigma.

Somente a partir da década de 70 é que a estratégia de marketing passou a ser orientada ao consumidor. As empresas passaram a executar pesquisas de mercado para identificar os desejos dos consumidores, criar produtos em sintonia com as informações reveladas e posteriormente utilizar técnicas de publicidade para informar as pessoas sobre o produto.

No cenário atual, a principal abordagem de marketing é orientada ao consumidor na qual desejos de consumo são os principais condutores das decisões estratégicas. Cada aspecto de uma oferta de produto ou serviço é orientado pelas necessidades de potenciais consumidores. O ponto de partida é sempre o consumidor. A principal justificativa para essa abordagem é que não há razão para gastar recursos em Pesquisa e Desenvolvimento de novos produtos que as pessoas não irão comprar.

2.1.2 Ciclo de Vida do Cliente

A compreensão da jornada realizada pelo usuário durante o período de relacionamento com a empresa permite captar e nutrir as necessidades específicas do consumidor em cada momento do seu ciclo de vida. O ciclo de vida do usuário, desde o início do seu relacionamento com a empresa até o abandono, apresenta variações específicas de acordo com as características do negócio, tais como área da indústria, tipos de produto e serviço ofertados e perfil do público-alvo. Para efeitos didáticos, nós apresentamos um modelo genérico de ciclo de vida do cliente para descrever as principais etapas da jornada encontrada na maioria das indústrias.

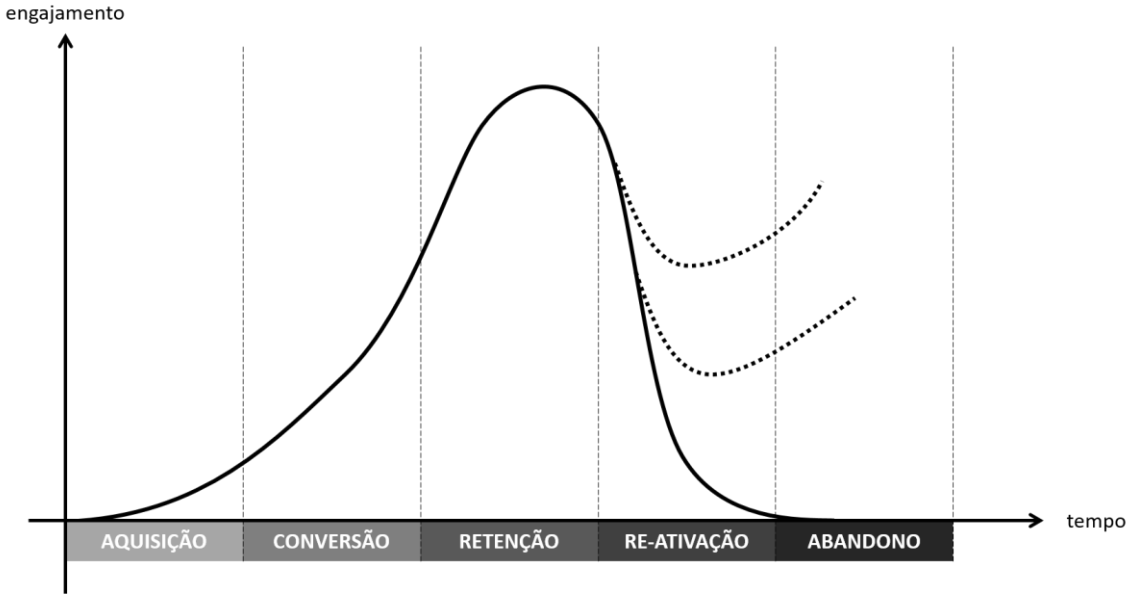

Figura 2-2: Etapas do ciclo de vida do cliente. A jornada do consumidor é representada através de gráfico que relaciona o período de relacionamento do usuário com a empresa (eixo X) com o engajamento apresentado pelo consumidor ao longo do ciclo de vida (eixo Y) em cada uma das etapas descritas sobre esse eixo.

As etapas expostas no eixo X da Figura 2-2 estão descritas em maiores detalhes a seguir. A etapa de **aquisição** consiste em atrair potenciais clientes para a esfera de influência de uma organização. Nessa fase, as atividades de marketing incluem a definição dos potenciais clientes através da pesquisa e qualificação do perfil do público-alvo com potencial para aquisição do

produto/serviço ofertado pela empresa. Após essa definição, ações de marketing promocional direcionadas aos potenciais clientes são realizadas com o propósito de atrair a atenção das pessoas e direcionar os usuários impactados para iniciarem relacionamento com a companhia.

O abandono pode acontecer neste ponto do processo quando a empresa, produto ou serviço não é interessante ou útil o suficiente para manter a atenção de um potencial cliente. O abandono nessa fase é classificado como *bounce*. É crítico que o primeiro contato contenha bastante valor para que esse usuário impactado deseje continuar o relacionamento até a conversão.

A etapa de **conversão** é um sinônimo para venda. Essa etapa representa o momento em que os potenciais clientes na esfera de influência de uma companhia decidem realizar a sua primeira compra. Esses usuários, portanto, são convertidos de usuários não-pagantes a clientes pagantes. A etapa de **retenção** envolve a realização de ações para alongar ao máximo o relacionamento com o cliente para, assim, propiciar a aplicação das práticas de *upselling* e *cross-selling*. *Upselling* é a prática em que uma empresa tenta persuadir os clientes a comprarem um produto mais sofisticado ou realizar um upgrade do seu serviço visando a uma venda mais lucrativa. Por exemplo, na indústria de telecomunicações é comum as operadoras de telefonia oferecerem aos seus clientes planos com recursos adicionais como maior quantidade de minutos, internet mais veloz e maior quantidade de dados de internet. No C*ross-selling,* o vendedor sugere a compra de produtos adicionais à venda principal. Ainda no mesmo exemplo na indústria de telefonia móvel, também é comum que as operadoras abordem seus clientes para oferecer outros produtos como telefone fixo e TV por assinatura. É benéfico para as empresas a utilização de ambas as técnicas, a fim de aumentar a receita e proporcionar uma experiência de consumo valorizado.

Os consumidores podem abandonar uma empresa por diversos motivos, como encontrar alternativas mais econômicas ou até mesmo não precisar mais do produto/serviço ofertado. Em todos os casos, as empresas podem realizar ações de **reativação,** como promoções e oferta de novos produtos/serviços, para reter os usuários e evitar o abandono. O **abandono** representa o fim do relacionamento entre a companhia e o usuário através do cancelamento do contrato de prestação de serviço. Após o abandono, a única maneira de atrair novamente o consumidor consiste na realização de ações de reaquisição com o objetivo de iniciar um novo ciclo com o cliente.

O custo de aquisição e também de reaquisição de usuários é muito maior do que o custo de retenção dos usuários, como já mencionado no início do capítulo. Pesquisas recentes encontraram uma correlação negativa entre a parcela de clientes regulares que desertaram e o

lucro das empresas (Bauer et al. 2002/2). Dessa maneira, a aplicação de processos sistemáticos e efetivos para retenção de usuário torna-se forte aliada no sucesso de empresas.

2.1.3 Importância da Retenção de Clientes

A atração de novos clientes é importante em todos as empresas, especialmente em negócios nascentes. Entretanto, conquistar clientes novos custa até 7 vezes mais caro do que manter os mesmos clientes que já se possui (Fox & Kotler 1994). E mais, segundo pesquisas recentes (Stillwagon 2014), um incremento em 5% na retenção implica em 50% de aumento no lucro por cliente, em média, considerando múltiplas indústrias, e chegando a 90% de aumento em indústrias específicas, como a de Seguros.

Se o cliente atual de uma empresa já possui histórico de compras, o custo para anunciar e vender um novo produto ou serviço para ele é consideravelmente menor do que o custo necessário para vender um produto a um novo usuário. Dado que o cliente já experimentou o processo de compra, identificou valor no produto ou serviço ofertado e continua utilizando-o com frequência, isso significa que a maioria das barreiras à compra já foram cruzadas e, portanto, menos investimento é necessário para influenciar o usuário a realizar a compra seguinte. Segundo estudos, a probabilidade de converter um cliente existente é de 60 - 70%, enquanto a probabilidade de converter um novo cliente é de apenas 5 - 20% (Advisors 2015).

A retenção também permite que a empresa construa uma visão mais específica e detalhada do perfil dos seus clientes. Esse melhor conhecimento do cliente promove 2 benefícios. Em primeiro lugar, esse conhecimento possibilita a criação de promoções personalizadas e direcionadas a cada um dos usuários com o objetivo de maximizar as chances de sucesso na venda. E, além disso, os dados sobre os clientes ajudam a empresa em seus esforços de aquisição futura. É possível saber com maior precisão que tipo de pessoas são mais susceptíveis de se tornarem clientes leais - quem eles são, o que eles fazem online, o que os motiva a comprar e que tipo de mensagens promovem melhor taxa de resposta. Esses dados permitem realizar ações de marketing direcionadas à aquisição de usuários com maior potencial de sucesso.

Os motivos acima demonstram que a aplicação com sucesso da gestão do relacionamento para fidelização do consumidor implica em benefícios diretos e impacto positivo na receita da empresa. A previsão de abandono é uma ferramenta importante na gestão do relacionamento com o usuário.

2.2 A Previsão de Abandono

A antecipação da intenção de abandono do consumidor é uma atividade chave na estratégia de retenção das empresas para prolongar o tempo de vida dos usuários. O diagnóstico precoce auxilia nesse processo ao reduzir a agressividade da ação terapêutica necessária ao mesmo tempo em que aumenta as possibilidades de recuperação do cliente.

No entanto, a atividade de previsão de abandono tornou-se uma atividade cada vez mais complexa dada a expansão da quantidade de informação armazenada nos bancos de dados das empresas. As soluções tradicionais para processamento dos dados, a partir de métodos estatísticos simplificados, não conseguem extrair informação relevante sobre a situação do cliente para auxiliar na tomada de decisão gerencial.

Nesse contexto, as técnicas de mineração de dados aplicadas às informações do cliente tornam-se fundamentais para auxiliar na compreensão do funcionamento dos mecanismos de fidelização para redução dos efeitos da intenção de abandono e aplicação de ações orientadas à retenção. De acordo com a razão envolvida no abandono, os consumidores podem ser classificados da seguinte forma:

- **Cancelamento Involuntário**

Essa forma de abandono, também chamado de cancelamento ou abandono forçado, se refere à situação em que a empresa prestadora de serviço proativamente realiza a ação de cancelamento do contrato de prestação de serviço. Em geral, as empresas não consideram esse tipo de cancelamento como abandono em seus registros.

Esse tipo de cancelamento é comumente realizado quando a empresa prestadora de serviço detecta faltas cometidas pelo consumidor (atrasos, fraudes, ...) ou registra baixo retorno sobre o investimento. Em outras palavras, o custo de manutenção do cliente é superior ao retorno financeiro obtido. Nesses cenários, a empresa opta pelo fim do relacionamento com o consumidor através do cancelamento do contrato.

- **Cancelamento Voluntário**

Essa forma de cancelamento se refere à situação em que o próprio usuário realiza a ação de cancelamento do serviço / contrato. Há duas variantes para esse tipo de cancelamento:

- **Circunstancial**: O cancelamento acontece devido a circunstâncias do próprio usuário que não o permite manter-se clientes da companhia. Os motivos são os mais variados possíveis, como mudança de endereço, alteração no status civil, filhos, dentre outros. O cancelamento, nesse caso, é intrinsecamente imprevisível.
- **Deliberado:** O cancelamento ocorre quando o consumidor voluntariamente e proativamente procura a empresa para solicitar o cancelamento do contrato da prestação de serviço.

Na presente tese, estamos interessados especificamente no último cenário: abandono voluntário e deliberado. A previsão de abandono, portanto, implica na concepção e desenvolvimento de modelos preditivos para identificação da probabilidade de abandono de clientes.

A partir da revisão da literatura, identificamos a aplicação da previsão de abandono em diferentes áreas da economia. As soluções, em geral, aplicam conceitos, técnicas e abordagens provenientes da Mineração de Dados, Mineração de Dados Orientada ao Domínio (D3M) e Behavior Scoring (BS). Nós iremos detalhar essas abordagens a seguir, assim como outros conhecimentos relevantes da literatura.

2.2.1 Mineração de Dados

A expansão da quantidade de dados transacionais das empresas levou ao desenvolvimento e adoção dos *data warehouses* como solução para o armazenamento das informações relativas às atividades da organização em bancos de dados, de forma consolidada. Essas bases de dados das empresas são responsáveis por armazenar uma grande quantidade de dados, como informações sociodemográficas dos clientes, histórico de compras e pagamentos, experiência de navegação no site da empresa, dentre outros.

Os *data warehouses* foram projetados para favorecer a produção de relatórios gerenciais a partir da análise de grandes volumes de dados. Esses relatórios visam a obtenção de informações estratégicas para facilitar a tomada de decisão dentro das empresas. Com o crescimento da quantidade e complexidade das informações armazenadas nos sistemas empresariais, as soluções tradicionais para processamento de dados não conseguem extrair informação relevante para a tomada de decisão.

A mineração de dados surgiu como uma alternativa viável para descobrir regras, identificar fatores e tendências-chave, descobrir padrões e relacionamentos ocultos em grandes

bancos de dados para auxiliar a tomada de decisões sobre estratégia e vantagens competitivas dentro das organizações. Uma das áreas de aplicação da mineração de dados, relevante para o tema de pesquisa, consiste na construção de modelos representativos. Para aprofundarmos a discussão nessa área é importante a definição dos termos abaixo:

- **Modelo:** Em estatística, um modelo é uma representação de um relacionamento entre variáveis presentes nos dados. Ele descreve como uma ou mais variáveis nos dados estão relacionadas às outras variáveis.
- **Modelagem:** É o processo no qual uma abstração representativa é construída a partir dos dados observados.

Por exemplo, nós podemos desenvolver um modelo baseado na avaliação de crédito, nível de renda e valor do empréstimo solicitado pelo usuário (dados de entrada) para determinar a taxa de juros do empréstimo (dado de saída). Para essa tarefa, nós precisamos de dados referentes a observações passadas com as informações sobre avaliação de crédito (risco), renda pessoal, valor do empréstimo solicitado e taxa de juros do empréstimo. A Figura 2.3 mostra as entradas e saídas do modelo. Uma vez que o modelo representativo é criado, nós podemos usá-lo para prever o valor da taxa de juros do empréstimo baseado nos valores de entrada.

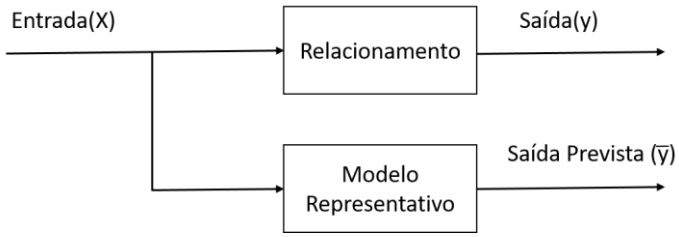

Figura 2-3: Representação conceitual do modelo representativo. A construção do modelo representativo do relacionamento da companhia com o cliente objetiva a previsão de uma variável de saída resultado a partir de um conjunto de dados de entrada.

Nesse contexto, a mineração de dados consiste no processo de construção de um modelo representativo que se ajusta aos dados observados. Esse modelo é usado para dois propósitos principais. De um lado, ele realiza a previsão de saída (taxa de juros) com base nas variáveis de entrada e, por outro lado, ele permite o entendimento do relacionamento entre a variável de saída e as variáveis de entrada. Por exemplo, a renda mensal do usuário, de fato, importa para a definição da taxa de juros do empréstimo? A renda pessoal importa mais do que a avaliação de crédito do usuário? O que acontece quando a renda do usuário dobra ou a avaliação de crédito do usuário cai 10 pontos?

2.2.2 Mineração de Dados Orientada ao Domínio

Embora muitos algoritmos e técnicas para mineração de dados tenham sido propostos nas últimas décadas, eles se concentram em técnicas independentes de domínio ou em problemas de domínio muito específicos. Como técnica independente do domínio, a mineração de dados é normalmente vista como um processo autônomo, de tentativa e erro, orientado a dados. E como técnica para resolução de problemas muito específicos, a mineração de dados é tida como um processo de análise de problema de negócios de maneira isolada, caso a caso. O resultado prático disso é que a descoberta de conhecimento geralmente não provê informações úteis e acionáveis para problemas de negócios reais (Cao 2008).

A realidade é que existe uma grande diferença entre os objetivos acadêmicos e as metas de negócios, assim como os resultados acadêmicos e as expectativas de negócios. Os pesquisadores em geral estão interessados na identificação de padrões novos e diferentes, enquanto os profissionais da indústria estão preocupados em resolver um problema prático.

A abordagem orientada ao domínio de aplicação (D3M) surgiu para tornar a mineração de dados viável no apoio a ações de tomada de decisão sobre problemas do mundo real nas empresas. A D3M busca a mudança de paradigma da descoberta de conhecimento orientada a dados para a descoberta de conhecimento acionável orientado ao domínio. Por conhecimento acionável entende-se aquele que pode ser usado na prática para tomada de decisões.

A ideia básica de D3M consiste em unir a mineração de dados tradicional ao conhecimento oriundo dos especialistas do domínio. Esses aspectos da metodologia D3M são relevantes ao problema de previsão de abandono por considerar as especificidades de cada domínio. O abandono nas indústrias de telecomunicações e varejo, por exemplo apresenta desafios distintos. A abordagem para tratar esse problema, nessas duas áreas, deve considerar as especificidades correspondentes para construção com sucesso de modelos representativos para aplicações preditivas (Dhamanwar & Murab 2016).

2.2.3 Behavior Scoring

A abordagem de *Behavior Scoring* (BS) tornou-se popular no domínio de instituições financeiras e também no varejo para a avaliação de novos clientes e previsão do comportamento de compra futuro de clientes já existentes. A aplicação de BS nesse domínio específico recebe os nomes de *Credit Behavior Scoring* e auxilia instituições financeiras na tomada de decisão sobre a concessão de crédito aos consumidores com base no risco de crédito de suas solicitações.

No contexto de mineração de dados, essa abordagem representa a solução para um problema de decisão binária, que é um problema de classificação em que a variável resposta é dicotômica, ou seja, corresponde a uma variável que possui apenas duas classes. O objetivo dessa abordagem é atribuir uma pontuação "score" que permita tomar uma decisão binária sobre o quão próximo o consumidor está de dois grupos: "bom" que é provável cumprir com suas obrigações financeiras ou um grupo de "mau", cujo pedido deve ser negado devido à sua alta probabilidade de faltar com seus compromissos na instituição financeira.

Em geral, o “score” é uma medida numérica contínua sobre a qual aplica-se um processo de discretização para valores binários com base na definição de um limiar, ou limite, para separação dos consumidores nos dois grupos desejados: “bons” e “maus” pagadores. A partir desse limite, as empresas definem políticas específicas para cada um dos grupos em relação as práticas de atendimento e satisfação do cliente para criação de barreiras de saída (ver Figura 2-4).

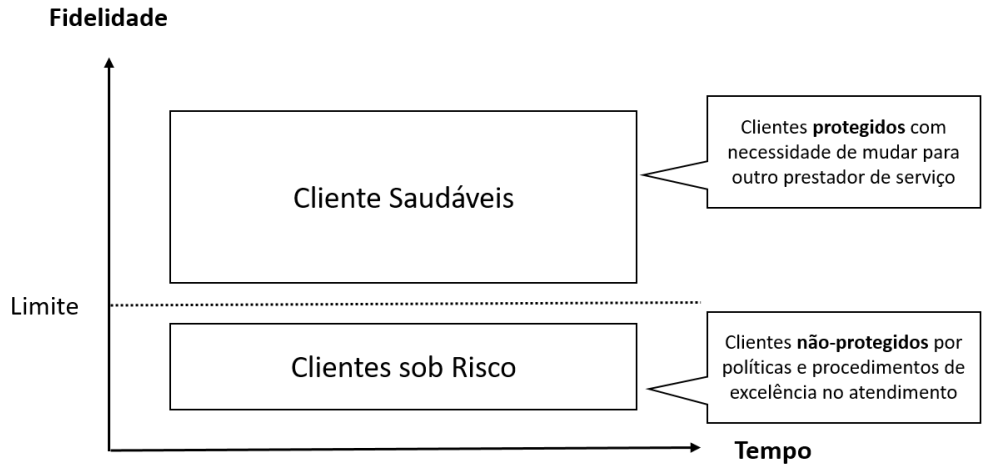

Figura 2-4: Ilustração do efeito do desenvolvimento de políticas e procedimentos de excelência no atendimento e satisfação do cliente para o desenvolvimento de barreiras de saída (abandono).

A construção de modelo representativo com uso de BS em geral utiliza uma abordagem pragmática e orientada ao domínio. Na área de crédito, por exemplo, a variável resposta é normalmente calculada entre 6 e 24 meses a partir da data de concessão de crédito. Em geral, um consumidor é considerado "mau" se possuir um atraso de 60 dias ou mais em uma parcela e considerado "bom" caso contrário.

Credit Behavior Scoring é utilizado quando um consumidor, que já possui histórico de transações na base de dados da instituição, está solicitando crédito. Neste caso, além das informações demográficas, informações comportamentais também são levadas em

consideração: o histórico de pagamentos em dia, em atraso, quantidade de empréstimos, entre outras informações.

O modelo de *Credit Behavior Scoring*, usado como uma ferramenta automática, fornece informação instantânea ao analista e aumenta a eficiência do analista de crédito. Os pontos fortes desse modelo são a precisão e a eficácia. Nas indústrias que utilizam BS, como crédito e telecomunicações, os modelos preditivos de melhor performance usam dados comportamentais, além de dados cadastrais, para construção dos respectivos classificadores. A saída de um modelo de *Credit Behavior Scoring* é interpretada como a propensão do cliente honrar sua dívida junto à instituição, ou seja, ser um bom cliente.

Para Kennedy e colegas (Kennedy et al. 2013/3), a primeira etapa do processo de BS corresponde à seleção de uma amostra de clientes, garantido que os dados referentes aos seus produtos e consumos estejam disponíveis em uma determinada data, chamada de ponto de observação. O período antes do ponto de observação é chamado de *janela de desempenho*. É preciso destacar que o significado do termo *desempenho* aqui se refere ao domínio de aplicação, não à qualidade dos modelos decisórios, como habitualmente usado pela área de Inteligência Artificial. Os dados contidos na janela de desempenho são estruturados em atributos que serão usados como entrada para o modelo de Behavior Scoring. Exemplos de variáveis criadas nesta janela são: máximo dias de atraso, quantidade de parcelas pagas em dia, número de ofertas recebidas, entre outras (McNab & Wynn 2000). A Figura 2-5 ilustra como os dados são particionados de acordo com a temporalidade.

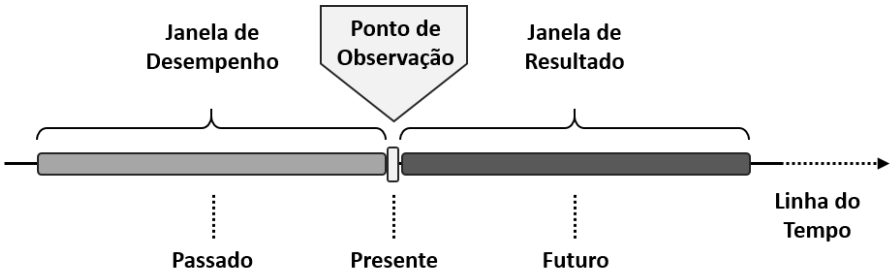

Figura 2-5: Particionamento dos dados na metodologia de Behavior Scoring adaptada de (Kennedy et al. 2013/3).

O período após o ponto de observação é chamado de *janela de resultado*. Os dados contidos na janela de resultado são estruturados em atributos que serão utilizados para avaliar a precisão do modelo. É nesta janela que a variável resposta binária ("bom" e "mau") é construída. A construção de variáveis em modelos de BS deve levar em consideração esta segmentação temporal. A Figura 2-6 ilustra o particionamento dos dados, em uma base de dados relacional, seguindo os conceitos de janelas de desempenho e resultado para uma data de

referência (ponto de observação) igual a 15/03/2015 para informações relacionadas às compras de um cliente.

cod_compra	cod_cliente	data	Valor
1	1	20/02/2015	R$ 100,00
2	1	03/03/2015	R$ 85,00
3	1	14/03/2015	R$ 111,00
4	1	16/03/2015	R$ 22,00
5	1	21/03/2015	R$ 50,00
6	1	01/04/2015	R$ 160,00

Figura 2-6: Exemplificação do particionamento de dados em função do tempo em uma base de dados relacional.

A forma de particionamento dos dados em Behavior Scoring é direcionada à criação do modelo de representação a partir de técnica de classificação via aprendizagem supervisionada. Nesse tipo de aprendizado, cada exemplo de treinamento é descrito por um conjunto de atributos que servem como dados de entrada (geralmente sob a forma de um vetor) que são associados a um valor de saída (também chamado de sinal de controle). A partir de um conjunto de entradas e saídas, o algoritmo pode produzir uma função de inferência capaz de gerar uma saída adequada a partir de uma nova entrada.

Na abordagem de Behavior Scoring, os dados de entrada representam os dados coletados anteriores ao ponto de observação (janela de desempenho ou performance) enquanto os dados posteriores (janela de resultado) são utilizados para definição da variável de saída. No cenário de previsão de abandono, as variáveis de entrada correspondem à informação sobre o comportamento do usuário em seu relacionamento com a empresa, como informações de pagamento, situação do contrato, histórico de uso do serviço, tipo de serviço utilizado, tempo desde a renovação do contrato, dentre outros aspectos.

A variável de saída, considerando o mesmo exemplo, indica a situação do cliente com relação ao abandono. Os dados após a data limite, ou ponto de observação, são utilizados para determinar a situação do usuário. Essa variável, em geral, é binária e pode assumir os valores *churner* em caso de abandono, ou *non-churner*, caso contrário. É comum adotar a regra em que o usuário é considerado churner caso não tenha mais dados de relacionamento após o ponto de observação. Esse cenário é ilustrado na Figura 2-7.

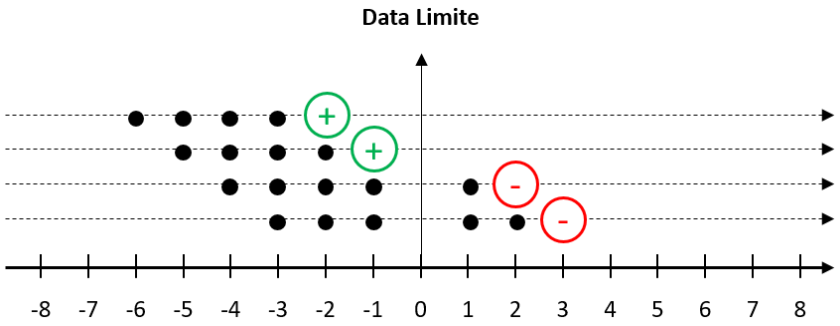

Figura 2-7: Representação da modelagem do problema de previsão de abandono. O eixo X representa o aspecto temporal, em dias, tendo o ponto 0 como a data limite, ou ponto de observação. As linhas representam o histórico de comportamento de clientes com a empresa e os pontos indicam a existência de uma ação relativa ao relacionamento como uma compra, pagamento ou ligação para o SAC. A indicação (+) representa a variável de saída do cliente e a sua indicação como churner. O sinal (-) indica a classificação do cliente como non-churner.

A janela é configurável com relação ao tamanho e disposição. Esses aspectos influenciam na forma de extração das informações dos usuários. Nós apresentamos esses conceitos abaixo.

2.2.3.1 Tamanho da Janela

Essa variável indicativa da duração da janela é utilizada para adaptação da técnica de BS em diferentes domínios. Em domínios como Telecomunicações e Crédito é comum encontrar janelas de performance com tamanho de 12 e 24 meses. Por outro lado, em jogos móveis, a duração das janelas varia entre 7 dias e 5 meses. O tamanho da janela influencia na quantidade de informação a ser extraída de cada um dos usuários. Além do tamanho, as janelas podem variar de acordo com a disposição.

2.2.3.2 Disposição da Janela

Essa variável indica como os dados serão particionados dentro da janela de performance para sua transformação e preparação dos atributos (variáveis independentes) a serem usados no treinamento do classificador. Há três principais configurações possíveis para a disposição da janela: Única, Múltipla Superposta (ou simplesmente “Superposta”) e Múltipla Disjunta (ou simplesmente “Disjunta”). Para efeito ilustrativo, a representação gráfica dessas alternativas é apresentada na Figura 2-8 usando como exemplo uma janela de tamanho 8 meses.

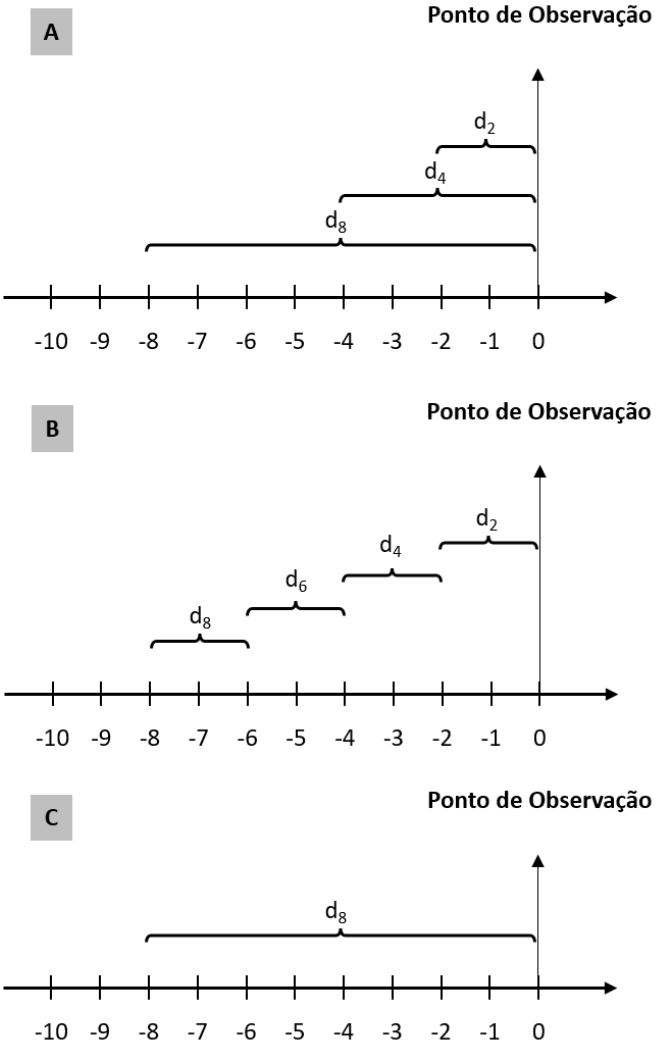

Figura 2-8: Tipos de disposição de janelas de performance. A) Janela Superposta. B) Janela Disjunta e C) Janela Única.

Abaixo descrevemos em maiores detalhes cada uma dessas configurações possíveis de janelas com relação à disposição.

- **Janela Superposta**

Nessa disposição de janela os dados são subdivididos em partições menores com tamanhos pertencentes a uma progressão geométrica. No exemplo da Figura 2-8(a), a janela de 16 meses é dividida em partições de tamanhos 2, 4 e 8 meses. Essas partições são superpostas tendo início no ponto de observação. Os atributos estipulados para treinamento do classificador, como por exemplo a frequência de uso, são construídos tendo como base essas partições. Para efeito de ilustração, consideremos que a frequência de uso será chamada de $sup\_freq\_d_x$ onde o termo d_x indica a partição de tamanho x meses. A frequência de uso é calculada para todas as partições com a produção dos atributos: $sup\_freq\_d_2$, $sup\_freq\_d_4$ e $sup\_freq\_d_8$. Todos

esses atributos são usados no treinamento do classificador para a janela com configuração de 8 meses.

- **Janela Disjunta**

Nessa disposição de janela os dados também são subdivididos em partições, porém com tamanhos pertencentes a uma progressão aritmética. No exemplo da Figura 2-8(b), a janela de 8 meses é dividida em partições de tamanhos 2, 4, 6 e 8 meses. Essas partições são disjuntas com a partição de menor tamanho tendo início no ponto de observação e as demais dispostas em sequência. Também para efeito de ilustração, consideremos que a frequência de uso será chamada de $disj\_freq\_d_x$ onde o termo d_x indica a partição de tamanho x meses. A frequência de uso é calculada para todas as partições com a produção dos atributos: $disj\_freq\_d_2$, $disj\_freq\_d_4$, $disj\_freq\_d_6$ e $disj\_freq\_d_8$. Todos esses atributos são usados no treinamento do classificador para a janela com configuração de 8 meses.

- **Janela Única**

Na janela Única os dados são divididos em uma única partição de tamanho igual ao tamanho da janela de performance. No exemplo da Figura 2-8(c), a janela única de 8 meses é usada para extração dos dados dentro desse período para construção da janela de dados. Com base no exemplo da frequência de uso, ilustrada nas disposições acima, a janela única produz somente um atributo $unica\_freq\_d_8$. É válido notar que $unica\_freq_{d_2} = sup\_freq\_d_8$.

Essas diferentes disposições particionam os dados da janela de performance de maneira diferente com impacto na forma de apresentação das informações a serem usadas na construção do modelo preditivo. A disposição Única é simples e rápida de ser produzida, porém despreza informações mais recentes dos usuários referentes aos últimos meses. A janela Superposta, por outro lado, extrai dados do comportamento mais recente do usuário e permite avaliar a evolução da relação. A janela Disjunta permite realizar a avaliação da evolução, porém por meio de uma abordagem mais próxima da análise de Domínio Temporal (Castro & Tsuzuki 2015).

A configuração desses aspectos relativos ao tamanho e disposição da janela de performance, assim como da janela de observação, é uma etapa importante do processo de construção do modelo preditivo. É também relevante a definição dos aspectos comuns às abordagens D3M como o entendimento dos dados, seleção dos tipos de dados a serem usados como variáveis de entrada, determinação das técnicas de modelagem a serem utilizadas para construção do classificador. A seção seguinte apresenta os detalhes envolvidos na construção

dos modelos em conjunto com revisão do estado da arte na indústria de maneira geral - sem considerar ainda jogos para dispositivos móveis.

2.3 A Construção de Modelos Preditivos e a Previsão de Abandono

A concepção de modelos inteligentes para previsão de abandono de usuários em domínios específicos considerando as abordagens D3M e BS pode ser alcançada por meio do uso de diferentes métodos para mineração de dados. O método CRISP-DM (Wirth 2000) é baseado nas principais abordagens existentes (KDD e SEMMA) e tornou-se o método padrão sendo amplamente usado em projeto acadêmicos e comerciais (Marbán et al 2009).

A metodologia CRISP-DM consiste em uma abreviação de *CRoss Industry Standard Process for Data Mining* que pode ser traduzida como Processo Padrão Inter-Indústrias para Mineração de Dados. É um modelo de processo de mineração de dados que descreve abordagens comumente usadas por especialistas em mineração de dados para atacar problemas. A CRISP-DM é composta de 6 fases executadas na ordem descrição na Figura 2-9.

Essa metodologia é idealizada como uma sequência de etapas organizadas em ordem. Na prática, as etapas podem ser executadas em uma ordem diferente. É comum a necessidade de voltar a uma etapa anterior para repetir determinadas ações. O modelo, portanto, não se propõe a capturar todas as possíveis rotas do processo de mineração de dados. As fases são descritas em maiores detalhes a seguir.

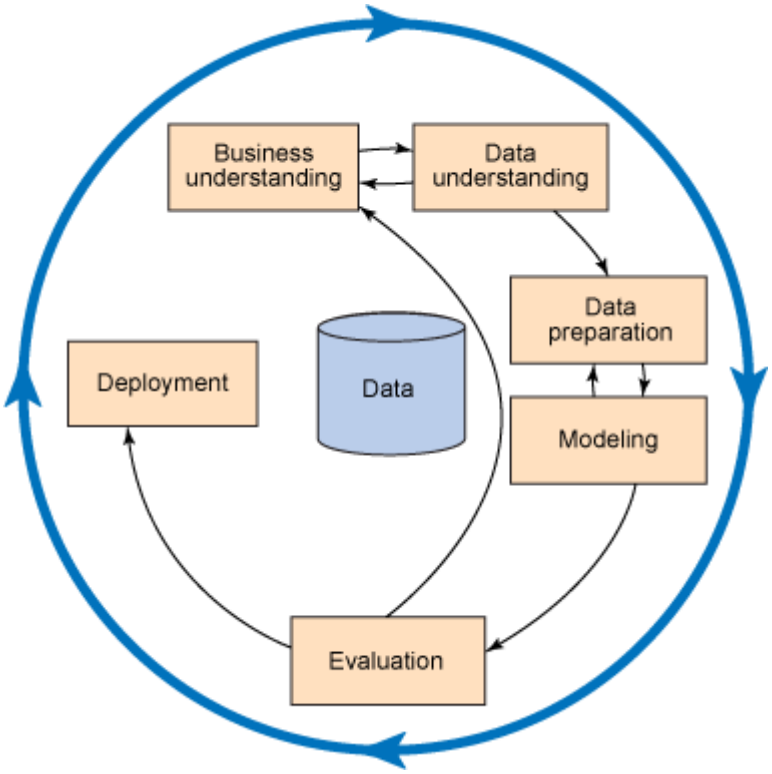

Figura 2-9: Metodologia CRISP-DM para projetos de mineração de dados.

2.3.1 Etapa 1: Entendimento do Negócio

Essa etapa inicial do processo concentra-se na compreensão dos objetivos e requisitos do projeto a partir da perspectiva empresarial e, em seguida, converter este conhecimento em uma definição de problema de mineração de dados e um plano preliminar concebido para atingir os objetivos almejados. Essa etapa inclui, dentre outras atividades, a realização das seguintes ações:

- Determinação dos objetivos de negócios
- Avaliação da situação
- Determinação dos objetivos de mineração de dados
- Produção do plano de projeto

O objetivo dessa etapa consiste em descobrir fatores relevantes que podem influenciar no resultado do projeto de mineração de dados. O negligenciamento dessa etapa pode implicar na alocação de esforço na produção das respostas corretas para as perguntas erradas. Os objetivos de negócios relacionados à retenção do usuário, segundo revisão da literatura proposta por (Ngai et al. 2009), estão detalhados abaixo:

- **Gestão de reclamação:** A gestão de reclamação é um elemento essencial em estratégias de negócios de sucesso para gestão das necessidades dos consumidores e mudanças em seus hábitos de consumo (Larivière & Van den Poel 2005; Bae et al. 2005/1).
- **Programa de fidelidade:** Essa atividade envolve campanhas e ações de marketing com o objetivo de auxiliar na manutenção de um relacionamento de longo prazo com o usuário. A análise de abandono além dos serviços de qualidade no atendimento e satisfação do cliente formam os programas de fidelidade (Cox 2002; Douglas et al. 2005; Larivière & Van den Poel 2005/8; Lejeune 2001; Kim 2006/1; Buckinx & Van den Poel 2005; Chu et al. 2007).
- **Marketing *one-to-one*:** Essa atividade refere-se à realização de campanhas de marketing direcionadas aos usuários de acordo com suas necessidades e características individuais. Essa ação é suportada pela análise, detecção e previsão de mudanças no comportamento do cliente (Chen et al. 2005/5; Jiang & Tuzhilin 2006; Kim & Moon 2006).

2.3.2 Etapa 2: Entendimento dos Dados

Essa etapa inicia-se com a coleta inicial de dados e prossegue com as atividades para familiarização com os dados, identificação de problemas relativos à qualidade dos dados, realização de descobertas preliminares nos dados e ainda a detecção de subconjuntos de dados relevantes para formação de hipóteses sobre informações ocultas na base de dados. Essa etapa inclui, dentre outras atividades, a realização das seguintes ações:

- Coleta de dados iniciais
- Descrição dos dados
- Exploração dos dados
- Verificação da qualidade dos dados

A revisão da literatura apresenta uma seleção de diferentes tipos de dados usualmente aplicados à análise e previsão de abandono. Uma grande parcela dos estudos baseia a criação dos seus modelos preditivos em dados cadastrais do usuário (ex: sociodemográficos) e em informações comportamentais acerca do uso/consumo de serviços (ex: histórico de uso). Uma compilação dos dados usados em diversos trabalhos (Madden et al. 1999/7; Hwang et al. 2004/2; Au et al. 2003; Rosset et al. 2002; Ng & Liu 2000; Verhoef & Donkers 2001; Hsieh

2004; Chiang et al. 2003; Jenamani et al. 2003; Van Den Poel 2003; Hung et al. 2006; Neslin et al. 2004; Slater & Narver 2000; Zhao et al. 2005; Coussement & Van den Poel 2008/1; Nie et al. 2009; Wang et al. 2010; Larivière & Van den Poel 2005/8) é mostrada a seguir. É importante atentar para o fato de que a maioria dos trabalhos identificados é da área de telecomunicações ou crédito e, portanto, os tipos de dados são relacionados a essas áreas.

- Dados cadastrais: economia (renda familiar, gasto mensal e forma de pagamento), sociodemográficos (idade, sexo, ocupação, localização geográfica e número de residentes)
- Histórico de uso: meses de uso, acesso a e-mail, acesso via browser, uso de serviços (ligações, sms, internet móvel, etc.), registros de alterações de serviço, tempo de uso, consumo total, tipo de equipamento, tendências de uso
- Histórico de compra e interação com a empresa: histórico da conta, situação de pagamento, chamadas ao serviço de atendimento ao consumidor, tempo desde a renovação, número de reclamações, tempo desde a última reclamação e resposta às ações de marketing.
- Dados de instituições parceiras: Bureau de Crédito.

Um número relevante de autores (Chen et al. 2012; Ha et al. 2002/9; Hsieh 2004; Jonker et al. 2004/8; Liu & Shih 2005/3; MacKay 1996; Pfeifer & Carraway 2000; Van Den Poel 2003; Verhoef et al. 2003/3) utilizaram um mesmo subconjunto de dados, amplamente conhecidos como RFM (*Recency*, *Frequency* e *Monetary*), na tarefa de construção do modelo:

- Tempo decorrido desde a última compra
- Frequência de uso
- Consumo monetário realizado em um determinado período de tempo

2.3.3 Etapa 3: Preparação dos Dados

Essa etapa cobre todas as atividades para construção do conjunto de dados final, a serem usados para alimentar a etapa de modelagem, a partir dos dados brutos inicialmente coletados. As atividades para preparação dos dados em geral são realizadas inúmeras vezes e não seguem uma ordem pré-determinada. Essas atividades incluem a seleção da tabela, registros e atributos,

assim como a transformação e limpeza dos dados para uso na ferramenta de modelagem. Essa etapa inclui, dentre outras atividades, a realização das seguintes ações:

- Seleção dos dados
- Limpeza dos dados
- Construção dos dados
- Integração dos dados
- Formação dos dados

A forma de execução de cada uma dessas ações é diretamente dependente de vários aspectos relacionados ao domínio como objetivos de negócios, tipos de dados disponíveis, organização da base de dados da aplicação, dentre vários outros fatores. O resultado prático é que não existe uma fórmula padrão para realização dessas ações. Entretanto, a aplicação com sucesso da metodologia CRISP-DM em situações reais para suporte à decisão dependem mais da representação das variáveis de entrada, ou seja, da preparação dos dados, do que da seleção e especificação dos parâmetros da técnica (Zhao et al. 2009/3).

A representação das variáveis é resultado da aplicação das atividades previstas nessa etapa referente à seleção, pré-processamento e transformação dos dados. A execução dessa etapa é custosa e, em geral, consome entre 50 e 80% do tempo total do projeto e normalmente é dependente do domínio da aplicação (Hall et al. 2011). Além disso, as atividades dessas etapas exigem uma conjugação de competências em banco de dados (extração, integração, consultas), estatística (análise de dados) e análise de sistemas (entendimento do domínio de aplicação).

A preparação dos dados é realizada a partir da configuração da janela de tempo relativa à abordagem de Behavior Scoring. A transformação e construção das variáveis independentes é realizada com base nos dados dentro da janela de performance. Na indústria de telecomunicações essa janela, em geral, é configurada como estática e com duração entre 6 a 24 meses (Jeannt 2016; Hsieh 2004; Castanedo et al. 2014; Ruta et al. 2009; Kennedy et al. 2013). Os valores múltiplos de 12 meses são normalmente utilizados por permitir a inclusão de informação independentemente de fatores sazonais (Chen et al. 2012).

Por outro lado, há pesquisas nessa indústria que utilizam janelas mais curtas para verificar a possibilidade de construção de modelos preditivos a partir de uma quantidade menor de dados. Nas pesquisas conduzidas por (Zhao et al. 2005; Neslin et al. 2004) os dados foram compilados com base em uma janela de 3 meses. Já na pesquisa conduzida por (Hung et al. 2006), também na indústria de telecomunicações, os autores analisam a performance de

classificadores construídos com tamanhos de janelas de 1, 2 e 3 meses. A janela mais curta de 1 mês apresentou os melhores resultados, com base no *Lift*. O *Lift* é uma medida de desempenho do modelo calculada como a razão entre a confiança da classificação e a confiança da amostra de dados (Victor 2002).

A transformação e construção das variáveis dependentes referente à situação do usuário (*churner* ou *non churner*) é realizada com base nos dados dentro da janela de performance. Na indústria de telecomunicações, o cancelamento do contrato e encerramento da prestação de serviço define a situação do usuário. Essa janela em geral assume a duração aproximada de 3 meses (Neslin et al. 2004; Zhao et al. 2005) tendo casos entretanto de janelas de resultado com 6 meses (Chen et al. 2012) e até 12 meses de duração (Madden et al. 1999).

2.3.4 Etapa 4: Modelagem

Nessa etapa as técnicas de modelagem do problema são selecionadas e aplicadas, e seus parâmetros são calibrados para atingir os valores ótimos. Dado que a previsão de abandono consiste essencialmente em um problema de classificação binária (*churner* e *non-churner*), o número de técnicas de classificação binária sugeridas pela academia é bastante numerosa (Kennedy et al. 2013).

As técnicas em geral apresentam requisitos específicos com relação ao formato dos dados e, portanto, muitas vezes é necessário retornar à fase de preparação dos dados. A técnica de regressão logística, por exemplo, não suporta atributos nominais, enquanto redes neurais (MLP) e árvores de decisão provêm suporte a este formato de dado. Essa etapa inclui, dentre outras atividades, a realização das seguintes ações:

- Seleção das técnicas de modelagem
- Geração de projeto de teste
- Construção do modelo
- Avaliação do modelo

A revisão da literatura apresenta uma seleção de diferentes técnicas e algoritmos de modelagem usualmente aplicados à análise e previsão de abandono. As técnicas mais populares na gestão de relacionamento com o usuário, segundo revisão da literatura realizada por Ngai e colegas (Ngai et al. 2009), estão detalhados abaixo. Nessa revisão foram considerados 87

artigos dentro do período de 2000 a 2006. No total foram identificadas o uso de 34 técnicas distribuídas da forma abaixo.

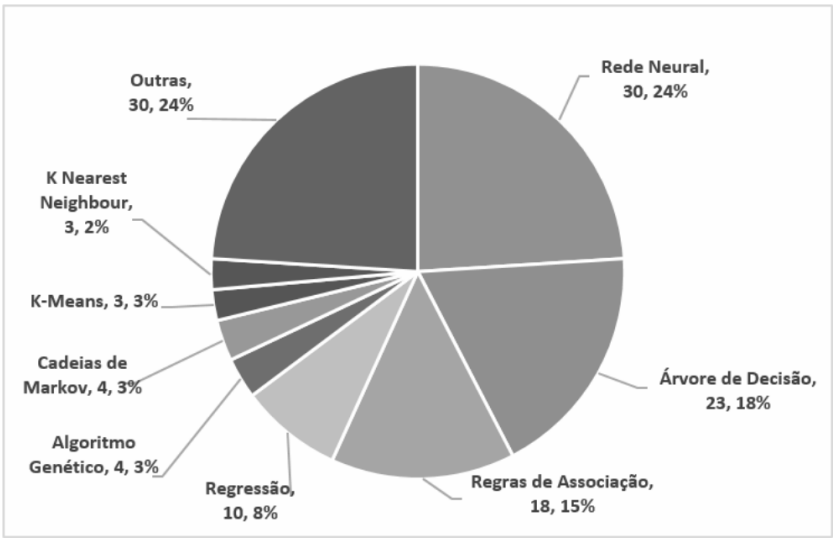

Figura 2-10: Distribuição dos artigos segundo a técnica de mineração adotada (Ngai et al. 2009/3).

Especificamente com relação à retenção de usuários e previsão de abandonos, as principais técnicas adotadas nas pesquisas são detalhadas a seguir.

- Redes Neurais: (Datta et al. 2000; Kim 2006/1; Buckinx & Van den Poel 2005; Koh & Gerry 2002; Mozer et al. 2000; Smith et al. 2000; Kuo et al. 2005; Chang et al. 2006)
- Árvores de Decisão: (Cox 2002; Douglas et al. 2005; Larivière & Van den Poel 2005; Datta et al. 2000; Hung et al. 2006; Koh & Gerry 2002; Mozer et al. 2000; Smith et al. 2000; Chu et al. 2007; Kim et al. 2005)
- Regras de Associação: (Adomavicius & Tuzhilin 2001; Au & Chan 2003; Chen et al. 2005; Demiriz 2004; Jiao et al. 2006; Lee et al. 2001; Wang et al. 2004; Hsieh 2004; Ha et al. 2002; Ha 2006; Liu & Shih 2005; Liao & Chen 2004)
- Regressão: (Kim 2006; Buckinx & Van den Poel 2005; Koh & Gerry 2002; Mozer et al. 2000; Smith et al. 2000; Cassab & MacLachlan 2006; Van den Poel & Buckinx 2005)

2.3.5 Etapa 5: Avaliação

Nesta etapa do projeto o modelo preditivo já foi construído e aparenta apresentar alta qualidade como previsto na atividade de avaliação do modelo. Antes de prosseguir com a implantação final do modelo em ambiente empresarial é importante avaliar mais

detalhadamente o modelo e rever as etapas executadas na sua construção para ter certeza de que ele atinge adequadamente os objetivos de negócios.

Um dos principais objetivos dessa fase é determinar se existe alguma questão comercial importante que não tenha sido suficientemente considerada. No final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada. Essa etapa inclui, dentre outras atividades, a realização das seguintes ações:

- Avaliação dos resultados
- Revisão do processo
- Determinação dos próximos passos

A partir da revisão da literatura, nós identificamos a ausência de padrão com relação à avaliação da performance dado o uso de diferentes métricas nos artigos avaliados. Essa diversidade de métricas, além do uso de diferentes bases de dados, torna complexa a comparação de performance entre trabalhos distintos. As métricas mais utilizadas são detalhadas abaixo.

- Uma parcela significativa dos trabalhos pesquisados utiliza um único tipo de métrica para avaliação da performance de classificação dos modelos preditivos.
    - Precisão (acurácia): (Chang et al. 2006; Kim 2006; Chu et al. 2007; Smith et al. 2000; Liao & Chen 2004; Demiriz 2004; Verhoef & Donkers 2001) .
    - Suporte e Confiança (regras de associação): (Chen et al. 2005; Au & Chan 2003; Adomavicius & Tuzhilin 2001; Chiang et al. 2003; Hsieh 2004; Ng & Liu 2000; Liu & Shih 2005).
    - Lift: (Datta et al. 2000; Cox 2002; Hung et al. 2006)
    - Precision / recall / f1-metric: (Liu & Shih 2005/3; Wei & Chiu 2002)

- A maioria das pesquisas consideradas na revisão bibliográfica aplicam dois ou mais métodos de avaliação para medição e comparação da performance de classificadores.
    - Accuracy e AUC / ROC: (Van den Poel & Buckinx 2005; Buckinx & Van den Poel 2005; Coussement & Van den Poel 2008; Larivière & Van den Poel 2005)

- Accuracy, AUC / ROC e Lift: (Mozer et al. 2000; Chen et al. 2012; Van Den Poel 2003; Wang et al. 2010)

A métrica de precisão, ou taxa de acerto, mede a proporção de casos classificados corretamente. Essa métrica tem sido considerada como um critério significativo na avaliação da capacidade de classificação dos modelos preditivos e, de fato, corresponde ao avaliador de desempenho mais comumente utilizado pelos autores. A revisão realizada por nós, e também por Ngai e colegas (Ngai et al. 2009), apresenta uma grande proporção de artigos que utilizam essa métrica. Em outra revisão (Marques et al. 2012), mais de 60% dos trabalhos avaliados utilizam essa métrica como principal indicador da performance. Essa métrica, entretanto, não é a mais apropriada para problemas de Behavior Scoring devido a dois principais fatores:

- Dados desbalanceados: De acordo com evidências empíricas e teóricas, o uso da métrica precisão é fortemente tendencioso em relação ao desequilíbrio na distribuição de classes (Fawcett & Provost 1997). Em aplicações reais, a classe de clientes considerados ruins geralmente está sub-representada em comparação com a classe de clientes bons.
- Custo de classificação incorreta: Os custos associados aos erros do tipo II (clientes ruins classificados incorretamente como bons) são muito mais elevados do que os erros do tipo I (clientes bons classificados como ruins) (Baesens et al. 2003; West 2000).

O trabalho de Marques e colegas (Marques et al. 2012) revela que mesmo com esses inconvenientes relativos ao uso da métrica de precisão, somente cerca de 80% dos trabalhos relatam as suas taxas de erros do tipo I e II e apenas dois artigos (Abdou 2009; Oreski et al. 2012) avaliam o impacto do custo de classificação incorreta. E mesmo nesses casos, a estimativa de custo de classificação incorreta geralmente está incorreta dada a dificuldade em se obter estimativas de custo confiáveis.

As métricas Receiver Operating Characteristic (ROC), Área sob a Curva ROC e teste de Kolmogorov-Smirnov são robustas com relação a esses problemas e permitem ainda a visualização e comparação da performance de classificadores. Apesar da sua utilidade, essas métricas ainda são pouco utilizadas nos trabalhos acadêmicos identificados em nossa revisão. E mesmo na pesquisa conduzida por Marques e colegas (Marques et al. 2012), somente 3 dos 54 trabalhos avaliados no período de 2000 a 2012 utilizam essas métricas.

2.3.6 Etapa 6: Implantação

A realização de todas as etapas anteriores para criação e avaliação do modelo não implica no fim do projeto. Mesmo nos casos em que o propósito do modelo é aumentar o conhecimento sobre os dados, esse conhecimento adquirido precisa ser organizado e apresentado de uma forma que seja útil ao cliente para auxiliar na tomada de decisão. A depender dos requisitos de negócios, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto integrar um processo de mineração de dados em uma empresa. Essa etapa inclui, dentre outras atividades, a realização das seguintes ações:

- Plano de implantação
- Plano de monitoramento e manutenção
- Produção do relatório final
- Revisão do projeto

Em muitos casos, o cliente, e não o analista de dados ou o pesquisador, será o responsável pela execução das atividades de implantação. Mesmo que o analista implemente o modelo, é importante que o cliente compreenda as ações que precisam ser realizadas para usar corretamente os modelos criados. E justamente por esse motivo, há poucas referências com relação a essa etapa.

2.4 As Diretrizes em Domínios Estabelecidos

A avaliação dos fundamentos da previsão de abandono e análise do estado da arte permitiu identificar que em domínios mais estabelecidos os principais desafios são conhecidos, assim como as soluções adotadas em sua resolução. A área de telecomunicações é uma dessas áreas estabelecidas com um grande número de pesquisas realizadas dentro das últimas duas décadas (Ngai et al. 2009; Hashmi et al. 2013). Esse amplo conjunto de pesquisas exploram diferentes alternativas para as especificidades e desafios relativos ao domínio e criam uma série de diretrizes que facilitam o reuso de boas práticas para resolução de problemas comuns à indústria de telecomunicações. Nesta seção nós analisamos mais a fundo esses aspectos, conceituando o que entendemos por diretriz e dando exemplos de diretrizes identificadas na indústria de telecomunicações.

2.4.1 Definição do Conceito de Diretriz

Segundo a definição encontrada em (Ferreira 1994), uma diretriz é "conjunto de instruções ou indicações para se tratar e levar a termo um plano, uma ação, um negócio". As diretrizes são orientações, recomendações e guias que servem para auxiliar e orientar as pessoas na realização de determinadas atividades. As diretrizes visam, de uma forma geral, apresentar um conjunto de orientações que contribuam para que haja o reuso de boas práticas analisadas e aprovadas em problemas recorrentes no desenvolvimento de aplicações, evitando desperdício e duplicação de trabalho. A Figura 2.10 apresenta, de forma simplificada, a relação semântica entre área de conhecimento, processo e diretrizes.

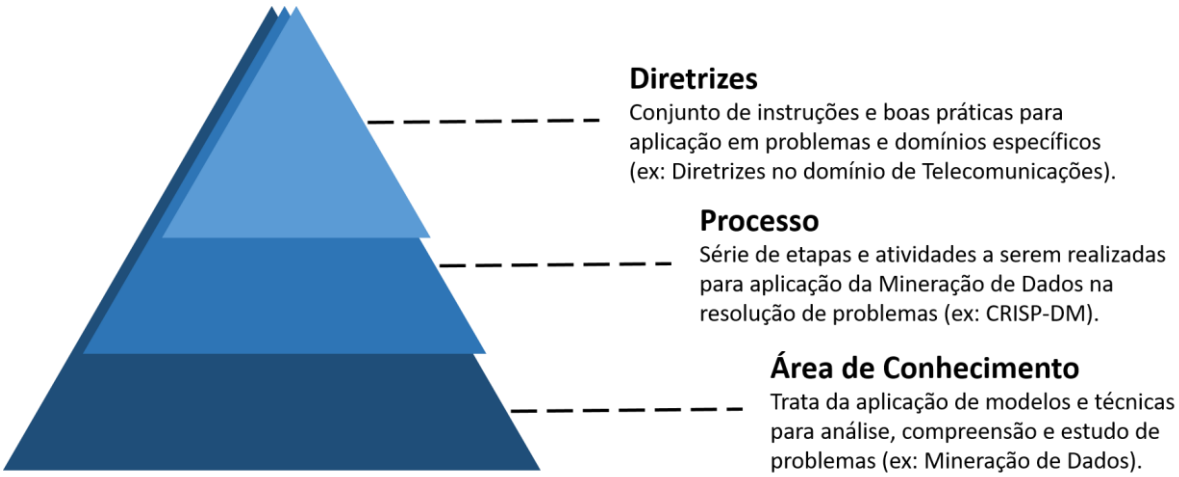

Figura 2-11: Relação semântica entre área de conhecimento, processo e diretrizes.

As diretrizes podem ser aplicadas em diversos contextos. Por exemplo, há diretrizes que indicam como redigir uma boa redação, como fazer uma boa apresentação de projeto, dentre outras. Na área de computação, nós podemos citar uma série de diretrizes relativas às boas práticas de programação (Balena and Dimauro 2005), à modelagem de um domínio (Trujillo 2006) e à pesquisa empírica em engenharia de software (Kitchenham et al. 2002).

No contexto de mineração de dados, nós identificamos trabalhos que objetivam lidar com os problemas de construção, avaliação e exploração de modelos de mineração de dados em domínios específicos como medicina clínica (Bellazzi & Zupan 2008), farmácia (Shen 2013) e redes de comunicação. Esses trabalhos avaliam as especificidades dos seus domínios de aplicação e propõem um conjunto de instruções e orientações para ajudar os pesquisadores e profissionais na seleção adequada de alternativas relativas à preparação dos dados, construção

dos classificadores e validação de modelos preditivos. No domínio de previsão de abandono, as diretrizes se remetem à:

1) Identificação das especificidades do domínio responsáveis pela apresentação de desafios à aplicação do processo “padrão” para construção do modelo preditivo. Essa etapa envolve responder à seguinte questão:

    - Quais as principais especificidades dessa indústria relevantes para a modelagem do problema?

2) Determinação das melhores práticas para tratamento dos problemas identificados e construção dos modelos preditivos da melhor maneira possível. Essa etapa envolve responder às seguintes questões:

    - Como essas especificidades podem ser tratadas no processo de construção do modelo preditivo?
    - Qual o peso dessas especificidades, e dos seus possíveis tratamentos, na performance da previsão?

3) Validação da aplicabilidade das melhores práticas em problemas similares. Essa etapa envolve responder à seguinte questão:

    - A solução proposta é replicável para outras bases dentro do mesmo domínio?

Em áreas de pesquisa com grande número de trabalhos publicados, os desafios relacionados à modelagem da previsão assim como as diretrizes para tratamento desses problemas em geral já são conhecidos. É justamente essas informações que iremos detalhar na próxima seção para a indústria de telecomunicações com base na revisão crítica do estado da arte.

2.4.2 As Especificidades e Diretrizes para Previsão de Abandono em Telecomunicações

A análise da indústria de telecomunicações e investigação dos trabalhos publicados na construção de modelos preditivos de abandono de clientes permitem identificar características a respeito do relacionamento estabelecido entre as empresas dessa área e seus consumidores. Essas características influenciam no tipo e qualidade dos dados existentes para realização da mineração de dados e treinamento de classificadores para previsão de abandono (Tabela 2-1).

Especificidades	**Impacto**	**Diretrizes**
Taxa de abandono: 20 - 40% ao ano; Ciclo de vida do Cliente: 2 - 4 anos	Configuração da Janela de Performance	Janela de performance estática com tamanho entre 6 a 12 meses.
Modelo de Relacionamento: Assinatura Taxa de Conversão: 100% Tipos de Dados: Fixo (independente da aplicação)	Dados para construção das variáveis independentes	Pessoais Comportamentais Monetários
Definição do rótulo: Fixo / Explícito (cancelamento da assinatura)	Configuração da Janela de Resultado	Janela de resultado com tamanho de 3 meses

Tabela 2-1: Análise das especificidades e diretrizes na aplicação de previsão de abandono no domínio de Telecomunicações.

De acordo com pesquisas de mercado, as taxas de abandono mensais na indústria de telecomunicações estão próximas a 2% (Berson and Smith 2002). A pesquisa conduzida por (Neslin et al. 2004) corrobora essa informação ao reportar taxas anuais de abandono entre 23,4% e 46%. Esses valores são similares seja para planos de telefonia pós-pagos ou pré-pagos. Essa característica da indústria implica em um tempo de vida do cliente entre 2 a 4 anos, em média. Esse valor de tempo de vida é considerado elevado quando comparado com outras indústrias como E-Commerce (Collins 2012) e Software as a Service (SaaS) (Consulting 2013) que apresentam taxas de abandono médias próximas a 10% e 5% ao mês, respectivamente.

A forma de relacionamento típica com usuário em planos pós-pagos consiste no contrato formal para assinatura do serviço com pagamentos mensais. A taxa de conversão dos usuários, ou seja a proporção de uantos usuários tornam-se pagantes, nesse caso, é de 100%. Essas características indicam, dentre outras coisas, que a empresa de telefonia possui informações cadastrais e pessoais do usuário necessárias ao estabelecimento do contrato formal.

A taxa de 100% de conversão implica ainda que todos os clientes estão com contrato de telefonia ativo e, consequentemente, esses consumidores usufruem dos serviços ofertados: ligações de curta e longa distância, envio de SMS, acesso à internet móvel, dentre outros. Todas essas informações relativas ao uso dos serviços são armazenadas para acompanhamento do consumo dos clientes e realização das devidas cobranças de acordo com as regras do contrato vigente. Até mesmo informações sobre interação com a empresa para atendimento, sugestão e reclamação são armazenadas nas bases de dados das empresas de telecom. Assim como os dados transacionais e monetários referentes aos pagamentos mensais (consumo, valores, data de vencimento, etc.) também são armazenadas em data warehouses.

As informações armazenadas, especialmente as comportamentais, são específicas do domínio de telecomunicações. Dados como quantidade de chamadas, duração total de ligações, frequência de envio de SMS, dentre outros, são comuns somente a empresas dessa indústria. Por outro lado, os serviços ofertados pelas empresas de telecomunicações são praticamente idênticos e direcionados aos mesmo tipos de usuários e, portanto, há pouca, ou nenhuma, diferenciação em relação aos tipos de dados armazenados. Em outras indústrias, por outro lado, os dados costumam ser dependentes da aplicação. É de se esperar, por exemplo, que um SaaS de armazenamento de dados (ex: Dropbox) e outro de comunicação (ex: Slack) produzam dados comportamentais completamente diferentes. Na própria indústria de jogos é de se imaginar que os dados também são extremamente dependentes da aplicação.

A definição da situação do usuário (rótulo) é uma das atividades da preparação de dados para permitir a realização do treinamento supervisionado do classificador. Na indústria de telecomunicações, o cancelamento da relação com o cliente na maioria dos casos acontece por cancelamento voluntário deliberado. O rótulo é definido com base em uma ação clara e objetiva para encerramento do contrato de prestação de serviço. A informação sobre o cancelamento é armazenada na base de dados junto às demais informações do cliente. Em outras indústrias como E-Commerce, em que não há assinatura de serviço envolvida, a definição do rótulo torna-se uma atividade complexa.

2.5 Conclusão e Observações

Este capítulo apresenta o contexto do problema de gestão do relacionamento com clientes e ressalta a relevância da previsão de abandono para a realização de ações preventivas

de retenção como forma de expandir o lucro das empresas. Resumidamente, as motivações para retenção de clientes são:

- O custo de retenção de clientes é menor do que o de aquisição de novos clientes;
- O custo da próxima compra é radicalmente menor do que o custo da primeira compra;
- O aumento do conhecimento sobre os clientes para ações personalizadas;
- A fidelização do cliente permite criar potenciais clientes defensores da marca para indicar e recomendar a marca, produtos e serviço da empresa para novos clientes.

A abordagem utilizada para a construção de modelos preditivos baseia-se na aplicação de mineração de dados por meio de aprendizagem supervisionada para classificação dos clientes de acordo com as chances de abandono. A revisão bibliográfica para avaliação da aplicação em indústrias estabelecidas, como crédito e telecomunicações, revela o uso das abordagens D3M e BS para tratamento das características únicas de cada domínio durante o processo de construção dos modelos preditivos.

Os domínios avaliados, em especial crédito e telecomunicações, apresentam configurações específicas similares com relação a determinados aspectos como tamanho das janelas de performance e resultado e também as regras para definição do rótulo. Por outro lado, os aspectos referentes à seleção dos dados, pré-processamento dos dados e seleção da técnica de mineração diferem na maioria dos trabalhos identificados. Não há um padrão, sistematização ou diretriz específica a ser seguida, o que ressalta a relevância da abordagem D3M para construção de modelo representativo específico para cada domínio.

A construção do modelo para previsão de abandono em jogos para dispositivos móveis, portanto, deve considerar as características específicas do relacionamento entre o usuário e as empresas de jogos dessa indústria para aplicação com sucesso das metodologias D3M e BS. Esse é o tema do próximo capítulo.

3 Previsão de Abandono em Jogos Móveis

A previsão de abandono de clientes é uma atividade dependente do domínio e de suas peculiaridades, como exemplificado nos domínios de telecomunicações e crédito anteriormente. De maneira similar, a indústria de jogos para dispositivos móveis também apresenta suas especificidades. Nesse capítulo, discutimos o problema do abandono de usuários no contexto da indústria de jogos móveis. Apresentamos também a revisão bibliográfica com os trabalhos identificados na área de previsão de abandono em jogos para dispositivos móveis.

3.1 A Indústria de Jogos Móveis

A indústria de jogos digitais mundial foi responsável por movimentar cerca de 99,6 bilhões de dólares em 2016 (Minotti 2016), com um crescimento médio global de 8,5% ao ano. Essa receita é proveniente do consumo realizado por cerca de 2,1 bilhões de usuários ao redor do mundo, ou 28,4% da população mundial. As previsões de mercado (Newzoo 2016) indicam ainda que a receita dessa indústria deve alcançar 118,6 bilhões de dólares em 2019 com um crescimento médio estimado de 6,6% ao ano (ver Figura 3-1).

Em termos absolutos, essa indústria é responsável por movimentar uma receita 2,5 vezes maior do que a bilheteria da indústria cinematográfica que faturou 38,3 bilhões de dólares em 2015, segundo dados de mercado (McClintock 2016). Ainda em termos absolutos, a indústria de jogos movimentou uma receita 6 vezes maior do que a indústria fonográfica (Ingham 2015; IFPI 2015) que contabilizou 15 bilhões de dólares em 2015 (IFPI 2016). Essas pesquisas de mercado indicam ainda que a indústria de filmes deve crescer a taxas de 4,3% ao ano e a fonográfica a taxas inferiores a 1% ao ano, ambas mais modestas que as projeções para indústria de jogos.

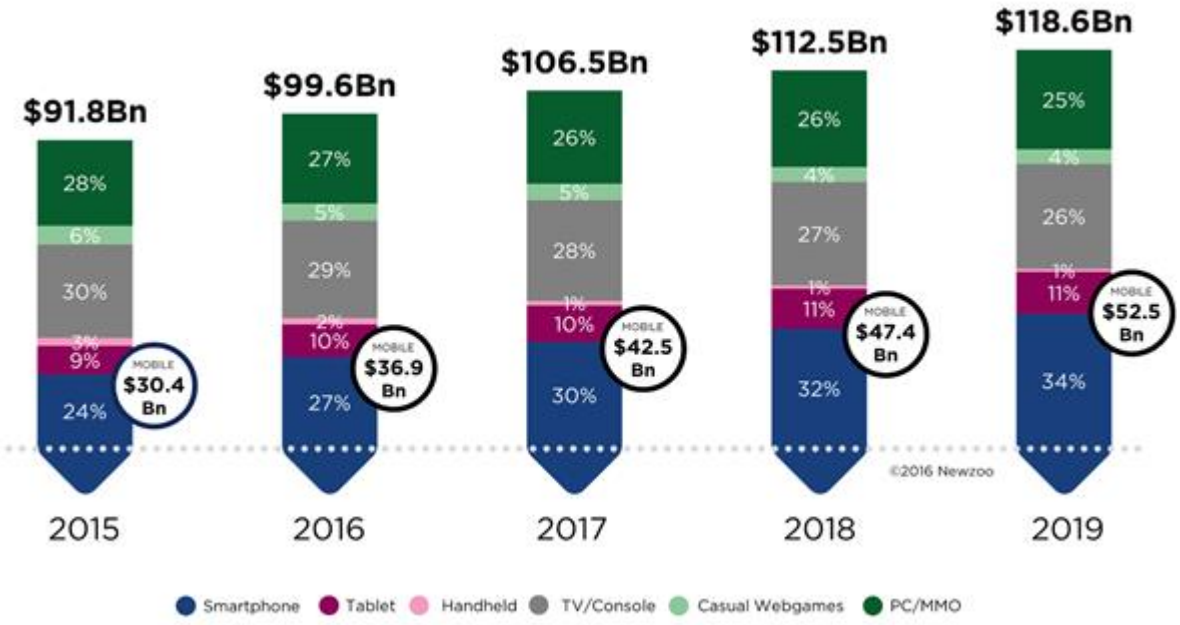

Figura 3-1: Faturamento da indústria mundial de jogos para os anos de 2015 e 2016 (real) e previsão de faturamento para os anos de 2017, 2018 e 2019 (Newzoo 2016). O faturamento apresentado está subdividido conforme s plataformas de distribuição e consumo dos jogos digitais – de cima para baixo estão PC/MMO, Casual Webgames, TV/Console, Handheld, Tablet e Smartphone.

A área de jogos móveis se destaca atualmente por ser a principal plataforma, em termos de faturamento total e expectativa de crescimento, da indústria global de jogos. Os jogos móveis são responsáveis por movimentar 38,7 bilhões de dólares em 2016, sendo responsável por 39% do faturamento da indústria global, com previsão de atingir 45% em 2019. Além disso, o mercado de jogos para smartphones e tablets deve crescer a taxas anuais de 23,7% e 15,1%, respectivamente, sendo as principais plataformas responsáveis pelo crescimento no faturamento global da indústria. Essa quantia é superior à receita advinda da comercialização de jogos para Computadores e Consoles (Takahashi 2016), individualmente.

O mercado de jogos móveis também é enorme ao ser comparado com os demais conteúdos disponíveis nas lojas de aplicativos, representando mais de 85% da receita total (Takahashi 2016).

3.2 Evolução dos Modelos de Negócios em Jogos Móveis

A origem dos jogos para dispositivos móveis está intimamente relacionada ao progresso da telefonia móvel, em termos de tecnologia de comunicação para transmissão de dados, assim como também da tecnologia embarcada nos aparelhos celulares. A evolução foi responsável pela criação da base tecnológica necessária para desenvolvimento da indústria de jogos móveis.

A expansão da distribuição digital dos jogos permitiu o desenvolvimento de diferentes modelos de negócios para comercialização de jogos móveis. Por motivos didáticos, o conteúdo está distribuído dentro das áreas de modelos de distribuição e modelos de negócios.

3.2.1 Modelos de Distribuição

No final dos anos 90, apareceram os primeiros telefones celulares para os quais um terceiro (não fabricante) poderia construir aplicativos. Nesse cenário de alta fragmentação, tanto em termos de sistema operacional quanto em configuração de hardware, a produção de jogos para dispositivos móveis era uma atividade extremamente complexa. Os desenvolvedores de jogos, para atingir uma parcela significativa dos aparelhos disponíveis, precisavam desenvolver e manter centenas, e até milhares, de versões diferentes dos seus jogos com código-fonte específicos para cada uma das plataformas. Além disso, era necessário testar cada uma dessas versões para garantir o correto funcionamento de todas as versões nos respectivos aparelhos-alvo.

Além do desafio em termos de produção de jogos, os desenvolvedores ainda enfrentavam um cenário extremamente competitivo para distribuição dos jogos em dispositivos móveis. Os principais modelos de distribuição adotados apresentavam várias barreiras para disponibilizar os jogos e alcançar o público consumidor final.

- **Distribuição pelas fabricantes**

Inicialmente a distribuição dos jogos foi realizada direto pelas fabricantes através da pré-instalação dos jogos nos aparelhos. Os jogos eram instalados nos celulares junto com o sistema operacional ainda durante o processo de fabricação. Do ponto de vista do consumidor, esse processo permitia acessar jogos logo após a aquisição dos aparelhos. Porém, os consumidores não podiam remover os jogos já previamente instalados e tampouco instalar novos jogos (ver Figura 3-2).

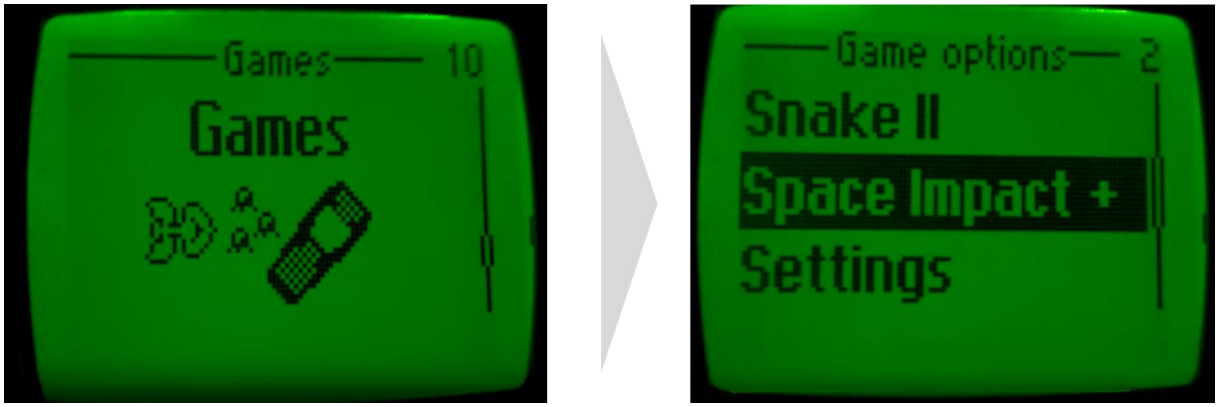

Figura 3-2: Menu de jogos no Nokia 1100 com espaço para somente 2 títulos (Snake e Space Impact).

Do ponto de vista do desenvolvedor, a pré-instalação através dos fabricantes representava o único caminho viável para conseguir alcançar os consumidores. Entretanto, os aparelhos apresentavam restrições em termos de memória para permitir a pré-instalação de uma grande quantidade de jogos. E dada a quantidade também limitada de fabricantes, a distribuição de jogos através desse modelo era algo acessível para um número extremamente restrito de desenvolvedores de jogos. Esse foi o formato de distribuição mais comum até por volta do ano 2000.

- **Distribuição pelas operadoras de telefonia**

Após o advento da tecnologia WAP para internet móvel, as operadoras de telefonia também iniciaram a distribuição de jogos digitais. A tecnologia originalmente criada para prover serviços como e-mail, notícias e download de música passou a também ser utilizado para o download e instalação de jogos e aplicativos.

Para o consumidor, a distribuição pela operadora não era de usabilidade fácil e não ganhou escala. Esses jogos eram comprados direto pelas operadoras e o custo incluído na conta de telefone. A baixa adoção da tecnologia pelos consumidores levou as operadoras a não realizar grandes investimentos para a criação de um vasto portfólio de jogos para os usuários. Aliado a isso, as operadoras não desejavam gerir a relação com uma grande quantidade de desenvolvedores para disponibilizar novos conteúdos para seus usuários. O resultado é que esse tipo de distribuição não vingou.

Para o desenvolvedor, a distribuição através das operadoras de telefonia representava uma alternativa ao monopólio das fabricantes. Esse canal de distribuição de conteúdo, em teoria sem restrição quanto ao limite de espaço físico, permitiria incluir uma quantidade maior de jogos e abrir o mercado para um número maior de desenvolvedores. Na prática, entretanto, essa previsão não se concretizou.

- **Sites de Conteúdo**

Praticamente em paralelo com a distribuição de conteúdo pelas operadoras surgiram os sites de conteúdo distribuídos através do protocolo de comunicação WAP, como o site GetJar.com. Esse serviço de distribuição enfrentou os mesmos problemas das operadoras. A baixa qualidade desse serviço, aliado ao alto custo cobrado pelas operadoras na conexão WAP, também inviabilizou esse modelo de distribuição.

- **Publicadores de Jogos**

Os publicadores de jogos surgiram para suprir a lacuna existente entre os desenvolvedores de jogos e os principais distribuidores de conteúdo: fabricantes e operadoras. As operadoras não estavam interessadas em lidar com potencialmente centenas de desenvolvedores para buscar conteúdo para popular as lojas de aplicativos WAP. Os publicadores de jogos emergiram, portanto, para intermediar esse relacionamento comercial ao custo de uma porcentagem da receita gerada pelo jogo.

Essa nova peça na curva de valor da indústria de jogos para dispositivos móveis foi responsável por reduzir ainda mais a receita dos desenvolvedores. A negociação com as operadoras e fabricantes em geral consistia em um modelo de compartilhamento de receita com somente 30% da receita sendo destinada para o desenvolvedor. Essa fração passou a ser dividida também com os publicadores.

- **Loja de Aplicativos**

O cenário mudou com o lançamento da loja de aplicativos do iPhone, chamada de iPhone App Store, inaugurada em julho de 2008 através de uma atualização do aplicativo iTunes. Essa loja permite à Apple controlar a qualidade do conteúdo distribuído e cobrar uma porcentagem da receita proveniente da comercialização dos aplicativos, incluindo jogos. Os usuários do iPhone podem baixar somente aplicativos distribuídos através dessa loja.

Ainda em 2008, em conjunto com o lançamento do sistema Android, o Google lançou também a sua loja de aplicativo chamada Android Market, e posteriormente renomeada para Google Play. A loja de aplicativos do Google possui funcionamento similar à loja da Apple com a disponibilização de jogos e aplicativos compatíveis com o sistema operacional Android. O modelo de negócios também é similar, os desenvolvedores distribuem seus jogos diretamente na loja sem a necessidade de intermediários e em contrapartida pagam uma porcentagem (30%) da receita proveniente dos jogos e aplicativos disponibilizados à respectiva loja virtual.

Para o desenvolvedor, a iPhone App Store inovou também com o modelo de negócios para distribuição de aplicativos e jogos para o iPhone ao permitir que os desenvolvedores distribuíssem diretamente na loja sem intermediários. Essa mudança de paradigma permitiu aos desenvolvedores evitar as longas negociações com publicadores, fabricantes de aparelhos e operadores de telefone e ainda aumentar a sua participação na receita.

E do outro lado, para o consumidor, a forte integração da iPhone App Store, posteriormente renomeada para Apple App Store, com o dispositivo em si levou muitos consumidores a experimentar jogos e aplicativos. A facilidade para encontrar, baixar e instalar os conteúdos foi responsável por promover um crescimento substancial no consumo de jogos para dispositivos móveis.

A iPhone App Store foi lançada com incríveis 500 jogos e aplicativos para o iPhone. Esse é uma quantidade de conteúdo imensa para os padrões da época. Na semana do lançamento da loja foram contabilizados o download de mais de 10 milhões de aplicativos (Bowcock 2008). De lá para cá, a quantidade de conteúdo só cresceu. As lojas possuem, juntas, praticamente 4 milhões de jogos e aplicativos disponíveis para download (ver Figura 3-3).

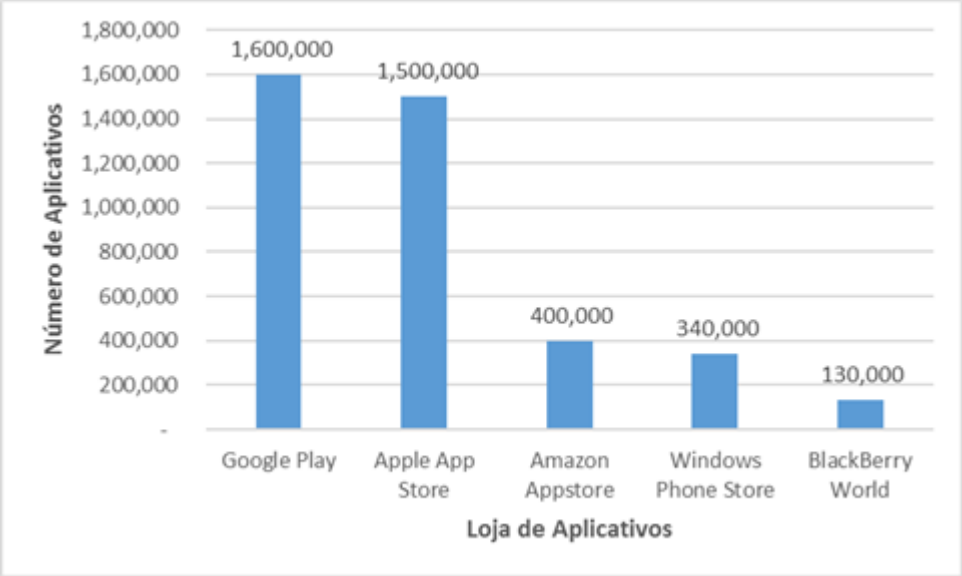

Figura 3-3: Número de aplicativos disponíveis nas principais lojas de aplicativos ao final de 2016 (Costello 2016; AppBrain 2017)

O surgimento das lojas de aplicativos acarretou no encerramento do modelo de negócios de distribuição por meio das operadoras de telefonia e fabricantes de aparelhos. A distribuição se tornou mais simples, rápida e barata. Além disso, também favoreceu a redução da pirataria. A exigência de um identificador único para um usuário impediu as pessoas de jogarem sem pagar ou de repassar o jogo para terceiros evitando os balcões de revendas de jogos.

3.2.2 Modelos de Negócios

Os modelos de negócios dentro da indústria de jogos digitais (Van Dreunen 2011) também evoluíram junto com a evolução na tecnologia e distribuição digital de conteúdo. Os principais modelos de negócios utilizados pelos desenvolvedores para geração de receita são hoje:

- **Pay-to-Play:** Nesse modelo, também chamado de Premium, o usuário é cobrado antes da realização do download e instalação do jogo em seu dispositivo móvel.
- **Free-to-Play:** Este modelo disponibiliza o jogo gratuitamente para usuários baixarem e instalarem em seus dispositivos móveis. A receita nesse modelo é proveniente principalmente da comercialização de conteúdo exclusivo, como níveis extras e itens virtuais, disponível somente através de compras realizadas dentro do próprio jogo. A comercialização de espaços publicitários dentro do jogo e solicitação de doações dos jogadores também são práticas comuns nesse modelo.
- **Paymium:** Esse modelo é uma versão híbrida do Pay-to-Play e Free-to-Play em que jogos pagos também disponibilizam conteúdo extra para compra dentro do jogo.
- **Subscription:** O modelo de assinatura, comum na indústria de jogos Multijogador Massivo Online (MMO), também se tornou popular em dispositivos móveis. O pagamento de uma assinatura periódica, normalmente semanal ou mensal, permite ao usuário usufruir do jogo sem limitações de conteúdo ou tempo.

No início da indústria de jogos móveis, o pay-to-play foi o principal modelo de negócios adotado. Essa realidade mudou somente no início de 2011 com o lançamento da funcionalidade de microtransação pela Apple, que permitia a compra de itens virtuais através de micropagamentos realizados diretamente dentro do jogo ou aplicativo, e não mais somente na loja de aplicativos. Por exemplo, o jogo Candy Crush oferece uma série de itens virtuais para auxiliar o usuário em caso de falha na finalização dos quebra-cabeças do jogo. É possível comprar desde movimentos extras, para permitir a realização de mais ações para aumentar as chances de finalizar o quebra cabeça com sucesso, até itens especiais para facilitar a resolução dos níveis.

Uma pesquisa de mercado da Flurry (ver Figura 3-4) mostra o crescimento da fatia de mercado do modelo *Free-to-Play* em comparação ao modelo *Pay-to-Play*. Em janeiro de 2011 a receita do modelo *Free-to-Play* representava 39% da receita total da indústria de jogos

móveis. Em junho de 2011, após o lançamento da funcionalidade de micropagamentos, o modelo *Free-to-Play* passou a representar 65% da receita proveniente das lojas de aplicativos (Valadares 2011).

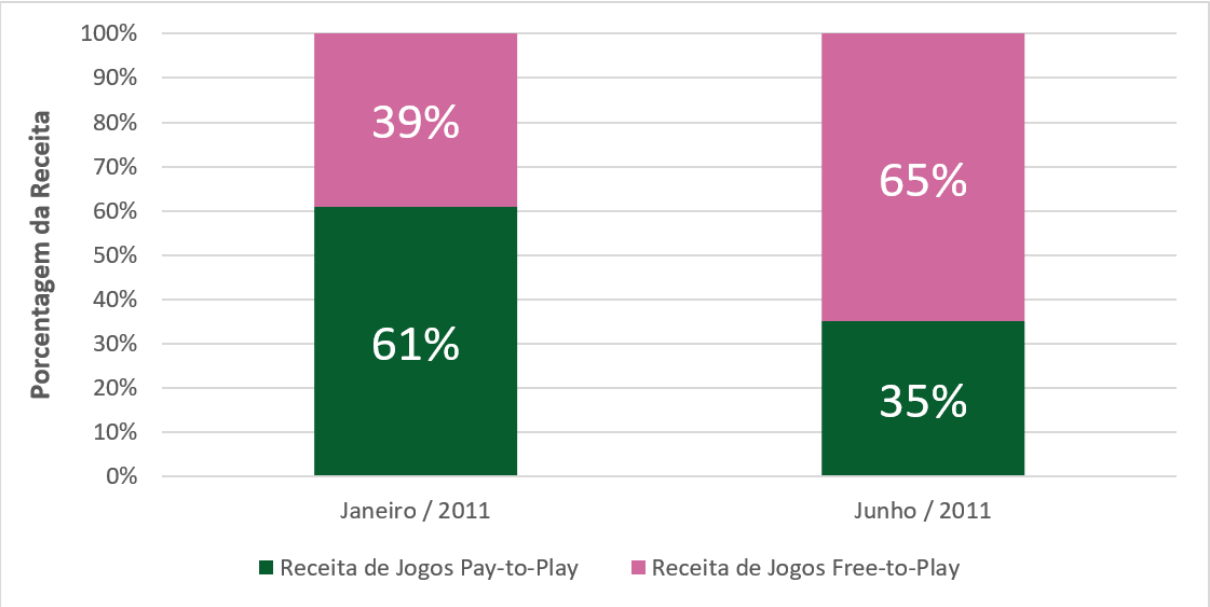

Figura 3-4: Modelo de negócios free-to-play supera o pay-to-play em junho de 2011 (Valadares 2011).

Desde então, o modelo *Free-to-Play* segue como modelo predominante na indústria de jogos para dispositivos móveis, sendo responsável pela maior parcela da geração de receita. Em recente relatório de mercado, o modelo *Free-to-Play* segue em franca expansão no mercado, se estabelecendo como o principal modelo de negócios dessa indústria sendo responsável por mais de 90% do faturamento (Jameson 2016).

3.3 O Modelo de Negócios Free-to-Play

A última década presenciou a ampla adoção do modelo de negócios free-to-play em que o consumidor evoluiu de uma transação (único pagamento adiantado adotado no modelo Pay-to-Play) para um relacionamento (pagamentos constantes realizados ao longo de vários meses). Essa mudança de paradigma remete à mesma evolução ocorrida no mercado de serviços de maneira geral. De maneira similar ao modelo *shareware* adotado na indústria de software, o faturamento no modelo Free-to-Play tem diferentes modelos de receitas:

- Donationware – Esse é um modelo de licenciamento que fornece uma versão completa e sem restrições do jogo ao usuário e solicita uma doação opcional a ser paga ao

desenvolvedor do jogo. O valor da doação pode ser estipulado pelo autor ou ser deixada a critério do usuário, baseado nas percepções individuais de valor do jogo.

- Trialware – Esse modelo limita o tempo de acesso em que o usuário pode testar o jogo até o fim do período experimental. Nesse momento, o programa para automaticamente de funcionar, a menos que o usuário pague uma taxa para ativar uma versão registrada e liberar o acesso completo ao jogo, sem restrições.
- Adware – O termo é um acrônimo de "advertising-supported software" e baseia a sua geração de receita na apresentação de publicidade ao usuário.
- Freemium – Esse é um modelo de negócios no qual o jogo é fornecido gratuitamente (*free*), mas pagamentos (*premium*) são cobrados para liberação de conteúdos extras e itens virtuais. O termo *freemium* é um neologismo cunhado em 2006 a partir da junção das palavras *free* e *premium*.

Esses modelos de receita foram adotados com diferente intensidade pela indústria de jogos para dispositivos móveis e apresentaram resultados distintos. O modelo Donationware, por exemplo, obteve baixa aceitação da indústria por não representar um modelo de geração de receita sustentável e confiável. O modelo Trialware, o mais próximo do modelo pay-to-play foi afetado pela ampliação na oferta de jogos gratuitos (free-to-play) de alta qualidade. Na corrida competitiva, os modelos *freemium* e *adware* levaram a melhor. A explicação para esse fato é que os jogadores preferem jogos gratuitos similares (Brasil 2015).

A avaliação dos principais modelos de negócios e receita mostra a importância dos modelos Freemium e Adware, atualmente responsáveis por 88.7% e 6.5% da receita total da indústria, respectivamente. A rápida adoção ao modelo Freemium demonstra o potencial de geração de receita por intermédio de microtransações. De fato, nos modelos Pay-to-Play e Trialware, a receita estava restrita ao valor pago pelo usuário para acesso completo ao jogo. o que em geral fica dentro da faixa de 2-10 dólares. Já no modelo Freemium, o usuário pode realizar repetidas compras dentro do jogo de diferentes valores, sem nenhum teto.

Os jogos de maior sucesso atualmente faturam mais de 20 dólares por usuário pagante, em média. O jogo Candy Crush (ver Figura 3-5a) fatura cerca de 40 dólares ao ano por usuário pagante, em média. Já o jogo Game of War (ver Figura 3-5b) faturou cerca de 550 dólares por usuário pagante em 2015 (Grubb 2016). E mais, uma parcela significativa dos usuários (10%) chegam a gastar milhares de dólares na compra de itens virtuais (Carmichael 2013).

Figura 3-5: Jogos para dispositivos móveis de sucesso. a) Imagem da esquerda referente ao jogo Candy Crush, um dos jogos casuais para dispositivos móveis mais populares da história com mais de 1 bilhão de downloads e com faturamento diário acima de 3 milhões de dólares. b) Imagem da direita referente ao jogo Game of War, um dos jogos hardcore para dispositivos móveis mais populares com faturamento diário acima de 2 milhões de dólares.

A avaliação das *taxas de conversão* no modelo Freemium, entretanto, demonstra que, em média, somente 2% dos usuários não-pagantes decidem pela compra e viram usuários pagantes. A taxa de conversão indica a porcentagem de usuários ativos de jogos free-to-play que realizam compras, ou seja, tornam-se usuários pagantes. O modelo Adware surgiu como uma forma alternativa e inteligente para geração de receita sobre os demais 98% dos usuários não convertidos a pagantes. É comum, portanto, encontrar os dois modelos sendo utilizados conjuntamente.

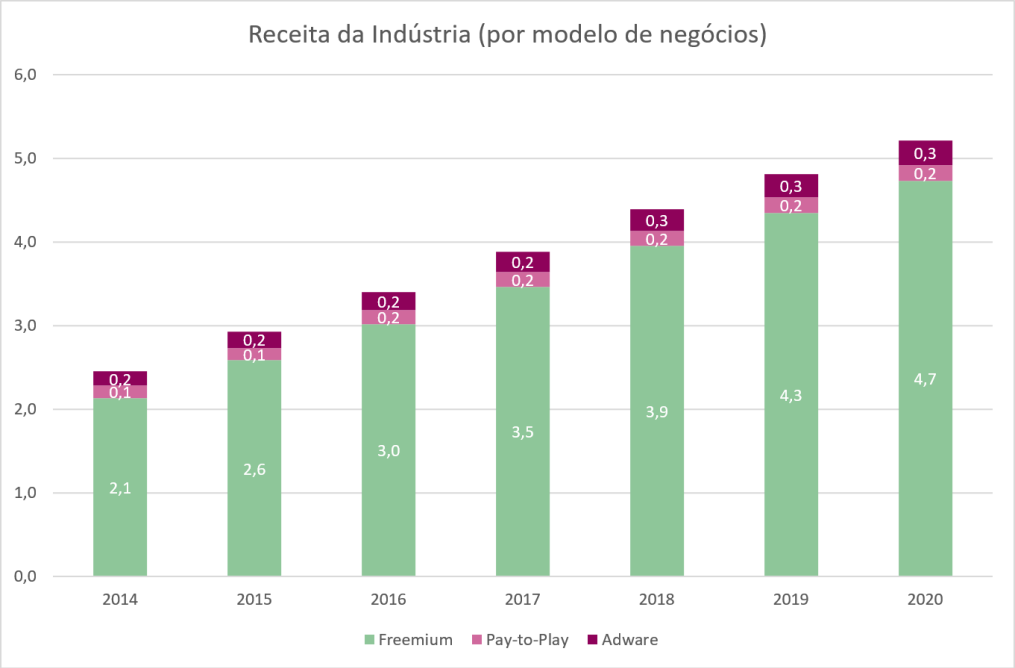

Figura 3-6: Receita da indústria de jogos para dispositivos móveis por modelo de negócios nos 4 principais países europeus (Reino Unido, Alemanha, França e Itália). O eixo X apresenta a linha do tempo com a receita realizada até 2016 e projetada até 2020. O eixo Y apresenta a receita em bilhões de dólares americanos. Fonte: (CyberAgent 2016).

3.4 A Gestão do Relacionamento com o Jogador

A mudança do modelo de negócios Pay-to-Play para os modelos Freemium e Adware implica na evolução do relacionamento com o consumidor. Enquanto no modelo Premium a relação com o usuário é finalizada no momento da compra do jogo, nos modelos Freemium e Adware o relacionamento começa no momento da instalação do jogo. É preciso manter o usuário ativo para aumentar as chances de convertê-lo a um usuário pagante (freemium) ou de apresentá-lo a um volume maior de campanhas publicitárias (adware).

3.4.1 Ciclo de Vida do Jogador

O modelo *freemium* é normalmente representado pelo funil ARM - acrônimo para Aquisição, Retenção e Monetização – proposto pela empresa de pesquisa de mercado Kontagent (Draganov 2014). Em termos práticos, essa representação permite visualizar os jogadores atravessando o funil durante as etapas do seu ciclo de vida no jogo.

Figura 3-7: Funil ARM.

Essa visão simplificada do ciclo de vida do usuário proposta pelo funil ARM permite entender o relacionamento macro do usuário com o jogo, porém não especifica etapas importantes tipicamente realizadas pelos desenvolvedores e tampouco inclui as etapas presentes no modelo Adware.

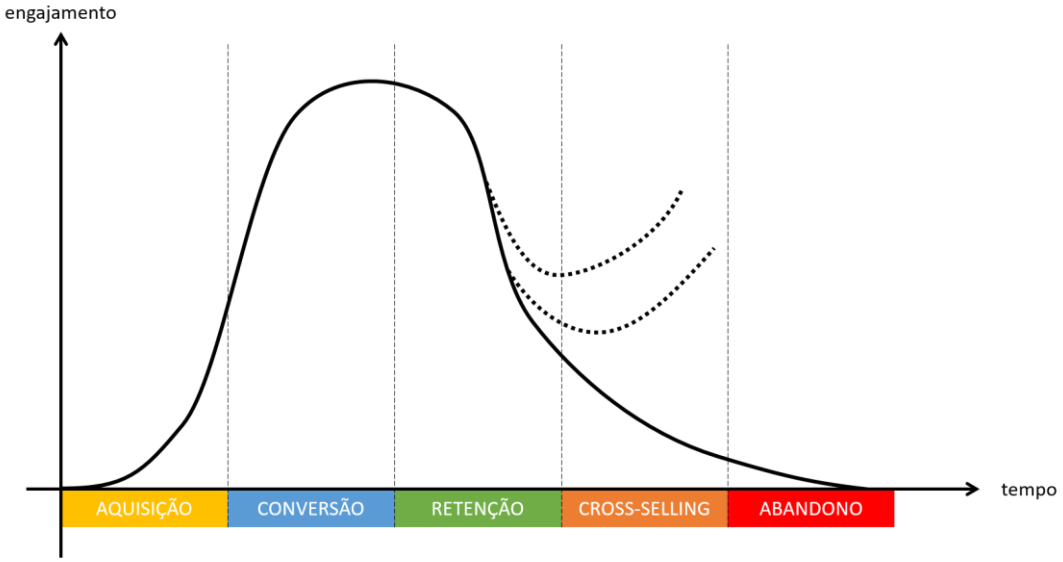

Figura 3-8: Ciclo de vida do usuário em jogos free-to-play para dispositivos móveis. O eixo X indica o tempo de relacionamento enquanto o eixo Y representa o engajamento do jogador.

Além disso, o formato de apresentação do ciclo de vida é diferente dos normalmente utilizados em outras indústrias. Para contornar esses problemas e explicar o ciclo de vida do usuário nos modelos Freemium e Adware em maiores detalhes, nós elaboramos um gráfico acima (Figura 3-8) e explicamos cada uma das etapas descritas.

1) **Aquisição:** ações de marketing (campanhas de publicidade ou funcionalidades virais como “convide um amigo”) para apresentação do jogo a novos usuários com o propósito de promover novos downloads e instalações;
2) **Conversão:** ações para converter o usuário de não-pagante para usuário pagante. Normalmente, nos primeiros momentos do ciclo de vida do usuário, quando o jogador está no ápice de engajamento com o jogo, itens virtuais são apresentados para o usuário em diferentes situações de jogo na busca pela oferta mais adequada para o sucesso na conversão;
3) **Retenção:** Após alcançar o ápice de engajamento, o relacionamento com o jogo inicia sua fase de decadência, o que requer ações de retenção e reengajamento para manter o usuário ativo no jogo e evitar o abandono precoce;
4) **Cross-Selling:** Caso as ações de retenção não surtam o efeito desejado e o engajamento do usuário permaneça em queda, ações de cross-selling são colocadas em prática para direcionar o usuário para outros jogos do portfólio da empresa desenvolvedora, forçando assim o início de um novo ciclo de vida do usuário;
5) **Abandono::** Caso todas as iniciativas anteriores para conversão do usuário, retenção e cross-selling não funcionem, o engajamento do usuário seguirá em queda até o abandono completo do jogo. Nesse caso, os jogos apresentam publicidade ao usuário para faturar através do modelo Adware como uma forma de reduzir a perda financeira resultante do futuro abandono do usuário.

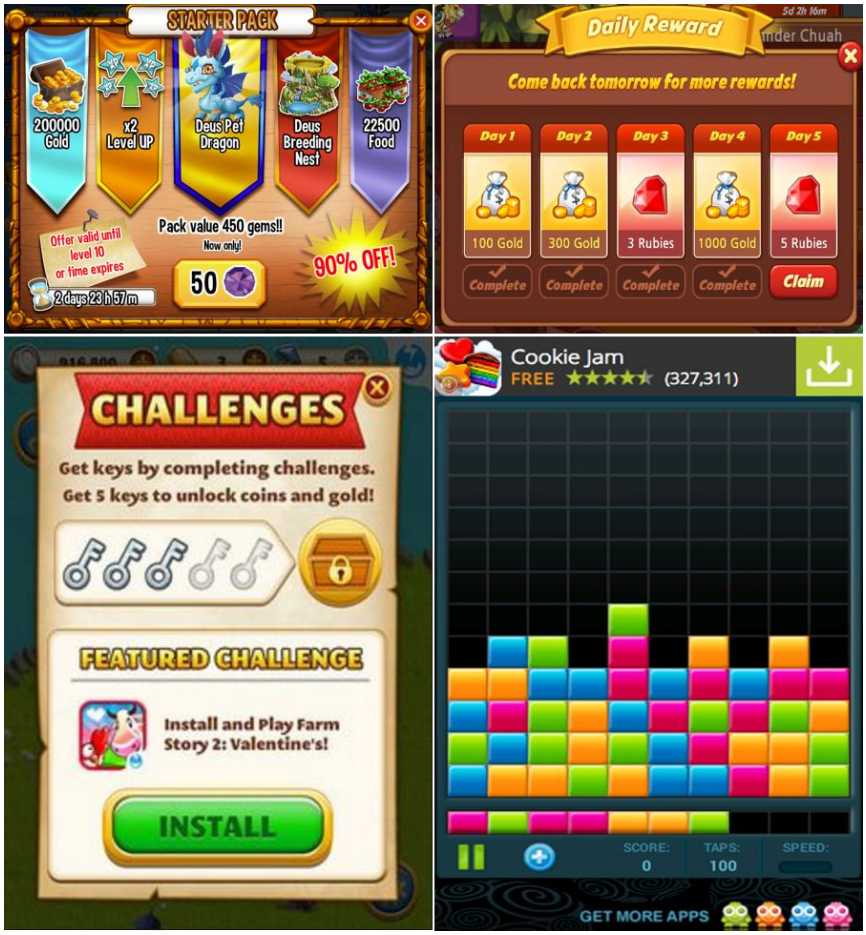

Figura 3-9: Demonstração da aplicação do modelo de negócios em jogos reais. Superior-Esquerda) Tela de jogo com oferta de itens virtuais para jogadores iniciantes com foco na conversão do usuário não-pagante para usuário pagante. Superior-Direta) Tela de jogo com ação de retenção para incentivar o usuário entrar diariamente no jogo para obter benefícios (itens virtuais). Inferior-Esquerda) Tela de jogo com demonstração do cross-selling. Inferior-Direita) Tela de jogo com banner na parte superior para geração de receita através de publicidade.

A análise do ciclo de vida do jogador proporciona uma visão geral da interação média existente do usuário com o jogo. Essa interação pode apresentar variações a depender do perfil do usuário, especialmente com relação ao consumo. Segundo análises de mercado (TapJoy 2016), os usuários podem ser classificados nas categorias abaixo, de acordo com o seu perfil de consumo.

- **Baleias** (do inglês, Whales): Esse grupo de usuários é formado pelos 10% dos usuários mais gastadores do jogo. Esses usuários realizam, em média 7,4 compras e gastam mais de 300 dólares por jogo. Esse grupo é responsável por cerca de 60% de toda a receita do jogo.
- **Golfinhos e Peixinhos** (do inglês, Dolphins e Minnows): Esses usuários são formados pelos demais 90% dos usuários convertidos a pagantes. Esses usuários possuem hábitos de consumo mais econômicos que as Baleias e, em média, realizam menos de 2 compras

e gastam 18 dólares por jogo. Esses usuários usuário, assim como as Baleias, tornam-se pagantes na Etapa 2 do ciclo de vida (Figura 3-8).

- **Tomadores de Ofertas de Anúncio** (do inglês, Offer Takers): Esse grupo é formado pelos usuários que interagem com anúncios publicitários, em especial anúncios com recompensas. Nesse formato de anúncio, o usuário recebe itens virtuais por assistir e interagir com anúncios publicitários. Esses usuários são responsáveis por cerca de 16% de toda a receita do jogo. Esses usuários acessam os anúncios na Etapa 5 do ciclo de vida.

Os perfis dos usuários complementam a avaliação da jornada do jogador a qual permite identificar uma peculiaridade com relação ao ciclo de vida de outras indústrias. Na indústria de jogos só existe o abandono voluntário e deliberado. O abandono, ou cancelamento, involuntário só acontece em casos raros de cancelamento de projeto ou término do suporte ao jogo. Essa avaliação também demonstra o papel fundamental da retenção para estender a relação com o jogador e criar mais oportunidades para conversão do usuário para pagante. A próxima seção aprofunda essa discussão sobre a importância da retenção na indústria de jogos para dispositivos móveis.

3.4.2 Importância da Retenção

A aquisição e retenção de usuários assumem papel de destaque para o sucesso de empreendimentos, independente da indústria em que atuam. Na indústria de jogos, especificamente, essas etapas apresentam diferentes grau de importância e impacto ainda maior.

O resultado prático da expansão do mercado de jogos para dispositivos móveis é a alta concorrência. De acordo com relatório da BBC (Lee 2013), aproximadamente 60% dos aplicativos nunca foram baixados. A Gartner também realizou uma pesquisa (van der Meulen & Rivera 2014) em que prevê que menos de 0,01% dos aplicativos obterão sucesso financeiro em 2018. De acordo com dados mais recentes, cerca de 79,6% dos aplicativos são considerados “zumbis”, ou seja, não foram listados em nenhuma das 39.171 listas das lojas de aplicativos (Valadares 2014).

Por esses motivos, a maioria dos desenvolvedores certamente irá, em algum momento, enfrentar o problema de levar as pessoas a conhecerem e baixarem seu jogo sem realizar gastos excessivos, em termos de marketing. Os desenvolvedores precisam adquirir usuários a um custo que torne possível obter um Retorno sobre o Investimento (ROI) positivo. Dada a sua

importância, essa métrica é comumente utilizada para determinar o sucesso de um jogo, e geralmente é o fator mais importante na decisão de continuar ou cancelar um projeto.

Em essência, apenas duas variáveis influenciam o ROI:

$$ROI = LTV - CPI$$
$$LTV = Valor\ do\ tempo\ de\ Vida$$
$$CPI = Custo\ por\ Instalação$$

O CPI é uma medida do custo de aquisição do usuário, ou no caso de jogos para dispositivos móveis, o custo para fazer um usuário realizar o download do jogo. O CPI é uma variável externa dependente da relação entre oferta e procura no mercado de publicidade para dispositivos móveis. Segundo pesquisas de mercado (Pearson 2015), o custo de aquisição de novos usuários nos sistemas operacionais iOs e Android cresce a taxas anuais de 75% e 169%, respectivamente. A tendência é que o CPI aumente consideravelmente nos próximos anos (David 2015; Kimura 2014; Cailean 2014). O cálculo do CPI é realizado conforme a fórmula abaixo.

$$CPI = \frac{Investimento\ em\ marketing}{\#\ Instalações\ relacionadas\ ao\ período\ de\ investimento}$$

O LTV, por outro lado, é uma medida que indica quanto cada usuário gasta no jogo. O LTV é uma variável interna fortemente dependente das características do jogo. Essa métrica de marketing estima o lucro futuro gerado pelo relacionamento com um jogador. Na sua forma mais simples, o LTV é uma função da Monetização e Retenção, como apresentado abaixo.

$$LTV = f\ (Monetização, Retenção)$$
$$onde, Monetização = Receita\ média\ por\ usuário\ do\ jogo$$
$$Retenção = Duração\ média\ do\ ciclo\ de\ vida\ do\ usuário$$

O cálculo formal do LTV pode ser realizado de diferentes formas, mas, por motivos didáticos, a fórmula descrita abaixo foi selecionada como ponto de partida para compreensão do LTV.

$$LTV = \sum_{i=1}^{n} m \cdot r^{i} \cdot \frac{1}{(1+d)}$$

$$onde, m = fluxo\ de\ caixa\ líquido\ no\ período$$
$$r = taxa\ de\ retenção$$
$$d = taxa\ de\ desconto$$
$$n = horizonte\ de\ cálculo$$

Onde, ***m*** denota a contribuição líquida por usuário no período; ***r*** corresponde à taxa de engajamento dos jogadores com o jogo; ***d*** é o desconto, frequentemente ignorado nos cálculos do LTV utilizados na indústria; e ***n*** representa o horizonte de tempo considerado para o cálculo do LTV.

A partir da fórmula, nós realizamos a análise da relação entre as variáveis mais importantes (***m*** e ***r***) e o seu impacto no valor do LTV. O valor de ***m*** é diretamente proporcional ao LTV. O aumento em ***m*** implica em um aumento linear proporcional no valor do LTV, mantidas constantes as demais variáveis. Por outro lado, a retenção (***r***) apresenta uma relação diferente com o valor do LTV. O aumento na variável ***r*** implica em um aumento exponencial do LTV, mantidas constantes as demais variáveis, como é possível ver na Figura 3-10.

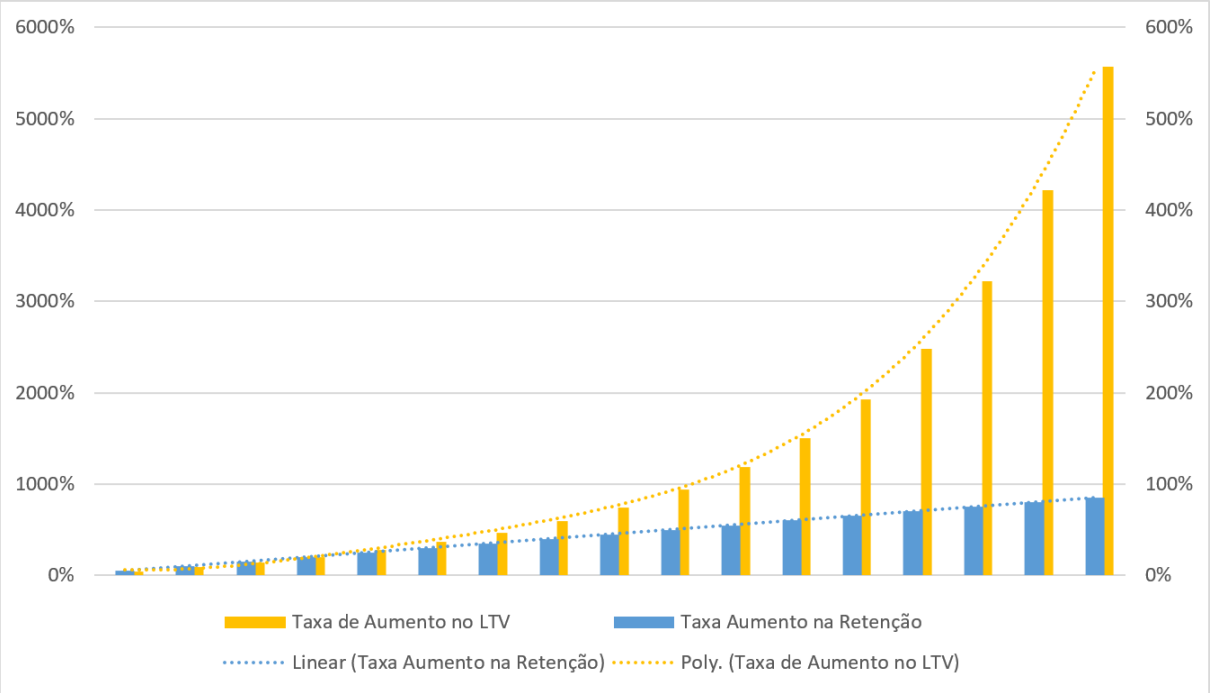

*Figura 3-10: Relação entre a taxa de retenção e o LTV. O eixo da esquerda apresenta a taxa de aumento no LTV provocado pelo respectivo aumento no eixo da direita referente à taxa de retenção (**r**). A simulação mantém constante os valores das demais variáveis (m=10; d=10; n=10).*

Nesse cenário, é possível observar que o investimento na melhoria da retenção dos usuários proporciona aumento exponencial no LTV. Com base na simulação realizada (Figura 3-11), a mudança na taxa de retenção mensal de 15% para 20%, por exemplo, resulta em um acréscimo de 42% no LTV. A identificação desse comportamento não desperta surpresa dado que fenômeno semelhante acontece em outras indústrias (Stillwagon 2014).

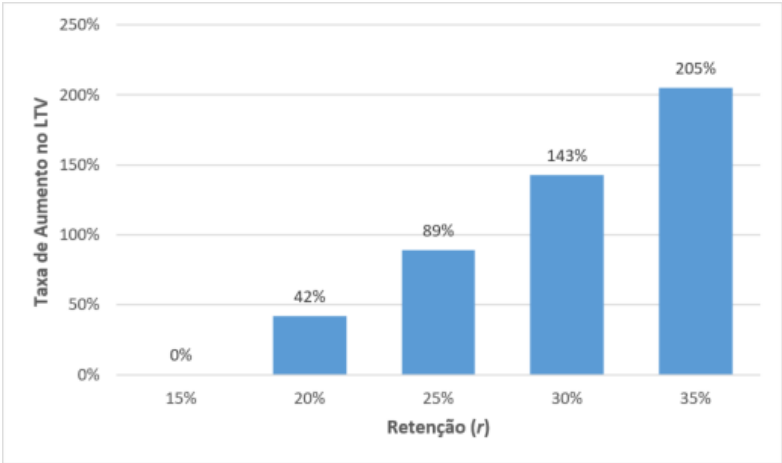

Figura 3-11: Relação entre taxa de retenção e a taxa de aumento no LTV. A simulação da taxa de aumento no LTV (eixo y) utiliza os valores de retenção (eixo x) e mantém constante os valores das demais variáveis (m=10; d=10; n=10). A taxa de aumento no LTV usa como base a retenção de 15%.

Essa relação exponencial entre a retenção e o LTV demonstra a importância de evitar o abandono de usuários como estratégia para aumentar o LTV. Na próxima seção, nós apresentamos a revisão bibliográfica dos trabalhos acadêmicos com relação à construção de modelos preditivos para a indústria de jogos free-to-play para dispositivos móveis.

3.5 A Construção de Modelos Preditivos em Jogos Móveis

A presente revisão bibliográfica da literatura sobre o domínio de revisão de abandono em jogos móveis adotou princípios do método de pesquisa proposto por Cooper (Cooper 1998) e Randolph (Randolph 2009). A classificação da pesquisa é apresentada na Tabela 3-1.

Característica	**Categorias**
Foco	Resultados da Pesquisa; Métodos de Pesquisa
Objetivo	Integração, Crítica e Identificação de Questões Centrais.
Perspectiva	Representação Neutra.
Cobertura	Exaustiva com seleção de citação
Organização	Metodológica
Audiência	Pesquisador

Tabela 3-1: Classificação da revisão bibliográfica.

O mapeamento dos artigos relevantes para esse estudo dentro do domínio de pesquisa considerou a execução de 03 (três) etapas. Em primeiro lugar foram definidas as bases de dados acadêmicas para busca de trabalhos relevantes. Em seguida o termo de busca foi definido para encontrar os artigos dentro dessas bases. Os artigos identificados, por fim, são avaliados de

acordo com critérios de inclusão e exclusão para determinação do estado da arte. As etapas são descritas em detalhes a seguir.

Na primeira etapa, a análise preliminar de trabalhos acadêmicos com base em um método *ad-hoc* de revisão de literatura permitiu mapear as principais conferências, jornais e base de artigos relativos ao domínio da pesquisa. A partir dessa informação, a revisão buscou nos seguintes bancos de dados acadêmicos:

- ACM Digital Library (ACM)
- IEEE Xplore Digital Library (IEEE)
- Science Direct (SD)
- CiteSeerX (CiteseerX)
- Engineering Village (Ei Compendex)
- Elsevier Scopus (Elsevier)

Na segunda etapa, a consulta às bases de dados por artigos relevantes à pesquisa considerou o termo de busca, descrito na Tabela 3-2, elaborado a partir da identificação dos principais termos encontrados nos artigos durante a revisão *ad-hoc* da literatura.

```
        ("games") AND ("player") AND ("churn")

        AND

        ("churn prediction" OR "churn detection" OR "churn prevention" OR
"customer churn" OR "player churn" OR "behavior analysis" OR "behavior scoring"
OR "behavior")

        AND

        ("prediction") AND ("detection")
```

Tabela 3-2: Termo de busca.

A etapa final consistiu em agrupar os artigos identificados através do uso do termo de busca nas bases de artigos e filtrar de acordo com os critérios de inclusão e exclusão. A inclusão de um trabalho é determinada pela relevância (acredita-se que o trabalho é um potencial candidato a tornar-se um estudo primário) em relação às questões de investigação, determinada pela análise do título, palavras-chave, resumo e conclusão. Os seguintes critérios de inclusão foram definidos:

1) O artigo foi escrito em Inglês ou Português.
2) O artigo foi publicado dentro dos últimos 10 anos (antes de 2016).

3) O artigo usou dados extraídos de jogos reais.

Os artigos encontrados foram analisados a partir do título, palavras-chave, resumo e conclusão, para exclusão dos estudos que se enquadrem em qualquer dos casos abaixo:

1) Estudos irrelevantes para a pesquisa.
2) Estudos Repetidos: se determinado estudo estivesse disponível em diferentes fontes de busca, somente a primeira pesquisa será considerada.
3) Estudos Duplicados: caso dois trabalhos apresentem estudos semelhantes, apenas o mais recente e/ou o mais completo seria incluído, a menos que tivesse informação complementar;
4) Estudos que apresentassem texto, conteúdo e resultados incompletos.

O mapeamento sistemático foi executado de acordo com o método apresentado e identificou um total de 29 artigos relacionados ao tema de pesquisa, após a aplicação dos critérios de inclusão. Desse total de artigos foram identificados 12 artigos repetidos, ou seja, disponíveis em diferentes fontes, e 8 irrelevantes para a pesquisas por tratarem de problemas distintos de previsão de abandono em jogos. Após a realização de todas as etapas foram selecionados 9 artigos para a revisão bibliográfica.

A Figura 3-12, em sua imagem esquerda, mostra a quantidade de trabalhos retornados por cada uma das bases científicas utilizadas na pesquisa considerando todos os 29 artigos previamente selecionados. Já a Figura 3-12, em sua imagem direita, apresenta a área de pesquisa dos artigos e demonstra a existência de somente 9 artigos dentro do domínio de pesquisa dessa tese.

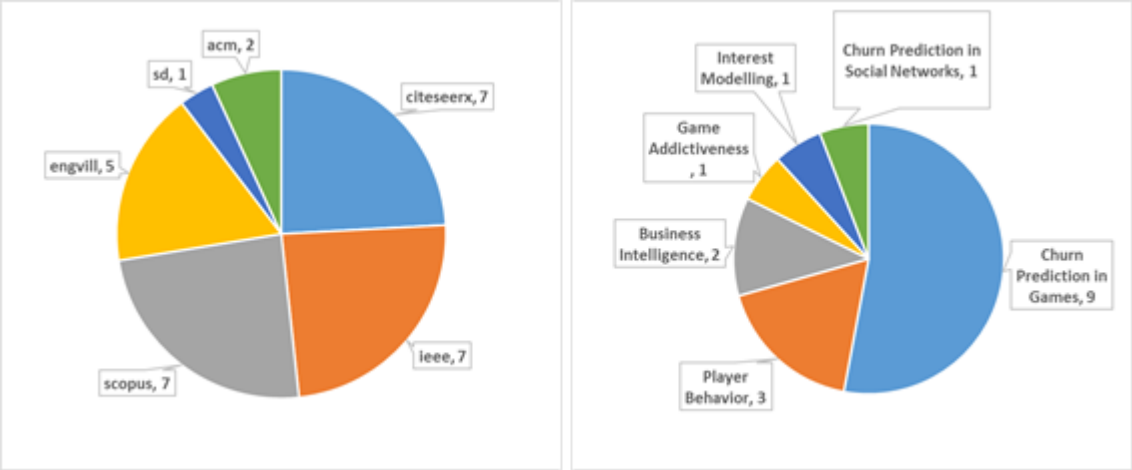

Figura 3-12: Distribuição dos artigos pesquisados em relação à fonte de dados e ao domínio.

Os artigos excluídos da revisão classificados como pertencentes aos temas *Player Behavior*, *Game Addictiveness* e *Interest Modelling*, apesar de fugirem da temática de pesquisa, apresentam aspectos interessantes com relação ao conhecimento sobre o domínio para modelagem do comportamento do usuário. Essa informação é útil na transformação dos dados para criação de atributos com maior poder discriminatório e será utilizada na determinação da solução proposta para o problema. Por outro lado, o foco da revisão consiste na avaliação dos trabalhos acadêmicos envolvendo exclusivamente a previsão de abandono em jogos. Esses artigos selecionados com base no método proposto são apresentados na Tabela 3-3.

Título	**Ano**	**Tema**	**Referência**
Churn Prediction in MMORGPs: A Social Influence Based Approach	2009	MMOG; Subscription	(Kawale et al. 2009)
Churn Prediction in MMORPGs using Player Motivation Theories and an Ensemble Approach	2011	MMOG; Subscription	(Borbora et al. 2011)
User Behavior Modelling Approach for Churn Prediction in Online Games	2012	MMOG; Subscription	(Borbora & Srivastava 2012)

Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning	2013	Online; Gambling	(Coussement & De Bock 2013/9)
Predicting Player Churn in Multiplayer Games using Goal-Weighted Empowerment	2013	MOBA; Free-to-Play	(Edge 2013)
Departure Prediction of Online Game Players	2014	MMOG; Subscription	(Savetratanakaree et al. 2014)
Predicting Player Churn in the Wild	2014	Mobile; Free-to-Play	(Hadiji et al. 2014)
Churn Prediction for High-Value Players in Casual Social Games	2014	Online e Mobile; Free-to-Play	(Runge et al. 2014)
Churn Prediction in Online Games Using Players' Login Records: A Frequency Analysis Approach	2015	MMOG, MOBA e TPS; Free-to-Play	(Castro & Tsuzuki 2015)

Tabela 3-3: Distribuição dos artigos de acordo com o ano e domínio de pesquisa, onde MMOG = Massive Multiplayer Online Games e MOBA = Multiplayer Online Battle Arena.

A distribuição dos artigos com relação ao ano de publicação, plataforma e modelo de negócios é apresentada na Figura 3-13. O gráfico de distribuição superior revela o caráter recente dessa área de pesquisa com todos os artigos tendo sido lançados a partir de 2009 e apresentando maior concentração entre o biênio 2013-14. O gráfico da esquerda demostra que a maioria dos artigos foram produzidos para as plataformas PC e Online. Somente 2 artigos avaliaram dados de jogos para dispositivos móveis. E com relação ao modelo de negócios, 4 dos 9 artigos avaliados consideraram jogos comercializados através de assinatura. Outros 4 dos 9 artigos estudaram jogos distribuídos sobre o modelo de negócios Free-to-Play.

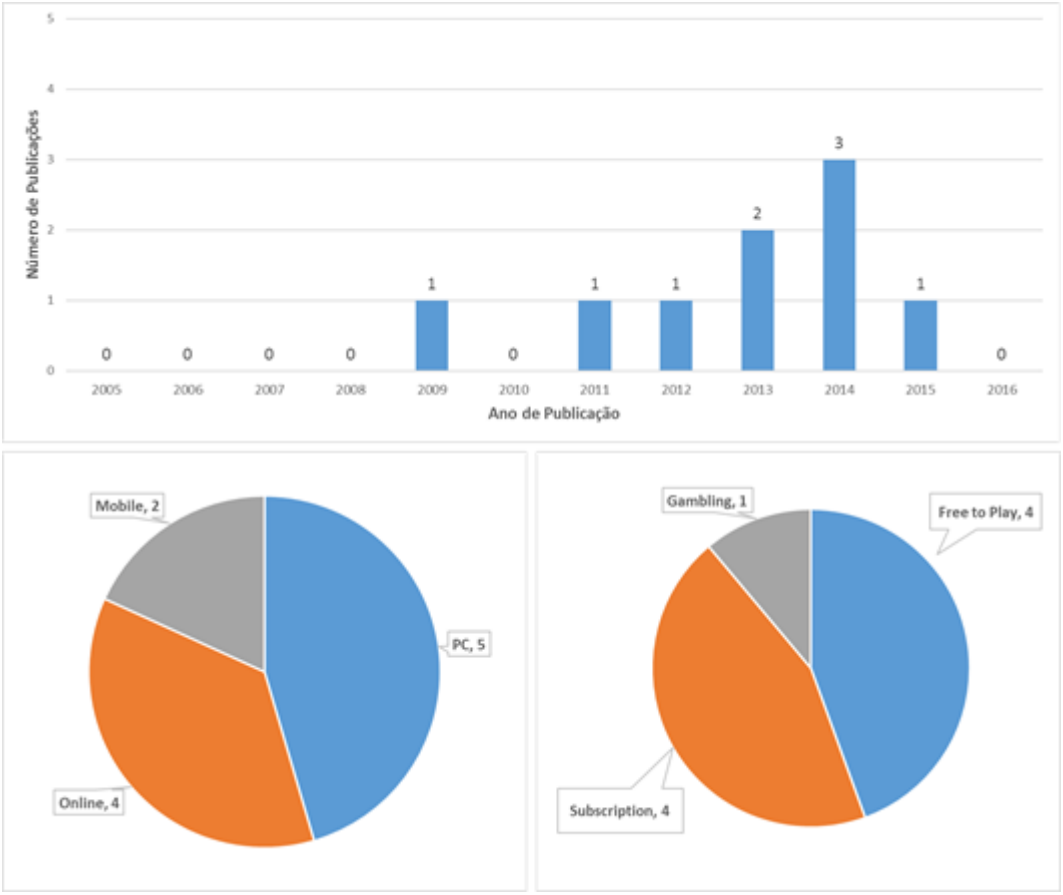

Figura 3-13: Distribuição dos artigos com relação à data de publicação (cima), plataforma (esquerda) e modelo de negócio (direta) dos jogos avaliados nos trabalhos.

A interseção entre os dados demonstra que dos 9 artigos pesquisados somente 2 deles atuam exatamente dentro do domínio de pesquisa da tese, ou seja, jogos para dispositivos móveis distribuídos sobre o modelo Free-to-Play. Dado que os demais artigos apresentam métodos e abordagens importantes para o tratamento e modelagem dos dados em domínio relacionado, estes também foram incluídos na revisão bibliográfica. A revisão bibliográfica é apresentada a seguir organizada a partir das etapas do processo de construção do modelo preditivo.

3.5.1 Etapa 1: Entendimento do Negócio

Não analisaremos a etapa de entendimento de negócios porque é difícil extrair o quanto cada artigo entendeu do domínio. No entanto, este entendimento tem consequências concretas nas outras etapas que são analisadas por nós a seguir.

3.5.2 Etapa 2: Entendimento dos Dados

Não analisaremos a etapa de entendimento de negócios porque é difícil extrair o quanto cada artigo entendeu do domínio. No entanto, este entendimento tem consequências concretas nas outras etapas que serão analisadas por nós.

Os atributos, também chamados de variáveis e características, são as informações utilizadas como input dos modelos preditivos para classificação do usuário entre churner e non-churner em função da probabilidade de abandono do jogador. Nesta seção, apresentamos as variáveis de entrada usadas nos artigos analisados.

Os artigos utilizam essencialmente dados do usuário relacionados ao comportamento (histórico de uso), assim como dados monetários (histórico de compras e pagamentos) no jogo. Os dados pessoais, usualmente extraídos no momento do cadastro, em geral não estão disponíveis nesse domínio e, portanto, não são mencionados nos artigos avaliados nessa revisão.

Alguns artigos (Borbora et al. 2011; Kawale et al. 2009; Borbora & Srivastava 2012) que utilizaram a base de dados do mesmo jogo, Sony EverQuest II, tiveram acesso ao mesmo tipo de informação referente principalmente a dados de sessão. Esses dados apresentam tempo e *duração das sessões*, *missões completadas* e *pontos conquistados* pelos usuários. A diferença entre os artigos reside na maneira com que o conhecimento do especialista foi introduzido para derivação de atributos semânticos.

No artigo de Kawale e colegas (Kawale et al. 2009), por exemplo, a probabilidade de abandono do usuário é determinada como uma função baseada em dois principais fatores: *engajamento* e *influência* social. O atributo derivado engajamento foi proposto para medir a distribuição do tempo gasto no jogo durante a janela de observação. Essa variável permite verificar que a duração média de sessão de jogo de usuários non-churners é maior do que os usuários churners (ver imagem da direita na Figura 3-14). O atributo influência social também foi criado para determinar a probabilidade de abandono do usuário dado que os seus amigos virtuais, descritos tecnicamente como vizinhos na rede social do jogo, abandonaram o jogo. É possível perceber a relação entre as variáveis na imagem da esquerda na Figura 3-14.

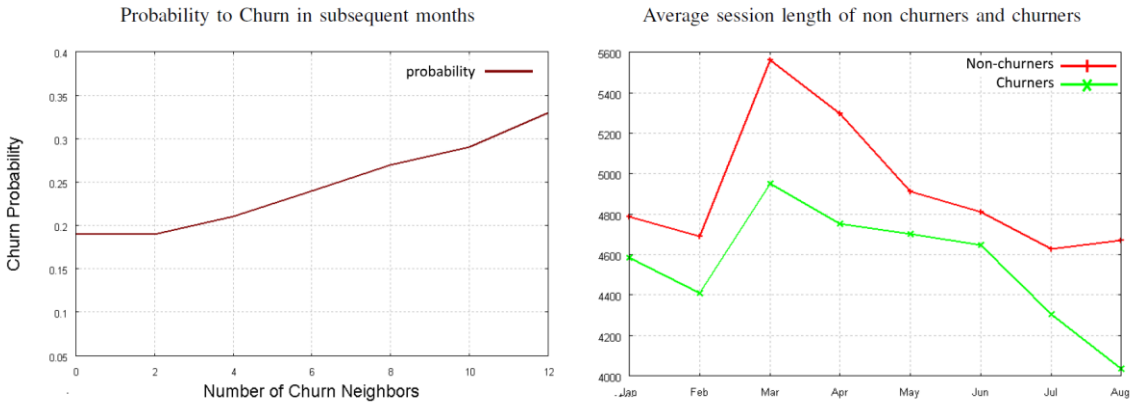

Figura 3-14: Probabilidade de abandono nos meses subsequentes (esquerda) e Duração média da sessão e usuários churners e non-churners (Kawale et al. 2009).

Já o artigo de Borbora e colegas (Borbora et al. 2011) define três categorias de atributos (engajamento, entusiasmo e persistência) derivados a partir de duas diferentes abordagens (direcionada à teoria ou domínio e direcionada a dados). A partir da abordagem direcionada à teoria, ou domínio, foram definidos 4 atributos orientados à conquista e socialização:

- Taxa de participação em quest;
- Taxa de mortes de monstros;
- Taxa de pontos de experiências ganhos;
- Taxa de interação com grupo;

E também foram definidos 14 atributos derivados a partir de abordagem direcionada à dados tais como:

- Duração total das sessões
- Pontos de experiências ganhos;
- Número total de participação em quests;
- Número total de mortes

A comparação do modelo preditivo com o uso dos dois diferentes grupos de atributos permite identificar que os atributos definidos com base em uma abordagem orientada ao domínio apresentam melhor desempenho (ver Figura 3-15).

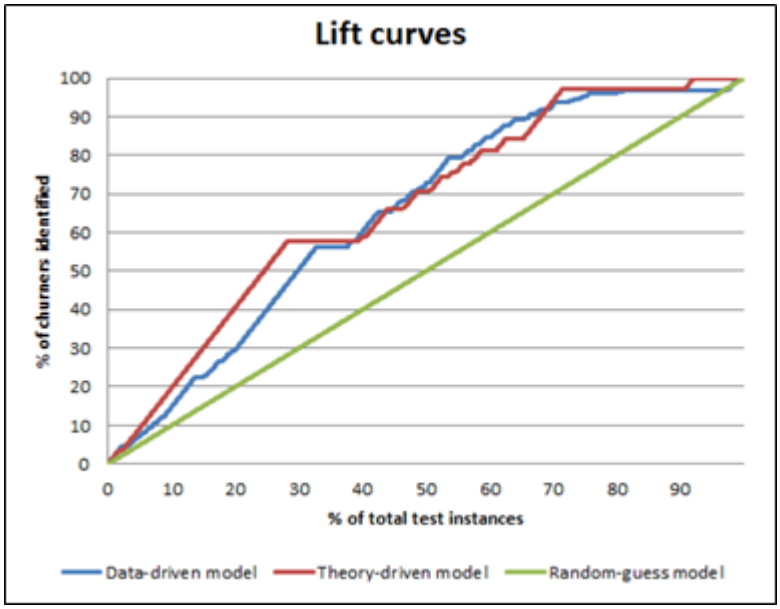

Figura 3-15:Avaliação dos modelos a partir da curva Lift (Borbora et al. 2011).

Em outro artigo (Borbora & Srivastava 2012), o histórico de dados relacionados à sessão de jogo dos usuários, em uma janela de tempo semanal, é usado para construção das variáveis independentes. O registro de atividades dos usuários extraídos do banco de dados do jogo apresenta somente ações dos usuários sem a informação sobre início e fim de cada sessão. Os autores definiram uma heurística particular, com base em informações de especialistas, para determinar que uma sessão consiste de um conjunto de ações realizadas pelo usuário separadas por um intervalo inferior a 30 minutos. Essa definição permite agrupar as ações em sessões de jogo e por sua vez em agregar as sessões em intervalos semanais para criar os seguintes atributos básicos:

$$v_1 = Número\ de\ sessões\ por\ semana$$
$$v_2 = Soma\ da\ duração\ das\ sessões\ por\ semana$$
$$v_3 = Soma\ da\ duração\ entre\ as\ sessões, referente\ à\ inatividade, por\ semana$$

A partir desses atributos mais básicos são derivados atributos semânticos como engajamento, entusiasmo e persistência. O atributo engajamento é autoexplicativo e calculado como a média do atributo básico. O atributo entusiasmo captura o aumento, ou redução, na magnitude do atributo básico. A direção da mudança no atributo básico, para cima ou para baixo, é capturada pelo atributo derivado persistência. Os atributos que proporcionaram melhor ganho de informação foram os relacionados ao número e duração de sessões:

$$num\_sessions\_enthu = aumento\ ou\ redução\ na\ magnitude\ de\ v_1$$

$$num_{sessions_{engage}} = média\ do\ número\ de\ sessões$$

$$sl\_mins\_enthu = aumento\ ou\ redução\ na\ magnitude\ de\ v_2\ (em\ minutos)$$

$$num\_sessions\_persist = direção\ da\ mudança\ de\ v_1$$

Outro artigo (Savetratanakaree et al. 2014) utiliza atributos similares ao considerar também o número de sessões e a duração das sessões de jogo agrupados em janelas diárias, ao invés de semanais. A diferença é que nesse artigo esses atributos são chamados de revisitações e tempo de permanência, respectivamente. Os autores avaliam diferentes tamanho de janelas (k dias) para identificar qual delas mostra melhor resultado, em termos de precisão. O trabalho considerou os tamanhos de 2 a 10 dias e o melhor resultado foi encontrado para a janela com tamanho de 2 dias (k = 2).

Em um outro artigo (Coussement & De Bock 2013/9), os dados foram pré-processados para a construção de 60 atributos relacionados aos usuários de jogos de azar. Os atributos foram classificados pelos autores como pertencentes a duas categorias: informações de comportamento e dados demográficos. Os atributos considerados mais importantes e estatisticamente significativos estão relacionados ao RFM:

- **Recência**: Quão recentemente o usuário realizou uma compra?
- **Frequência**: Qual a frequência de compra do usuário?
- **Valor monetário**: Quanto o usuário gastou?

E logo em seguida encontram-se os atributos relacionados à frequência de uso do jogo, como frequência de sessões (último mês, última semana e total) e frequência de apostas (último mês, perdidas).

Na pesquisa de Hadiji e colegas (Hadiji et al. 2014), os atributos considerados são, em geral, similares aos utilizados em outros trabalhos como atributos relacionados à sessão (ex: frequência e duração) e monetização (ex: número de compras e gasto médio por sessão). Entretanto, há atributos específicos derivados da teoria proposta por (Bauckhage et al. 2012) em que indica como o interesse do jogador diminui com o tempo seguindo uma Distribuição de Weibull:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

Os parâmetros k e λ da função, também chamados de *shape parameter* e *scale parameter* respectivamente, são obtidos a partir do ajuste da função aos dados de observação do usuário. Esses parâmetros são usados como atributos para treinamento do modelo preditivo. Os atributos considerados mais importantes na pesquisa, entretanto, são apresentados abaixo:

- Número de Sessões
- Número de Dias
- Tempo Médio por Sessão
- Tempo Médio entre Sessões
- Tempo de Inatividade

A lista contém principalmente atributos relacionados à sessão que apresentaram performance superior aos atributos relacionados à monetização e também ao interesse do usuário (parâmetros da Distribuição Weibull). Esse artigo apresenta um grande avanço com relação aos demais ao propor uma lista de atributos genérico, independentes do jogo (game-agnostic), e validar em 5 diferentes jogos.

No artigo (Runge et al. 2014), escrito por um dos funcionários da desenvolvedora de jogos sociais Wooga, os atributos considerados pertencem a três categorias. Os autores utilizaram dados de atividade *in-game* referente às séries temporais de *logins* por dia e precisão. Também foram usados dados relacionados à monetização, como as séries temporais de receita gerada por usuário. Por último, dados do perfil do jogador, como por quanto tempo o usuário tem acessado o jogo e o país de origem do jogador.

Ainda no artigo de Runge, foram estipulados centenas de atributos diferentes dentro dessas categorias. A aplicação da técnica de seleção de atributos (*feature selection*) permitiu identificar o conjunto de atributos a ser usado na construção do modelo preditivo. A lista final de atributos para cada um dos jogos analisados é apresentada na Tabela 3-4.

Base de Dados (Jogo)	Conjunto de Atributos (variáveis independentes)
Diamond Dash	Time series of rounds played, accuracy, invites send, days in game, last purchase, days since last purchase
Monster World	Time series of logins, level, in-game currency #1 balance (WooGoo), currency #2 balance (Magic Wands)

Tabela 3-4: Lista final de atributos utilizados para modelagem do classificador.

É interessante observar que apesar da lista inicial de atributos derivados para ambos os jogos ser similar, a lista final após a aplicação da seleção de atributos é disjunta, ou seja, não há repetição de atributos nas duas listas. Os autores não discursam sobre esse achado.

Outro trabalho (Castro & Tsuzuki 2015) apresenta como principal contribuição de pesquisa a análise de 4 (quatro) métodos de transformação de séries temporais em vetores de atributos com o objetivo de aprimorar a performance da previsão de abandono. O primeiro deles é baseado no RFM com a modificação do "M" de valor monetário para Intensidade. A modificação foi motivada pela ausência de dados monetários dos registros dos usuários na base de dados. Os atributos aqui considerados são:

- **Recência**: Intervalo de tempo, em dias, desde a última sessão de jogo registrada.
- **Frequência**: Número de sessões de jogos registradas durante a janela de observação.
- **Intensidade**: Soma da duração de todas as sessões de jogo registradas na janela de observação.

O segundo método é baseado no Domínio de Tempo (DT). Nesse método, a janela de observação é dividida igualmente em um determinado número n de fatias de tempo e todas as sessões de jogos registradas durante esse período são agrupadas em suas respectivas fatias. Esse esquema de discretização utilizou o valor de n=16 para dividir a janela de observação estipulada em 30 dias gerando assim fatias de tempo de 45 horas. Essas fatias contém a informação sobre o número total de sessões de jogo iniciadas durante ao sua respectiva sub-janela de tempo.

O terceiro método, chamado de Domínio da Frequência (DF), utiliza a técnica Wavelet Power Spectrum (WPS) que é baseada na Transformada de Haar. Os valores de entrada para essa técnica é o vetor resultante do método anterior (DT). O quarto e último método usa o Wavelet Packet Decomposition (WPD) para uma análise tempo-frequência TFPD. Esses métodos foram comparados através do uso classificador k-NN para identificar qual deles promove a melhor performance em relação à curva ROC (ver Figura 3-16).

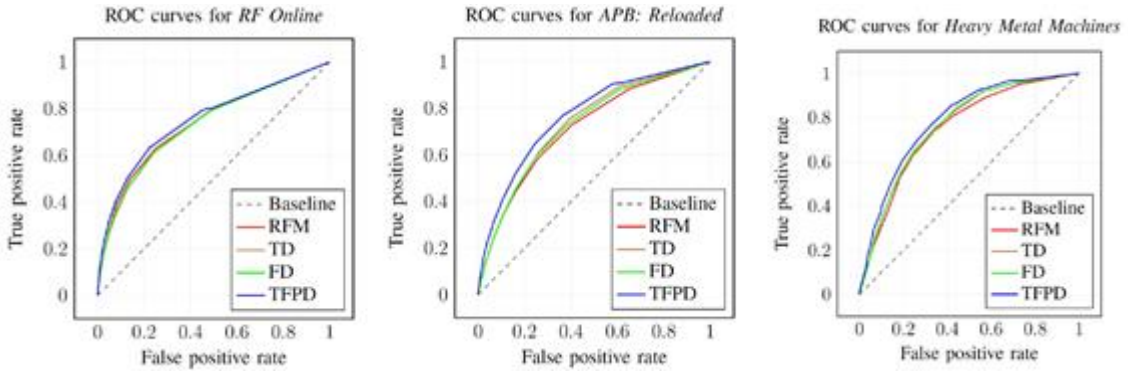

Figura 3-16: Análise dos métodos através da curva ROC aplicada aos três jogos sob análise (Castro & Tsuzuki 2015).

No artigo (Edge 2013), os dados básicos extraídos do jogo Dota 2 são essencialmente relacionados ao personagem do jogador (ex: identificador e nível) e às estatísticas ao final de cada partida (ex: número de inimigos mortes, número de assistências e quantidade de ouro). A performance do usuário na partida é uma informação importante, com base na visão de especialistas, dado que o jogador normalmente abandona a partida quando percebe que seu time possui poucas chances de vitória. Essa performance é medida através do atributo derivado KDA medido a partir da fórmula abaixo:

$$KDA = \frac{Kills + Assists}{Deaths}$$

O atributo derivado KDA é calculado não somente para o usuário sob análise como também para seu time e também o time adversário. A relação entre essas variáveis (ver abaixo) são usadas como atributos para determinar a intuição do usuário sobre a sua probabilidade de vencer a partida.

$$\left(\frac{KDA_{Player}}{KDA_{Team}}, \frac{KDA_{Player}}{KDA_{Opp}}, \frac{KDA_{Team}}{KDA_{Opp}}\right) \rightarrow P(Win \mid Actions)$$

O KDA_{Player} refere-se ao atributo KDA do jogador enquanto as variáveis KDA_{Team} e KDA_{Opp} são simplesmente o atributo KDA do time do jogador e do time oponente do jogador. Em resumo, as pesquisas utilizam métodos ad-hoc, baseados no conhecimento empírico de especialistas dos jogos sob análise, para definição e derivação de atributos. Nos artigos com

modelo de negócio de assinatura foram utilizados dados monetários, por outro lado os artigos com modelo Free-to-Play em geral utilizaram somente dados de comportamento.

3.5.3 Etapa 3: Preparação dos Dados

Todos os artigos realizaram transformação sobre os dados para derivar atributos mais relevantes com base na experiência de especialistas sobre o dinâmica de comportamento dos usuários. Do total de artigos, 4 utilizaram dados específicos dos jogos em análise (Borbora et al. 2011; Coussement & De Bock 2013/9; Edge 2013; Runge et al. 2014) enquanto os demais 5 optaram pelo uso de dados genéricos como frequência e duração das sessões de jogo.

As decisões de projeto para transformação e derivação de atributos são realizadas de maneira *ad-hoc* através da prática experimental. Não há estudo para comparar o impacto na performance do modelo preditivo com relação às possíveis configurações das variáveis abaixo:

- Categorias dos atributos (ex: pessoal, comportamento e monetização)
- Tipos de atributos (ex: específicos ou genéricos)
- Modelagem da frequência e duração de uso
- Agregação temporal dos atributos
- Agregação semântica dos atributos (ex: engajamento, interesse e entusiasmo)

Um artigo (Borbora et al. 2011) até apresenta uma abordagem voltada para a comparação de performance entre diferentes tipos de atributos no modelo, mas a análise atua em somente uma das variáveis acima listadas e com restrições sobre a quantidade de atributos existentes para teste. Além desses aspectos, a definição da configuração das janelas e também do rótulo do usuário pertencem a essa etapa.

3.5.3.1 Janelas de Observação e Resultado

O tamanho das janelas de observação varia de acordo com o modelo de negócios e características próprias do jogo como gênero, público e plataforma. Alguns artigos (Borbora & Srivastava 2012; Kawale et al. 2009; Borbora et al. 2011) utilizaram a base de dados do mesmo jogo, Sony EverQuest II, para modelar o problema de previsão de abandono. A base de dados possui 36 semanas de informações cadastrais e comportamentais dos usuários.

Na pesquisa de Borbora & Srivastava (Borbora & Srivastava 2012), a janela de desempenho foi estipulada em 13 semanas por corresponder ao ciclo de um trimestre. Os

autores estipularam um limite, chamado de *activity threshold*, usado para classificar os non-churners de acordo com seu nível de atividade. Os usuários com o último registro de atividade após essa data limite são considerados non-churners ativos. Já os usuários com a última atividade registrada, na janela de desempenho, antes dessa data limite são considerados non-churners inativos. Esse threshold foi estipulado em 1 mês dado que a periodicidade de pagamento da assinatura é mensal. Isso significa que o threshold identifica os non-churners que estiveram ativos desde o último pagamento realizado.

Já os churners são identificados, como já mencionado anteriormente, através do cancelamento da assinatura do jogo, pois se trata de MMOG. A pesquisa avalia ainda o impacto desses parâmetros utilizados na pesquisa. A análise do número de sessões por semanas, dentro da janela de observação estipulada em 13 semanas, indica uma redução expressiva no nível de atividade dentro do jogo nas últimas 2-3 semanas de observação. Essa informação indica, apesar de não comprovada na pesquisa, que a performance do classificador não deve ser degradada significativamente mesmo ao utilizar uma janela de observação com tamanho menor.

O artigo de Kawale (Kawale et al. 2009) não menciona explicitamente o tamanho das janelas de observação e resultado. Uma vez que o conjunto de treinamento do modelo preditivo utiliza os dados do mês de agosto de 2006 é razoável concluir que o tamanho da janela de observação é de 30 dias. A configuração das janelas não é informada também em outros artigos (Borbora et al. 2011). No trabalho desenvolvido por Savetratanakaree (Savetratanakaree et al. 2014) consta a informação sobre a configuração das janelas em 30 dias enquanto em outro artigo (Runge et al. 2014) é utilizada a janela de observação de 14 dias. O artigo informa que esse valor foi escolhido com base em experimentos empíricos que demonstram aumento na performance com base no Área sob a Curva ROC. Esses experimentos, entretanto, não são apresentados no trabalho e tampouco citados como pertencentes a uma pesquisa externa. O estudo realizado por Runge (Runge et al. 2014) não indica a configuração da janela de resultado.

As maiores janelas de tempo foram encontradas no artigo que avalia o abandono de jogadores na indústria de *online gambling* (Coussement and De Bock 2013). A janela de observação e resultado apresentam 17 e 5 meses, respectivamente. Já as menores janelas de tempo foram constatadas no trabalho de Hadiji (Hadiji et al. 2014) em que as janelas de observação e resultado foram ajustadas para o tamanho de 7 dias.

No artigo (Castro & Tsuzuki 2015), a configuração da janela é realizada de maneira independente para cada um dos três jogos analisados na pesquisa. A janela de resultado é fixada em 30 dias para todos os jogos. Não há indicação ou explicação para o uso desse tamanho de janela. Com relação à janela de performance, os jogos RF Online, APB: Reloaded e Heavy

Metal Machines utilizam janelas de 30, 15 e 7 dias, respectivamente. Esses valores foram determinados com base na distribuição acumulada de métricas como LTV por usuário e tempo máximo de inatividade entre sessões de jogo, também por usuário. O artigo, entretanto, não apresenta as distribuições ou gráficos para fundamentar as escolhas.

O trabalho de Edge (Edge 2013) avalia o abandono em partidas do jogo Dota 2 por meio de dados extraídos de jogo com uso de um grão diferente dos demais. O grão representa o nível de detalhamento em que a informação é armazenada para processamento via mineração de dados. Enquanto todos os artigos aqui avaliados consideram o grão usuário, esse estudo utiliza o grão partida, ou sessão de jogo. Nesse caso a janela de tempo considerada, incluindo observação e resultado, contém duração de 20 a 40 minutos.

Em resumo, a configuração da janela de observação é também realizada de maneira distinta em cada um dos trabalhos através de um processo ad-hoc. E mais, mesmo os artigos que apresentam a informação sobre a duração da janela utilizada, não é apresentada qualquer informação sobre o tipo de janela utilizada na pesquisa, se estática ou escalonada. É esperado que as janelas apresentem configurações, em especial o tamanho, diferentes para cada um dos jogos devido as suas peculiaridades, entretanto não há método para estipular qual o tamanho ideal dessas janelas. Além disso, também não existe estudo comparativo sobre o impacto na performance do modelo preditivo referente ao uso de diferentes configurações para as janelas de observação e resultado.

3.5.3.2 Rótulo do Usuário

Em geral, a definição do status do usuário está intimamente ligada ao modelo de negócios adotado pelo jogo considerado nos trabalhos científicos avaliados. No caso dos jogos com modelo assinatura (*subscription*, do inglês), os usuários que já cancelaram as suas assinaturas são considerados churners e os usuários non-churners são aqueles com a assinatura ainda ativa. Esses artigos (Borbora & Srivastava 2012; Borbora et al. 2011; Kawale et al. 2009) explicitam a adoção dessa definição, enquanto o artigo (Savetratanakaree et al. 2014) não informa o método utilizado.

O artigo (Coussement & De Bock 2013/9) analisa um jogo com modelo de negócios *gambling* no qual adota a definição de que o jogador é considerado churner caso não tenha acessado o jogo por um período de 4 meses. Na pesquisa (Hadiji et al. 2014) conduzida por Hadiji a definição de churn é realizada de duas maneiras diferentes utilizadas para comparar o impacto na performance do modelo preditivo.

1) O usuário sem nenhuma sessão de jogo após uma data estipulada, chamada de *cutoff date*, é considerado *churner*.
2) O usuário com baixo número de sessões de jogo ou quantidade reduzida de dias jogados após a *cutoff date* é considerado *churner*. A quantidade restante de dias ativos no jogo devem estar dentro de uma janela deslizante de tamanho pré-definido.

A primeira definição é considerada mais restrita e informa ao modelo preditivo para identificar usuários sem sessões de jogos restantes, ou seja, sem a possibilidade de serem impactados por uma ação de retenção pelos desenvolvedores do jogo. A segunda definição, considerada mais relaxada, por outro lado, permite treinar o classificador para identificar usuários ainda ativos, mas com propensão a abandonar o jogo. Essa identificação permite treinar um classificador para fornecer informação acionável para tratar problemas reais e complexos da indústria.

Na pesquisa conduzida por Runge (Runge et al. 2014), o objetivo consiste em prever o abandono dos usuários considerados de alto valor. A pesquisa analisa a contribuição dos *top percentile* usuários pagantes na receita total do jogo e identifica que os 7% usuários no topo da lista são responsáveis pela geração de praticamente 50% da receita do jogo. Essa proporção é comum na indústria, como é possível perceber nesse relatório sobre o mercado (Carson 2015). O foco dessa pesquisa é em estudar e investigar o fenômeno de abandono de jogadores pertencentes exclusivamente a esse grupo de usuários.

Nesse estudo, a definição de abandono é realizada com base no prazo de inatividade dos usuários considerados de alto valor. O valor limite (*threshold*, do inglês) utilizado para definir quais usuários abandonaram o jogo é determinado com base na distribuição dos dias de inatividade entre as sessões de jogos dos usuários. Por exemplo, dado que um usuário acessa o jogo nos dias t=1 e t=3, e novamente no dia t=7, essa informação fornece duas amostras de dias de inatividade: a primeira amostra apresenta 1 dia de inatividade (3-1-1 = 1) e a outra apresenta 3 dias de inatividade (7-3-1 = 3).

A Figura 3-17 demonstra o histograma com a curva de distribuição e também de distribuição acumulada dos dias de inatividade. O gráfico mostra que menos de 2% dos usuários de alto valor ficam inativos por mais de 14 dias. E com base nessa informação, os autores consideraram 14 dias de inatividade para definição da situação de abandono. Com essa definição, 98% dos jogadores definidos como churners de fato abandonaram o jogo.

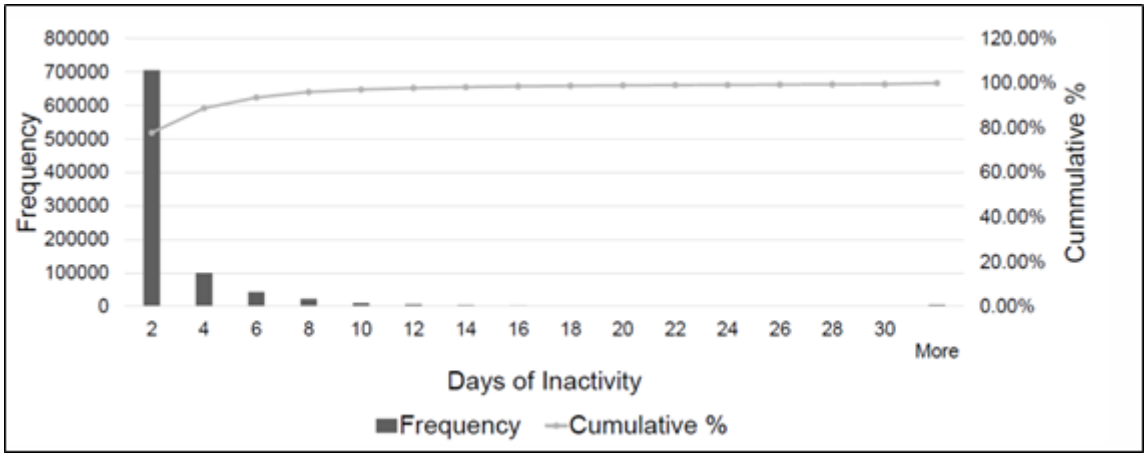

Figura 3-17: Histograma de dias de inatividade do jogo Diamond Dash (Runge et al. 2014).

Em outra pesquisa (Castro & Tsuzuki 2015), a escolha do método para definição de abandono não é explicitada. Os autores mencionam que a prática adotada na indústria de jogos é considerar o usuário ausente por um período superior a 30 dias como churner, mas não informam explicitamente que esse foi o método selecionada.

Na última pesquisa identificada realizada por Edge (Edge 2013), o objetivo é construir um modelo para prever os usuários propensos a abandonar uma partida no jogo Dota 2. Esse é um dos jogos mais populares do mundo do tipo MOBA. Dota 2 é baseado em partidas online, com o objetivo primário de cada partida sendo derrotar o time adversário destruindo o cristal que se localiza no centro da base adversária. Cada time é formado por exatamente 5 personagens, totalizando 10 personagens participantes da partida.

As partidas duram em média 20 a 40 minutos e a desistência de um ou mais dos 10 usuários em cada partida promovem o desbalanceamento do jogo impactando na experiência dos usuários remanescentes. Ao contrário das demais pesquisas que objetivam identificar o abandono do usuário de um jogo, o propósito dessa pesquisa é detectar quando um usuário vai abandonar uma partida do jogo. A definição de abandono, nesse caso, é simples: o usuário ao abandonar uma partida é considerado churner.

Em resumo, a definição do rótulo do usuário é realizada de maneira distinta em cada um dos trabalhos acadêmicos, em geral com base em um processo ad-hoc. Não existe consenso com relação ao método a ser utilizado para estipular o status do rótulo e tampouco estudo comparativo sobre o impacto na performance do classificador referente à utilização de diferentes métodos.

3.5.4 Etapa 4: Modelagem

A análise do comportamento do jogador proposta em (Borbora & Srivastava 2012) envolve o agrupamento de usuários com base nas informações históricas sobre o número de sessões semanais. O agrupamento é realizado com o uso da técnica *k-means* sobre a amostra de churner e non-churners usando 5 grupos como entrada do processo. A determinação do número de grupos foi obtida ao plotar o gráfico de Within Cluster Sum of Squared Errors e observar que o "joelho" das curvas, tanto para churners quanto para non-churners, acontece próximo de k=5.

Dos 5 perfis de comportamento criados, é possível identificar 4 padrões distintos tanto para churners quanto para non-churners. A principal diferença entre eles é que os perfis de churners apresentam inclinação (derivada) negativa nas últimas semanas de análise indicando a redução no número de sessões de jogo anterior ao abandono (Figura 3-18 - esquerda). Esse mesmo fenômeno não é observado nos perfis dos usuários classificados como non-churners (Figura 3-18 – direita).

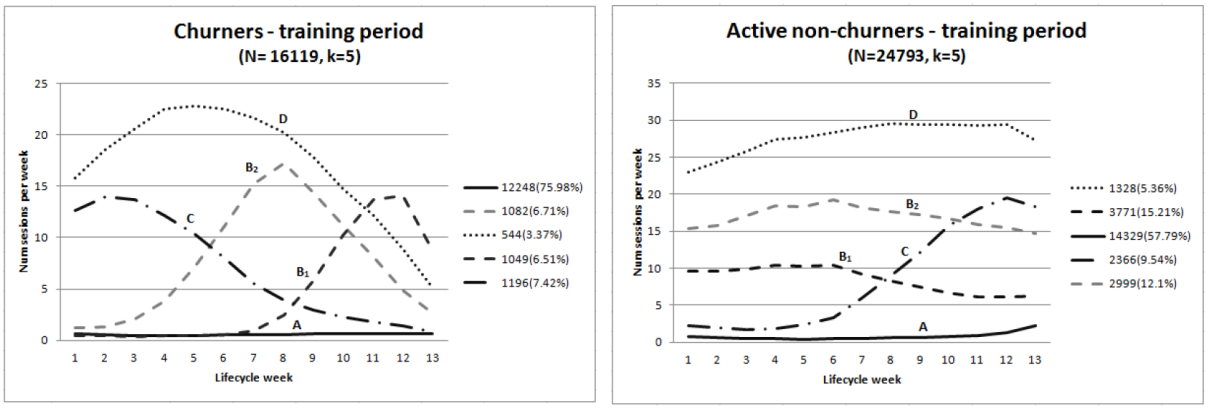

Figura 3-18: Centróides dos clusters produzidos a partir das amostras de dados (Borbora & Srivastava 2012). Esquerda) Churners. Direita) Non-churners. O eixo x apresenta a janela de desempenho com tamanho de 13 enquanto o eixo y mostra o número de sessões por semana.

O classificador foi modelado utilizando diferentes técnicas comparadas a partir da métricas de performance Recall, Precision e F-measure. A pesquisa considerou as técnicas JRip, J48, NaiveBayes, BayesNet, k-NearestNeighbor (k=3), Regressão Logística, Multiplayer Perceptron, SVM (com uso de kernel RBK) e wClusterDist. Esse último método, proposto no artigo, classifica uma nova instância ao comparar a soma ponderada das distâncias euclidianas para dois grupos ou classes.

Dentre os classificadores tradicionais, o SVM apresentou melhor Precision (71,6) NaiveBayes o melhor Recall (83,3) e também o melhor F-measure (57,8). A técnica proposta

no artigo, wClusterDist, apresentou o melhor F-measure geral (58,3) e o segundo melhor Recall (74,1).

O artigo (Kawale et al. 2009) também abordou a análise do comportamento do jogador a partir de duas verticais. A primeira delas com relação à duração média da sessão. A distribuição dessa informação sobre os 8 meses analisados revela que o churner apresenta, em geral, menor duração média durante todo o ciclo de vida. Ou seja, o usuário churner não apresenta o mesmo nível de engajamento que o usuário non-churner.

A segunda vertical refere-se à quantidade de vizinhos em uma estrutura de comunidade social. A redução no número de vizinhos devido ao abandono do jogo implica em aumento da probabilidade de abandono do usuário. Os autores consideraram que a previsão de abandono pode ser capturada efetivamente utilizando o engajamento do usuário definido com base nessas duas verticais. A construção do classificador é realizada a partir de três diferentes modelos:

- Simple Diffusion Model
- Classification Based on Network and Player Engagement
- Modified Diffusion Model (MDM)

A pesquisa revela que a técnica MDM é capaz de combinar a influência social com o engajamento do usuário e prover uma melhoria significativa na precisão da previsão de abandono. No artigo (Borbora et al. 2011) foi proposta a realização de um experimento para comparar a performance de classificação de modelos criados a partir de dados processados segundo a teoria da motivação dos jogadores. Em outras palavras, a pesquisa considerou dados dos logs de atividades dos usuários para geração de atributos dentro de 2 principais categorias com base nessa teoria: atributos relativos ao sucesso do jogador e atributos relativos à sociabilização do jogador. Todos esses atributos são gerados como taxas (porcentagens) pois os autores acreditam que esses atributos indicam melhor a intensidade do engajamento do usuário do que valores absolutos. Uma terceira categoria de dados foi gerada seguindo uma abordagem orientada a dados com valores absolutos. Os atributos gerados foram combinados em 4 (quatro) grupos diferentes e avaliados através de um processo de validação cruzada (10-fold) a partir da aplicação de árvore de decisão (C4.5) como classificador.

A melhor performance foi alcançada pelo grupo com atributos derivados da abordagem baseada em dados consistindo de atributos relacionados ao uso (ex: quantidade e duração de sessões), performance dentro do jogo (ex: número de missões e níveis finalizados, número de

mortes e quantidade de pontos de experiência conquistados) e aspectos sociais (ex: número de interações com outros usuários e número de amigos virtuais que abandonaram o jogo).

O artigo (Savetratanakaree et al. 2014) propõe um novo tipo de atributo chamado SLKDays. Esse termo é um acrônimo para Staytime Last K Days calculado a partir do logaritmo da média da variável duração da sessão considerando os últimos k dias.

$$SKLdays = log\left(\frac{\sum_{i=2}^{k} S_i}{k}\right)$$

$$Staytime(S) = Logout.time\ -\ Login.time$$

Os autores informam que a performance com a inclusão do atributo proposto SLKDays aumenta significativamente a performance do classificador modelado a partir da técnica SVM. A melhor performance foi encontrada com k=2 indicando que a informação mais recente, referente aos últimos dois dias da janela de observação, promovem maior ganho de informação. O artigo, entretanto, não indica quais os outros atributos usados na pesquisa para comparar a melhoria da performance com a suposta inclusão do atributo proposto.

O artigo (Coussement & De Bock 2013/9) utiliza dados do cadastro do usuário e informações históricas de uso do jogo de azar. Os autores validam os resultados a partir da repetição (5x) da aplicação de validação cruzada (2-fold). Os classificadores aplicados para construção do modelo e comparação de performance foram Árvores de Decisão (CART), Generalized Addictive Models (GAM), Random Forest e GAMens.

A escolha de GAM é realizada devido a pesquisa (107), dos mesmos autores, informar que esse classificador supera a regressão logística. O Random Forest é um classificador conjunto (ensemble) composto da combinação Bagging, RSM e Arvore de Decisão (CART). Já GAMens é um classificador conjunto (ensemble) composto da combinação de GaM, Bagging e Random Subspace Method (RSM).

A técnica Bagging é utilizada para combinar modelos em problemas de classificação com implementação simples e intuitiva. A base de dados é dividida em subconjunto de dados aleatoriamente criados com reposição. Cada subconjunto é usado para treinar um classificador do mesmo tipo. As saídas dos classificadores são combinadas por meio do voto majoritário com base em suas decisões. Para uma dada instância, a classe que obtiver o maior número de votos será então a resposta.

Já o RSM (Ho 1998) é um classificador conjunto composto por vários classificadores operando sobre um subespaço do conjunto de atributos original. As saídas dos classificadores

são combinadas para determinar o resultado da classificação. A avaliação de performance considera somente as métricas Top-decile Lift e Lift index normalmente aplicadas na indústria. Em termos acadêmicos essa decisão de pesquisa torna mais complicado a comparação de performance e definição de benchmarks.

No artigo (Hadiji et al. 2014) a construção do modelo foi avaliada a partir de validação cruzada (10-fold). Para efeito de comparação foram utilizados 4 (quatro) diferentes classificadores: Redes Neurais, Regressão Logística, Naive Bayes e Árvore de Decisão. A métrica F1-score foi utilizada para comparação dos classificadores. A performance apresentada é a melhor entre os demais artigos avaliados nessa revisão. Por outro lado, a configuração dos classificadores, como a topologia da rede neural ou o algoritmo utilizado na árvore de decisão, não é informada. A validação e comparação dos resultados dessa pesquisa em trabalhos posteriores torna-se inviável.

O artigo (Runge et al. 2014) utiliza validação cruzada (10-fold) para avaliar os modelos desenvolvidos na pesquisa. As técnicas de Redes Neurais (RNA), Regressão Logística (RL), Árvore de Decisão e SVM foram utilizadas para construção dos classificadores. A comparação de resultado entre os classificadores é realizada a partir da área sob a curva ROC (AUC_ROC).

Os classificadores RNA e RL apresentaram performance similar tendo apresentado performance de AUC_ROC de 0,815 e 0,814 para o jogo Diamond Dash, respectivamente. E a performance de 0,930 e 0,924 para o jogo Monster World, respectivamente. A rede neural foi configurada com uma única camada escondida e a quantidade n de nós nessa camada foi estipulada a partir da heurística abaixo.

$$n = \frac{\#attributes \ + \ \#classes}{2} + 1$$

Na pesquisa conduzida por Castro (Castro & Tsuzuki 2015) a principal contribuição proposta consiste na definição de métodos de transformação de séries temporais em vetores de atributos a serem utilizados para treinamento de classificadores. O classificador k-NN foi utilizado para comparar entre os diferentes grupos de atributos a partir da performance alcançada em cada grupo individualmente. As métricas utilizadas para medir a performance foram Top-Decile Lift e Área sob a curva ROC (AUC_ROC). O método Time-frequency Plane Domain (TFPD) proposto apresentou a melhor performance.

E por fim, o artigo (Edge 2013) escolheu o classificador k-NN para previsão do abandono do usuário. A técnica Naive Bayes também foi aplicada devido a sua simplicidade e

rapidez. E por fim, os autores aplicaram um classificador baseado em regras para extrair informações legíveis sobre a tomada de decisão interna do classificador. A performance é medida através da apresentação direta da matriz de confusão sem apresentação de qualquer das métricas relacionadas como precisão e F1-score. A ausência de informações na apresentação da matriz, entretanto, inviabiliza a definição das métricas de precisão, recall ou ainda F1-score.

Referência	**Técnica**	**Precisão**	**Recall**	**F1-score**
(Borbora & Srivastava 2012)	SVM (1) e Naive Bayes (2)	71,6 (1)	83,3 (2)	57,8 (2)
(Kawale et al. 2009)	Modified Diffusion Model com AdaBoostM1	50,1	29,8	37,3
(Borbora et al. 2011)	Decision Tree com J48	69,3	84	76
(Savetratanakaree et al. 2014)	SVM	98,2	84,5	90,8
(Hadiji et al. 2014)	Árvore de Decisão	N/A	N/A	91,6
(Edge 2013)	Naive Bayes	N/A	N/A	N/A

Tabela 3-5: Comparativo das técnicas via precisão, recall e F1-score.

Referência	Técnica	Top-decile Lift	Lift Index	AUC
(Coussement & De Bock 2013/9)	Árvore de Decisão com CART (1)	3,46 (1)	0,77 (1)	-
(Runge et al. 2014)	Rede Neural - Diamond Dash (1); Monster World (2)	N/A	N/A	0,815 (1) 0,930 (2)
(Castro & Tsuzuki 2015)	k-NN com TFPD – RF (1); APB (2); HMM (3)	4,03 (1) 3,32 (2) 2,87 (3)	N/A	0,750 (1) 0,759 (2) 0,791 (3)

Tabela 3-6: Comparativo das técnicas via Lift e AUC.

As Tabelas (Tabela 3-5 e Tabela 3-6) apresentam as informações detalhadas sobre as técnicas e métodos de avaliação utilizados nos trabalhos considerados. É importante notar que determinados artigos (Savetratanakaree et al. 2014; Kawale et al. 2009) não informam o F1-score que calculado para efeitos didáticos de comparação entre as pesquisas com base na fórmula abaixo:

$$F_1 = 2 \cdot \left(\frac{precision \cdot recall}{precision + recall}\right)$$

Em resumo, os artigos demonstram a visão orientada a dados da maioria das pesquisas ao apresentarem foco no desenvolvimento, demonstração e aplicação de algoritmos e modelos específicos, em detrimento da produção de conhecimento acionável para permitir a tomada de decisão dentro das empresas. A única exceção é o trabalho de Runge (Runge et al. 2014) que aplica os resultados da pesquisa na prática para detectar o impacto na taxa de resposta de ações de marketing para promoção da retenção de usuários.

Ainda assim, em todos os trabalhos, as escolhas das técnicas são realizadas através de um processo ad-hoc, assim como as métricas de performance. A performance geral da solução, entretanto, é dependente de todas as etapas do processo de mineração de dados envolvendo desde a extração dos dados, pré-processamento, modelagem e avaliação de performance. A ausência de diretrizes na área não permite comparar trabalhos distintos e entender, por exemplo,

porque pesquisas que utilizaram mesma técnica e mesma base de dados (Savetratanakaree et al. 2014; Borbora & Srivastava 2012) apresentam performances tão distintas.

3.6 Conclusão e Observações

A análise da indústria de jogos para dispositivos móveis permite identificar a retenção como peça fundamental no sucesso de jogos por influenciar diretamente na receita. As características únicas dessa indústria com alta taxa de abandono e restrição ao acesso a determinados tipos de dados do usuário tornam a previsão de abandono uma ferramenta importante para proativamente atuar na prevenção do abandono de usuários.

A revisão bibliográfica da previsão de abandono nessa indústria demonstra a viabilidade de aplicação de técnicas de mineração de dados para construção de modelos preditivos. A revisão revela ainda que esse tema de pesquisa é um território ainda pouco explorado com número incipiente de pesquisas na área. E mais, o aprofundamento na análise dos trabalhos demonstra que o processo de desenvolvimento dos modelos preditivos ainda é realizado de maneira ad-hoc na maioria dos casos. Essa situação torna complexo o entendimento da contribuição real das pesquisas assim como a avaliação comparativa dos resultados alcançados.

4 Avaliação das Especificidades de Jogos Móveis

A avaliação da previsão de abandono na indústria de jogos móveis revela a aplicação de processos ad hoc com a finalidade de produção de modelos preditivos para atender especificamente aos jogos avaliados nas pesquisas. Os artigos identificados na revisão bibliográfica aplicam soluções distintas para os desafios envolvidos no domínio de jogos móveis. Aliado a isso, a baixa quantidade de trabalhos publicados na área resulta em pouco conhecimento gerado sobre esse domínio para reuso entre os trabalhos.

Neste capítulo, investigamos as principais especificidades do domínio de jogos móveis. Essa análise examina as características únicas do relacionamento entre os usuários e os jogos móveis para compreensão das decisões-chave, e suas respectivas escolhas, no estado da arte da construção de modelos para previsão de abandono. Enfim, elencamos as escolhas que pretendemos avaliar experimentalmente a fim de obter possíveis diretrizes para área.

4.1 As Especificidades e seus Impactos

A análise do estado da arte na indústria de jogos móveis em comparação com outras indústrias revela as especificidades de cada um dos domínios com impacto direto nas decisões-chaves a serem tomadas durante o processo de construção do modelo preditivo de abandono. Na indústria de telecomunicações, área utilizada como exemplo de domínio maduro com diretrizes estabelecidas, nós identificamos as especificidades desse domínio com os respectivos impactos em decisões-chaves e as escolhas identificadas na literatura (Tabela 2-1).

Nesta seção nós realizamos uma análise similar para identificação das especificidades do domínio de jogos móveis, o impacto em decisões chaves dessas características e as escolhas

aplicadas, conforme a revisão bibliográfica realizada. A avaliação das especificidades em jogos móveis é realizada com base nas características de taxa de abandono e ciclo de vida, taxa de conversão e categorias de dados e ainda com relação ao modelo de negócios Freemium e o cancelamento voluntário implícito. Os detalhes estão descritos nas subseções abaixo.

4.1.1 Taxa de Abandono e Ciclo de Vida

Essas especificidades estão relacionadas, dado que a taxa de abandono define a duração média do ciclo de vida do usuário. O contexto da indústria de jogos móveis com alta taxa de abandono (85% ao mês) e consequentemente um curto ciclo de vida do jogador (1 a 3 meses) implica em restrições severas na construção do modelo preditivo. Essa curta duração do relacionamento provoca impactos diretos na configuração da janela de performance com relação ao seu tamanho e disposição. Na Tabela 4-1 apresentamos as especificidades, impactos nas decisões-chaves e escolhas identificadas na literatura. Os detalhes são discutidos em maior profundidade a seguir.

Especificidades	**Impacto em decisões-chaves**	**Escolhas identificadas na literatura**
Taxa de abandono ▪ 85% ao mês Ciclo de vida do Cliente ▪ 1 a 3 meses	Configuração da Janela de Performance	Janela Única: ▪ Não menciona ▪ 20 minutos ▪ 7 dias ▪ 15 dias ▪ 30 dias ▪ 3 meses ▪ 17 meses Janela Superposta: ▪ Não é relatada Janela Disjunta: ▪ 30 dias

Tabela 4-1: As especificidades, seus impactos em decisões-chaves e escolhas identificadas na literatura para o domínio de jogos móveis com relação à taxa de abandono e ciclo de vida.

4.1.1.1 Tamanho da Janela

Essas características do domínio dos jogos exigem a realização do balanceamento com relação ao tamanho da janela. De um lado, é importante a definição de um tamanho suficientemente longo da janela de performance para extração de dados relevantes para a classificação do usuário quanto à sua propensão ao abandono. A definição de tamanhos de janelas muito longos, entretanto, pode resultar na criação de sistemas pouco efetivos para

utilização prática na indústria. A utilização da janela de tamanho 90 dias para o domínio de jogos móveis implica que a maioria dos jogadores a serem avaliados já abandonaram o jogo dada a alta taxa de abandono neste domínio (85% ao mês). Neste caso, a aplicação do sistema preditivo em ambientes reais não fornece informação acionável dado que uma porcentagem significativa dos jogadores já abandonou o jogo dentro desse prazo.

Porém, a janela não pode ser pequena demais. Segundo pesquisas de mercado (Sonders 2016), os jogadores *hardcore* jogam por cerca de 4 horas por semana enquanto os jogadores casuais chegam a jogar por somente 30 minutos por mês, em média. A seleção de uma janela de tamanho de 1 dia implica que uma grande porcentagem de jogadores pode ter poucos eventos, ou até mesmo nenhum evento, capturado dentro dessa janela de tempo. Neste caso, a performance do modelo preditivo é impactada negativamente devido à ausência de dados em uma parcela significativa da base de usuários.

Não há uma fórmula para determinação do tamanho da janela. Segundo (Kennedy et al. 2013), a prática mais comum é a experimentação de diferentes configurações de tamanho (ex: 1, 30 e 90 dias) até a definição dos tamanhos com melhores desempenhos. Na avaliação dos artigos, identificamos trabalhos que sequer citam o tamanho selecionado para a janela de performance. Dentre os trabalhos que informam o tamanho, a divergência de valores é contundente. Há desde janelas com 20 minutos de duração até janelas com 17 meses. Não há padrão de configuração que indique a existência de melhores práticas ou diretrizes com relação à configuração da janela de performance.

4.1.1.2 Disposição da Janela

Além da configuração em termos de tamanho, a janela de performance pode ser ajustada com relação à disposição da janela. Os três tipos de disposição de janelas comumente utilizadas em problemas de Behavior Scoring são Única, Superposta e Disjunta – descritas em detalhes na seção 2.2.3.2. A análise do estado da arte permite identificar a maior adoção da **janela Única** nos trabalhos avaliados (Tabela 4-1), porém essa janela apresenta inconvenientes. A construção de dados a partir dessa disposição de janela pode apresentar os mesmos valores para usuários com comportamentos notavelmente diferentes. Os exemplos da Figura 4-1 ilustram esse comportamento a partir da distribuição da quantidade de sessões sobre o período de 16 dias para dois usuários hipotéticos.

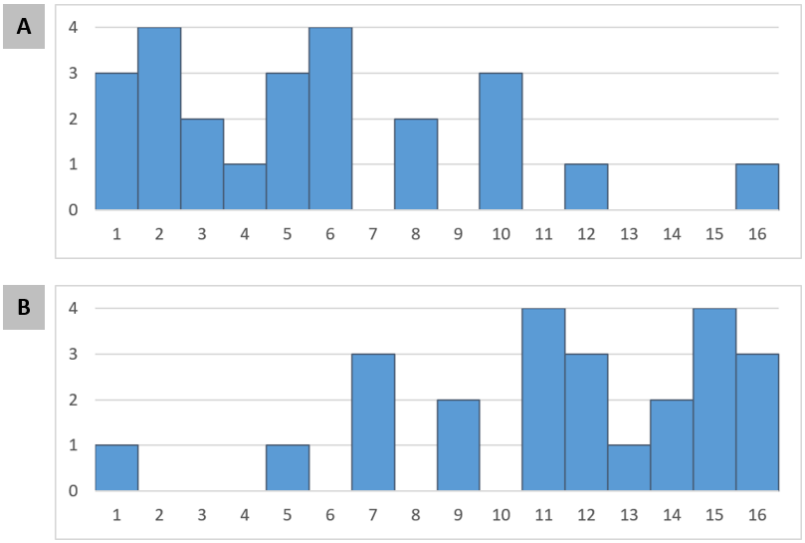

Figura 4-1: Distribuição da quantidade de sessões sobre o período de 16 dias para dois usuários hipotéticos. O eixo X representa a linha do tempo da janela de performance e o eixo Y indica o número de sessões realizadas em cada um dos dias. A) Usuário fictício nº 1. B) Usuário fictício nº 2.

A aplicação da transformação de dados RFM permite observar que ambos apresentam a mesma *recência* com sessões de jogos realizadas no último dia da janela de tempo de 16 dias considerada. Esses usuários também apresentam o mesmo valor de *frequência* pois apresentam um total de 25 sessões distribuídas pelo mesmo intervalo de tempo. Caso a duração das sessões de jogos também sejam hipoteticamente parecidas, ou ignoradas, a análise RFM desses usuários apresentará os mesmos atributos, mesmo com comportamentos completamente opostos - os exemplos são intencionalmente simétricos.

É interessante notar que, conforme estudos realizados na área (Feng et al. 2007; Chambers et al. 2010), os usuários com tendências a abandonar o jogo apresentam mudança em seu comportamento no período mais próximo ao ponto de observação. É comum identificar um aumento no tempo entre sessões e consequente redução na frequência de uso. O usuário A, portanto, apresenta uma tendência ao abandono maior do que o usuário B que reduz o tempo entre as sessões ao final da janela de performance. A partir da janela Única, entretanto, esse comportamento pode não ser identificado.

A **janela Superposta**, por outro lado, permite obter informações mais recentes do usuário ao mesmo tempo em que mantém uma visão mais global sobre o comportamento do usuário. No exemplo ilustrativo da Figura 4-1, os usuários apresentam mesma recência para as diferentes partições, por outro lado a informação de frequência e duração são distintas para as partições mais curtas (2, 4 e 8 dias).

Na **janela Disjunta,** a configuração de janela permite avaliar o comportamento do usuário sob o ponto de vista temporal. As partições apresentam dados comportamentais

referentes a momentos diferentes do ciclo de vida do usuário e fornece informação sobre a evolução da relação do usuário com o jogo.

Como visto, a janela de performance pode assumir diferentes configurações com relação ao tamanho e disposição da janela. É reconhecida a importância da experimentação nesses casos com a validação de diferentes candidatos (Kennedy et al. 2013), porém há poucos trabalhos publicados na área e nenhum consenso sobre a configuração da janela de performance para elevação da performance do modelo (Tabela 4-1).

4.1.2 Dados Dependentes da Aplicação e Taxa de Conversão

Essas especificidades estão relacionadas por impactarem diretamente nas categorias de dados disponíveis sobre o usuário e sobre seu relacionamento com o jogo a serem utilizados na construção do modelo preditivo. Em indústrias consolidadas, como telecomunicações e instituições financeiras, a revisão crítica do estado da arte revela uma grande quantidade de trabalhos publicados com a aplicação de múltiplas categorias de dados. Na maioria dos casos, os atributos construídos e transformados pertencem às mesmas categorias: cadastrais, comportamentais e financeiros. Essa informação fornece diretrizes para novos pesquisadores e profissionais da indústria construírem seus modelos para previsão de abandono nessas indústrias.

Na indústria de jogos, essas diretrizes ainda não existem. Os artigos investigados apresentam diferentes categorias de dados elaborados de maneira ad hoc para treinamento dos modelos preditivos. E mais, pouco se sabe sobre o impacto e contribuição de cada uma dessas escolhas. Dada a inexistência de diretrizes em jogos móveis, nós avaliamos as categorias de dados disponíveis à luz das diretrizes de Telecomunicações. A proposta é checar a viabilidade de aplicação de tipos de dados similares. Abaixo nós avaliamos como essas categorias podem ser replicadas em jogos móveis.

Especificidades	Impacto em decisões-chave	Escolhas Identificadas na literatura
Taxa de Conversão ▪ 2,2% Categorias de Dados ▪ Variável e Dependente da aplicação	Dados para construção das variáveis independentes	Não há dados pessoais. As pesquisas usam categorias diferentes e específicas de dados: ▪ Histórico de Uso ▪ Compras e Pagamentos ▪ Engajamento e Influência Social ▪ Específicos de Jogo ▪ RFM ▪ Entusiasmo e Persistência ▪ Análise Temporal e de Frequência

Tabela 4-2: As especificidades, seus impactos em decisões-chaves e escolhas identificadas na literatura para o domínio de jogos móveis com relação aos dados dependentes da aplicação e à taxa de conversão.

4.1.2.1 Dados Cadastrais

Os dados cadastrais comuns à área de Telecomunicações não estão disponíveis em jogos móveis. Os sistemas operacionais restringem o acesso de aplicativos a informações sobre o usuário (Boyles et al. 2012). Não é possível obter através dos sistemas operacionais (iOS e Android) informações sobre nome, sexo, idade ou qualquer outro dado sociodemográfico do proprietário do dispositivo móvel.

Uma alternativa viável consiste em obter essas informações através das redes sociais como Facebook. Nessa opção, a extração de dados do usuário demanda que o jogador realize o login no jogo através do Facebook. Essa opção, entretanto, é disponibilizada em somente uma pequena fração dos jogos móveis e mesmo os jogos com conexão ao Facebook apresentam baixa taxa de adesão, em geral em torno de 10-30% (Takahashi 2013). Em resumo, a maioria dos jogos móveis não possui acesso a informações cadastrais do usuário.

4.1.2.2 Dados Financeiros

Os dados monetários referentes ao histórico de compras e pagamentos, em geral associados ao valor do usuário (RFM), estão disponíveis para os desenvolvedores de jogos Freemium para dispositivos móveis. Todos os itens virtuais disponíveis para compra são gerenciados pelos próprios desenvolvedores de jogos que definem quais itens, e seus respectivos preços, estarão disponíveis para compra dentro do jogo. O histórico de compra pode

considerar tanto a compra de itens com moeda virtual assim como os pagamentos realizados com moeda real, via cartão de crédito.

Apesar dos dados estarem disponíveis, a realização de compra dentro de jogos para dispositivos móveis não é algo comum. Segundo pesquisas de mercado, a taxa de conversão de usuários não-pagantes a usuários pagantes é de somente 2,2% (Sinclair 2014). Os dados monetários, portanto, estarão disponíveis para somente uma parcela pequena da população implicando em desequilíbrio nos dados.

Nos jogos disponíveis sob o modelo Adware, os dados monetários estão relacionados ao faturamento do jogo de maneira global e raramente são associados aos jogadores individualmente. Nesse cenário não há como conectar o faturamento do jogo aos usuários para construção de dados a serem utilizados na modelagem do problema de previsão de abandono.

4.1.2.3 Dados Comportamentais

As informações comportamentais do usuário relativas às ações realizadas pelo usuário durante o relacionamento com o jogo estão disponíveis. A natureza digital dessa indústria associada ao alto nível de engajamento proporcionado por jogos em relação aos demais tipos de aplicativos são responsáveis pela geração de uma grande massa de dados comportamentais (Albuquerque et al. 2014).

Os tipos de dados comportamentais armazenados para análise nos jogos, entretanto, são em sua maioria variáveis e dependentes da aplicação. As ações realizadas pelo usuário em um jogo como o Pac-Man, por exemplo, estão relacionadas à movimentação do personagem (Pac-Man) sobre o labirinto para comer os pontos (*dots*) do estágio. Além disso, o usuário deve escapar dos fantasmas enquanto não está sobre o efeito das pastilhas maiores (*power pellets*) que permitem ao Pac-Man caçar os fantasmas. Nesse jogo, os dados comportamentais estão associados à movimentação do usuário, quantidade de *dots* comidos, número de *power pellets* consumidos, quantia de fantasmas caçados, número de mortes do jogador, entre outros.

Por outro lado, em um jogo como o Angry Birds, o usuário deve lançar pássaros para destruir os porcos. A dinâmica do jogo é totalmente diferente do Pac-Man e, portanto, os dados comportamentais também serão diferentes. No Angry Birds não há dados de movimentação, não há *dots* ou *powerpellets* e tampouco há fantasmas. O usuário não controla somente um personagem (Pac-Man), mas diferentes pássaros. Em resumo, os dados comportamentais gerados a partir do Angry Birds em nada se parecem com os dados provenientes do Pac-Man. Isso se deve à natureza dependente da aplicação desses dados.

Em outros domínios, como telecomunicações, os dados em geral são fixos e independentes da aplicação. Os serviços ofertados pelas empresas de telefonia se restringem quase que exclusivamente a realização de ligações, envio de mensagens de texto e à internet móvel. Isso significa que mesmo clientes com relacionamento com empresas diferentes geram dados similares entre si. Voltando ao domínio de jogos, as informações similares geradas por usuários em jogos diferentes se resumem aos dados independentes da aplicação como duração e frequência das sessões de jogos. Esses dados são passíveis de serem extraídos dos jogos móveis e auxiliam a identificar a tendência ao abandono.

A avaliação das soluções identificadas revela o uso de diferentes abordagens para tratamento dos dados e pouco conhecimento sobre seus impactos na performance da previsão (Tabela 4-2). Esse cenário reflete a ausência de diretrizes na definição dos tipos de dados com maior impacto na previsão.

4.1.3 Modelo Freemium e Cancelamento Voluntário Implícito

O modelo de negócios mais popular da indústria de jogos móveis (Freemium + Adware) oferece gratuitamente os jogos para os usuários sem amarras ou contratos envolvidos. Nesse relacionamento é dada ao usuário total autonomia para encerrar a relação a qualquer tempo, seja removendo o jogo do seu aparelho ou simplesmente não o acessando mais.

Essa especificidade implica em desafios na definição do rótulo do usuário e na configuração da janela de resultado. O método usado na maioria dos artigos avaliados consiste em classificar o usuário como *churner* na total ausência de sessões de jogo após o ponto de observação, ou seja, dentro da janela de resultado. E caso contrário, na existência de sessões de jogo após o ponto de observação, o usuário é considerado *non-churner*.

Essa definição é baseada nas informações sobre sessões de jogo obtidas na janela de resultado. Por definição, essa janela inicia-se após a janela de desempenho e estende-se pelo comprimento pré-determinado da janela de resultado. O valor a ser estipulado para essa janela apresenta impacto na performance do modelo preditivo. A definição de uma janela curta, por exemplo 7 dias, implica na classificação de usuários ausentes por mais de 7 dias como churners, mesmo aqueles que retornam após esse período. Essa configuração relaxa o conceito tradicional de abandono definitivo e o redefine para redução do engajamento. Dessa maneira o classificador é treinado para identificar a diminuição de engajamento e não necessariamente o abandono.

Por outro lado, a definição de uma janela longa, por exemplo 12 meses, acarreta em classificar como churners somente os usuários que não retornaram ao jogo por 1 ano. Esse prazo

implica na necessidade de obtenção de 14 a 15 meses de dados de comportamento dos usuários de um jogo para construção do modelo preditivo de abandono. Na indústria de jogos móveis, entretanto, os jogos são abandonados antes desse prazo caso o ROI não seja positivo. Isso significa que esse prazo não permite a extração de informação acionável para os momentos mais crítico da vida do jogo. Além disso, os usuários com baixa taxa de engajamento, digamos aqueles que retornam a cada 1 ou 2 meses, não são identificados como críticos por esse sistema preditivo.

A definição do tamanho da janela de resultado demanda o equilíbrio entre esses aspectos, entretanto esse não é o cenário percebido a partir da revisão bibliográfica (Tabela 4-2). Os trabalhos identificados apresentam configurações distintas para a janela de resultado com artigos sequer mencionando o tamanho da janela. Dentre os que apresentam a configuração percebe-se a aplicação de tamanhos variáveis, entre 40 minutos, 30 dias e até mesmo 5 meses.

4.2 As Escolhas a Serem Avaliadas Experimentalmente

A avaliação das especificidades e seus respectivos impactos na construção de modelos para previsão de abandono em jogos móveis permitem identificar a aplicação de soluções ad hoc e a ausência de relação entre as decisões de projeto tomadas nos trabalhos avaliados. Nesta seção nós elencamos as principais decisões na construção de um modelo preditivo para abandono em jogos. Com base nestas decisões, faremos um projeto experimental (explicado no próximo capítulo) que permitirá empiricamente apontar as melhores escolhas, constituindo, assim, um conjunto de diretrizes mais bem fundamentadas do que as decisões predominantemente ad hoc atuais.

4.2.1 Configuração da Janela de Performance

A configuração da janela de performance deve considerar os aspectos tamanho e disposição. Em termos de tamanho, a janela de performance deve ser suficientemente curta para viabilizar a realização de ações para evitar o abandono dos jogadores e também adequadamente longa para capturar eventos relevantes do usuário.

Para definir o tamanho da janela nós analisamos mais profundamente as métricas de mercado com relação ao abandono. Essa avaliação revela que apesar da retenção média de 30 dias ser 15%, a curva de redução da retenção com relação ao tempo não é linear (Walz 2015). Ao final da primeira semana a retenção já cai para 35% e após a segunda semana a retenção

chega a 22% conforme a Figura 4-2. A queda drástica da primeira semana inclui a taxa de rejeição, ou *bounce rate*, referente aos usuários que acessaram o jogo uma única vez e em seguida abandonaram, sem continuar a utilização (Peyton 2014).

Essa avaliação indica que quanto menor a janela de performance maiores são as chances de encontrar o usuário ainda presente no jogo e aplicar ações de retenção e fidelização do usuário. Nesse cenário nós propomos a configuração de janela com tamanho máximo de 16 dias para averiguar a possibilidade de identificação com alta precisão dos usuários com propensão ao abandono dentro desse prazo.

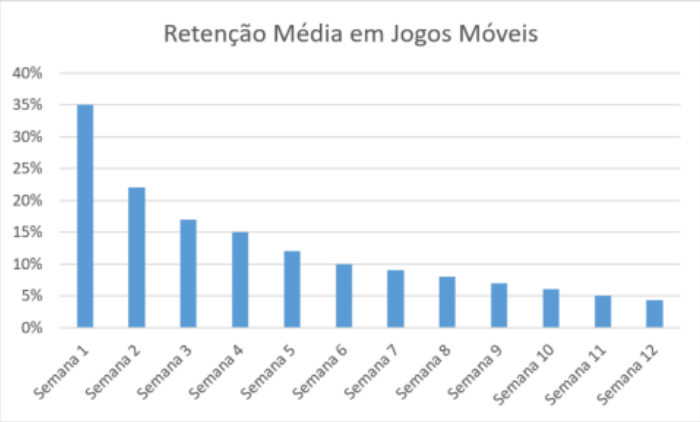

Figura 4-2: Taxa de decaimento da retenção com relação ao tempo.

Com relação à disposição da janela, nós propomos a aplicação das três variações apresentadas para identificação do efeito impacto de cada uma das variações na performance do modelo preditivo. Como as diferentes disposições das janelas demanda uma formatação própria, nós propomos a seguinte configuração:

- **Janela Superposta**: Para esse tipo, nós propomos a escolha de tamanhos pertencentes à progressão geométrica com elemento inicial 2 e quociente 2. A janela apresenta partições com tamanhos 2, 4, 8 e 16 dias.
- **Janela Disjunta**: Para esse tipo, nós propomos a escolha de tamanhos pertencentes à progressão aritmética com elemento inicial 2 e razão 2. A janela apresenta partições com tamanhos 2, 4, 6, 8, 10, 12, 14 e 16 dias.
- **Janela Única:** Nesse caso, o tamanho da janela é definido com base no tamanho máximo assumido pela janela de performance, ou seja, 16 dias.

A proposta é que a aplicação dessas combinações na construção dos modelos preditivos nos permita analisar o impacto dessas decisões na performance do modelo preditivo e também avaliar a interação existente entre a configuração da janela de desempenho e as demais decisões de projeto.

4.2.2 Tipos de Dados

O domínio de jogos móveis impõe restrições aos tipos de dados disponíveis. Os dados cadastrais são inexistentes e os dados monetários estão presentes somente para uma pequena parcela dos usuários, no caso de jogos Freemium. E esses dados inexistem para os jogos disponibilizados sob o modelo Adware. Os dados comportamentais, por outro lado, são amplamente utilizados no domínio de jogos móveis. Em geral os artigos avaliados aplicam desde dados independentes do jogo (ex: quantidade e frequência de sessões) como também dados dependentes do jogo (ex: quantidade de missões cumpridas e inimigos mortos (Srivastava 2011) ou convites enviados e precisão de movimentos (Runge 2014). A avaliação do impacto de dados específicos dos jogos, entretanto, está além do escopo da tese a qual visa identificar diretrizes passíveis de serem aplicadas em jogos independentemente do gênero e categoria do jogo móvel.

Para avaliação do impacto dos tipos de dados, na construção dos modelos preditivos de abandono, nós propomos a aplicação de dois tipos dados referentes à Análise RFE e Análise RFM. A análise RFE é uma variação da análise RFM comumente utilizada na análise do comportamento do consumidor em estratégias de marketing. Nessa versão modificada, o “M” referente à Valor Monetário é substituído por “E” referente à Engajamento. Em jogos móveis, o engajamento pode ser medido por duração das sessões de jogo, frequência de acesso, dentre outros tipos de dados possíveis. A lista abaixo de atributos representa os possíveis atributos para a Análise RFE.

- **Número total de sessões:** Número total de sessões de jogo do usuário dentro da janela de desempenho.
- **Duração total das sessões:** Duração total das sessões de jogo do usuário dentro da janela de desempenho. Esse atributo indica todo o tempo dedicado pelo usuário ao jogo.
- **Duração média das sessões**: Duração média das sessões de jogo do usuário dentro da janela de desempenho.

- **Duração da ausência:** Intervalo de tempo entre a última sessão de jogo do usuário dentro da janela de desempenhoe o ponto de observação.
- **Frequência de sessões (por dia):** Número médio de sessões de jogo do usuário, por dia, dentro da janela de desempenho.
- **Número de dias desde a última sessão de jogo:** Essa variável calcula o tempo de ausência de jogo, em dias, desde a última sessão.
- **Tempo médio entre as sessões:** Intervalo de tempo médio entre sessões de jogo do usuário dentro da janela de desempenho.
- **Tempo total de vida:** Essa variável indica a quantidade de dias desde a primeira sessão do jogo do usuário até o ponto de observação.

A análise RFM, por outro lado, incorpora todas as informações presentes na análise RFE e inclui ainda dados relativos ao comportamento de compra dos usuários. Os atributos abaixo foram projetados para fornecer informações históricas sobre as transações dos jogadores.

- **Quantidade total de compras:** Número total de sessões de jogo do usuário dentro da janela de performance.
- **Número de dias desde a última compra:** Essa variável calcula o intervalo de tempo desde a última compra realizada.
- **Frequência de compras (por dia):** Número médio de compras do usuário, por dia, dentro da janela de performance.

A proposta é que a aplicação desses dois tipos de dados represente semanticamente as decisões usualmente aplicadas nos artigos pesquisados para avaliação do impacto dos tipos de dados na construção dos atributos independentes usados para treinamento dos modelos preditivos de abandono.

4.2.3 Configuração da Janela de Resultado

A definição do prazo trata-se de uma questão de equilíbrio com relação ao tamanho da janela de resultado. As janelas curtas permitem a identificação de perda de engajamento, mas não necessariamente o abandono e si. Já as janelas mais longas tendem a se tornar pouco úteis em problemas reais como descrito na seção 4.1.3. Nós optamos por escolher o padrão utilizado para indústria. De acordo com as pesquisas realizadas (Lovell 2010; Lovell 2011), é comum a

aplicação do prazo de 1 mês para classificação do usuário como churner em caso de ausência de sessões de jogo. Caso o usuário retorne após esse prazo, as empresas consideram o jogador como um novo usuário e armazenam as informações relativas a esse novo período como um novo relacionamento. Nesse cenário, nós propomos o prazo de 30 dias como tamanho da janela de resultado para extração das informações e definição do rótulo do jogador.

4.3 Conclusão e Observações

A avaliação das especificidades do domínio evidencia desafios únicos relacionados ao domínio de jogos móveis. A avaliação das soluções aplicadas na indústria de jogos revela a inexistência de diretrizes para resolução desses desafios. Na busca de soluções candidatas às diretrizes, nós avaliamos as soluções aplicadas em outros domínios, especialmente telecomunicações, e também pesquisamos alternativas aplicadas na indústria de jogos. A partir dessas informações nós propusemos uma série de diretrizes para as decisões de construção do modelo preditivo de abandono em jogos.

5 Planejamento dos Experimentos

A identificação das especificidades de jogos móveis levanta questões importantes relativas ao tratamento dessas características únicas do domínio: como essas características podem ser abordadas no processo de construção do modelo preditivo? Qual o peso dessas características, e seus possíveis tratamentos, na performance da previsão? A solução proposta é replicável para outros jogos dentro do mesmo domínio? As respostas a essas perguntas são fundamentais para fornecer diretrizes relativas à construção de soluções para previsão de abandono nesse domínio.

Com o propósito de responder essas questões, nós propomos neste capítulo a condução de um projeto experimental em uma base de dados real para avaliação dos possíveis tratamentos aos desafios identificados no domínio e seus respectivos efeitos na performance do modelo preditivo. No início do capítulo nós apresentamos os conceitos e fundamentos da área de projetos experimentais com os respectivos termos, técnicas e processos usados para planejamento e execução. Em seguida, nós detalhamos o projeto experimental e os tratamentos a serem aplicados aos desafios encontrados com base nas escolhas possíveis para as decisões-chave de construção do modelo preditivo, conforme discutido no capítulo anterior.

5.1 Fundamentos de Projetos Experimentais

Segundo (Kinnear et al. 1991), um experimento é “um tipo de pesquisa científica no qual o pesquisador manipula e controla uma ou mais variáveis independentes e observa a variação nas variáveis dependentes concomitantemente à manipulação das variáveis independentes”.

O objetivo da condução dos experimentos, através da manipulação e medição das variáveis, é captar a causalidade. As variáveis independentes são responsáveis pelas possíveis

causas, e as variáveis dependentes sinalizam os efeitos. Os experimentos consistem em procedimentos no qual alterações propositais são feitas nas variáveis de entrada de um processo de modo que se possa avaliar as possíveis alterações sofridas na variável resposta, como também as razões de sua alteração (Figura 5-1).

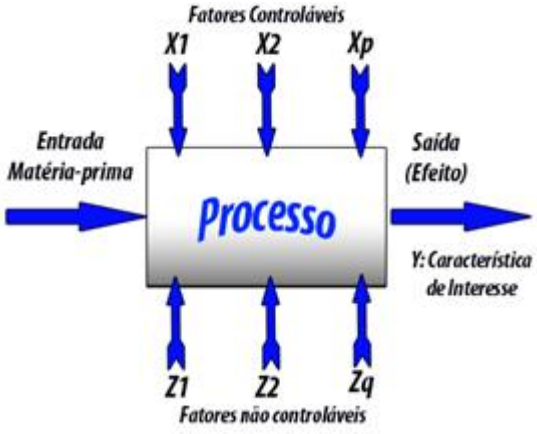

Figura 5-1: Esquema representativo de um experimento.

A performance de um modelo preditivo de abandono está condicionada a inúmeros fatores como a qualidade dos dados, a forma como os dados foram tratados, a configuração das janelas de performance e resultado, as técnicas de modelagem do problema, dentre inúmeros outros aspectos. Uma análise adequada, através de projeto experimental, requer que os efeitos de cada um desses fatores sejam isolados dos demais de modo que conhecimento relevante sobre a construção do modelo preditivo possa ser descoberto.

Por motivos didáticos, nós iremos considerar o seguinte projeto experimental como exemplo: construção de um modelo preditivo na qual diferentes ações podem ser realizadas durante o processo de D3M e BS. Em primeiro lugar, o tamanho da janela de observação deve ser determinado. As alternativas, usadas como exemplo, serão 30 e 60 dias. Em segundo lugar, os tipos de dados a serem considerados na modelagem do classificador podem ser pessoais (sociodemográficos) ou comportamentais. E, por fim, a técnica inteligente de modelagem pode variar entre Redes Neurais e Árvore de Decisão.

Os termos a seguir são comumente utilizados na concepção, execução e avaliação de projetos experimentais:

- **Variável Resposta:** A saída de um experimento é chamada de variável resposta. Geralmente essa variável é a medida de performance do sistema. No exemplo

considerado, a variável resposta pode ser a precisão da classificação ou outra métrica de performance;

- **Fatores**: As variáveis que afetam a variável resposta e podem assumir diferentes valores são chamadas de fatores. Há 3 fatores no exemplo dado referentes ao tamanho da janela, tipos de dados e técnica de modelagem;
- **Níveis do Fator**: Os valores ou estados que os fatores podem assumir nos experimentos são chamados de níveis. De volta ao exemplo, o fator tamanho da janela de observação apresenta dois níveis: 30 e 60 dias;
- **Tratamentos**: As diferentes combinações possíveis entre os níveis de cada fator são chamadas de tratamentos. A combinação dentre janela de observação (30 dias), tipo de dados (comportamentais) e técnica de modelagem (Redes Neurais) é um tratamento do projeto experimental proposto como exemplo;
- **Interação:** A interação entre dois fatores existe quando o efeito de um depende do nível assumido pelo outro fator.

Além da terminologia e definições básicas relativas ao projeto experimental, apresentamos a seguir os tipos de projetos existentes além das técnicas utilizadas para análise dos experimentos.

5.1.1 Princípios dos Experimentos

A condução dos experimentos será realizada tendo como base 2 princípios básicos (Fisher Ronald 1935).

- **Replicação:** As replicações são várias execuções experimentais realizadas com as mesmas configurações de fatores. As replicações estão sujeitas às mesmas fontes de variabilidade, independentemente umas das outras. A realização de experimentos com réplicas é muito importante por dois motivos. O primeiro é que isto permite obter o erro experimental e assim verificar se as diferenças observadas nos dados são estatisticamente diferentes. O segundo motivo se refere ao fato de que, se a média de uma amostra for usada para estimar o efeito de um fator no experimento, a replicação permite a obtenção de uma estimativa mais precisa desse efeito.
- **Aleatoriedade:** Os métodos estatísticos requerem que as observações, ou os erros, sejam variáveis aleatórias distribuídas independentemente. Os experimentos, com suas

réplicas, devem ser realizados de forma aleatória, de modo a garantir a distribuição equânime de todos os fatores não considerados. Por exemplo, na realização de um experimento para determinar as variáveis estatisticamente mais significantes na previsão de abandono de um usuário, deve-se atentar para a aleatoriedade na execução do experimento, pois fatores críticos que não estão no estudo, como versão do jogo e época do ano, podem influenciar as variáveis de interesse de forma diferenciada, o que compromete a independência e a variabilidade entre os erros experimentais.

Esses princípios permitem avaliar o erro experimental, reduzir o impacto de fatores externos não controláveis e oferecer proteção contra vícios no experimento.

5.1.2 Tipos de Projetos Experimentais

Há inúmeras maneiras de planejar e executar projetos experimentais. As técnicas mais frequentemente utilizadas consistem nos experimentos do tipo Simples e Fatorial Completo.

- **Simples**

Nesse tipo de projeto experimental, a execução dos experimentos é iniciada com uma configuração típica a partir da qual varia-se um fator de cada vez para identificar como esse fator afeta a performance.

No exemplo considerado, a configuração típica pode ser definida como o modelo preditivo construído com a janela de observação com tamanho de 30 dias a partir dos dados pessoais (sociodemográficos) com a técnica de modelagem de Redes Neurais. A performance dessa configuração é medida e posteriormente varia-se o primeiro fator - tamanho da janela - e compara-se a performance dos dois tratamentos, mantendo-se os demais fatores constantes. E assim determina-se qual tamanho de janela proporciona a melhor performance. De maneira similar é possível comparar as técnicas de modelagem ao variar entre Redes Neurais e Árvore de Decisão para comparar a performance e encontrar o valor (sub-)ótimo.

Nesse caso, um projeto experimental com k fatores, com o i-ésimo fator contendo n_i níveis, o projeto experimental simples requer somente a execução de n experimentos, onde:

$$n = 1 + \sum_{i=1}^{k} (n_i - 1)$$

No exemplo considerado, o número de experimentos seria:

$$n = 1 + (2\ tamanhos\ de\ janelas - 1) + (2\ tipos\ de\ dados - 1) + (2\ técnicas - 1)$$

$$n = 4$$

Esse tipo de projeto experimental, entretanto, não é estatisticamente eficiente. Além disso, o projeto experimental simples não permite calcular o efeito da interação entre fatores. Nesse caso, se os fatores apresentarem interação entre si o projeto pode levar a conclusões falsas. Esse tipo de projeto, portanto, é pouco recomendado e utilizado na prática.

- **Fatorial Completo**

Um Experimento Fatorial Completo inclui todas as possíveis combinações entre os níveis dos fatores do experimento. Nesse caso, um projeto experimental com k fatores, com o i-ésimo fator contendo n_i níveis, requer a execução de n experimentos, onde:

$$n = \prod_{i=1}^{k} n_i$$

No exemplo considerado, o número de experimentos seria:

$$n = (2\ tamanhos\ de\ janelas) \times (2\ tipos\ de\ dados) \times (2\ técnicas)$$

$$n = 8$$

A vantagem desse tipo de experimento é que todas as possíveis combinações entre os níveis dos fatores são examinadas. Dessa maneira é possível identificar o efeito de cada fator, de cada nível e também das possíveis interações. O principal problema aqui é o custo do estudo. É possível que a condução dos experimentos com todas as combinações consuma muitos recursos, seja tempo ou dinheiro.

5.1.3 Análise de Experimentos

A análise do resultado dos experimentos é comumente realizada através da Análise de Variância (ANOVA) e Análise de Regressão. Esses métodos de análise são apresentados em detalhes a seguir. Para ilustrar como as análises funcionam e detalhar a explicação, nós estendemos o exemplo ilustrativo e apresentamos todas as combinações possíveis para um experimento Fatorial Completo com as respectivas performances referente à precisão da previsão de abandono. As performances, assim como o experimento, são ilustrativas.

Tamanho da Janela	Tipos de Dados	Técnica de Modelagem	Performance
30	Pessoais	Rede Neural	0,81
30	Pessoais	Árvore de Decisão	0,72
30	Comportamentais	Rede Neural	0,93
30	Comportamentais	Árvore de Decisão	0,76
60	Pessoais	Rede Neural	0,82
60	Pessoais	Árvore de Decisão	0,73
60	Comportamentais	Rede Neural	0,95
60	Comportamentais	Árvore de Decisão	0,77

Tabela 5-1: Performance para os tratamentos em um experimento Fatorial Completo (exemplo ilustrativo).

Este exemplo ilustrativo dos experimentos, sem repetição, será utilizado para demonstração dos resultados obtidos com as análises e também da interpretação das saídas.

5.1.3.1 Análise de Variância (ANOVA)

A partir da análise de variância é possível quantificar a parcela da variabilidade total que é devida a cada fator e à interação entre eles. A análise visa, fundamentalmente, verificar se existe uma diferença significativa entre as médias e se os fatores exercem influência na variável resposta.

A análise de variância baseia-se na decomposição da variação total da variável resposta em partes que podem ser atribuídas aos tratamentos e ao erro experimental. Em outras palavras, a análise de variância é utilizada quando se quer decidir se as diferenças amostrais observadas são reais (causadas por diferenças significativas nas populações observadas) ou casuais (decorrentes da mera variabilidade amostral). Portanto, essa análise parte do pressuposto que o acaso só produz pequenos desvios, sendo as grandes diferenças geradas por causas reais. Os pressupostos básicos da análise de variância são:

- As amostras são aleatórias e independentes.
- As populações têm distribuição normal (o teste é paramétrico).
- As variâncias populacionais são iguais.

Na prática, esses pressupostos não precisam ser todos rigorosamente satisfeitos. Os resultados são empiricamente verdadeiros sempre que as populações são aproximadamente normais (isto é, não muito assimétricas) e têm variâncias próximas. O resultado da aplicação da ANOVA para o exemplo considerado (Tabela 5-1) é apresentado a seguir (Tabela 5.2).

Análise de Variância

Fonte	GL	SQ	Contribuição	Valor F	Valor-P
Modelo	6	0.052675	99.98%	702.33	0.029
Linear	3	0.049038	93.07%	1307.67	0.020
tamanho da janela	1	0.000312	0.59%	25.00	0.126
tipos de dados	1	0.013613	25.84%	1089.00	0.019
técnica de modelagem	1	0.035112	66.64%	2809.00	0.012
Interações de 2 fatores	3	0.003637	6.90%	97.00	0.074
tamanho da janela*tipos de dados	1	0.000012	0.02%	1.00	0.500
tamanho da janela*técnica de modelagem	1	0.000012	0.02%	1.00	0.500
tipos de dados*técnica de modelagem	1	0.003612	6.86%	289.00	0.037
Erro	1	0.000013	0.02%		
Total	7	0.052687	100.00%		

Tabela 5-2: Resultado da Análise de Variância (exemplo ilustrativo).

A ANOVA apresenta os resultados para os fatores individuais, além das possíveis interações entre fatores. A avaliação dos resultados apresentados é realizada através das variáveis descritas abaixo.

- **GL**: Os graus de liberdade são a quantidade de informação disponível a serem utilizados para estimar os valores de parâmetros populacionais desconhecidos e calcular a variabilidade dessas estimativas.
- **SQ:** A soma dos quadrados representa uma medida de variação ou desvio da média. Esse valor é calculado como a soma dos quadrados das diferenças da média. O cálculo da soma total dos quadrados inclui a soma dos quadrados dos fatores e aleatoriedade ou erro.
- **Contribuição**: A importância de um fator é medida pela proporção da variação total na resposta que é explicada pelo fator. Dessa forma, se dois fatores explicam 90 e 5% da

variação de uma resposta, o segundo fator pode ser considerado sem importância em muitas situações práticas.

- **Valor-F**: Essa métrica é utilizada para comparação de duas variâncias. As variâncias são uma medida da dispersão e indicam até que ponto os dados estão dispersos da média. Valores maiores indicam maior dispersão dos dados.
- **Valor-P**: O valor-p representa a chance, ou probabilidade, do efeito observado entre os fatores ser devido ao acaso, e não aos fatores sendo estudados. Em termos gerais, um valor-p pequeno significa que a probabilidade de obter um valor da estatística de teste como o observado é muito improvável, levando assim à rejeição da hipótese nula (H_0). Nós podemos definir o valor-p como a menor escolha que teríamos feito para o nível de significância (α), de forma que rejeitaríamos H_0. Em muitas aplicações da estatística, o nível de significância é tradicionalmente fixado em 0,05.

No exemplo, o resultado revela que o fator Tamanho da Janela não é estatisticamente significativo assim como as suas interações com os demais fatores (Valor-P > 0,05). O fator Técnica de Modelagem surge como fator o mais importante ao apresentar 66,64% de contribuição na alocação da variação. O fator Tipos de Dados e a interação com a Técnica de Modelagem apresentam 25,84% e 6,86% de contribuição, respectivamente. A análise dos gráficos de efeito, resultantes da ANOVA, também explicita essas informações (Figura 5-2 e Figura 5-3).

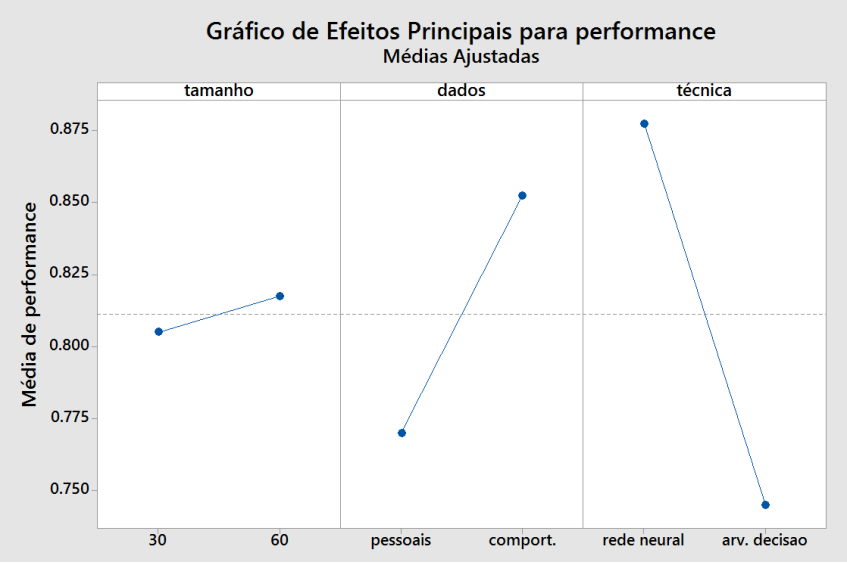

Figura 5-2: Gráficos de efeitos principais com apresentação das médias de performance ajustadas.

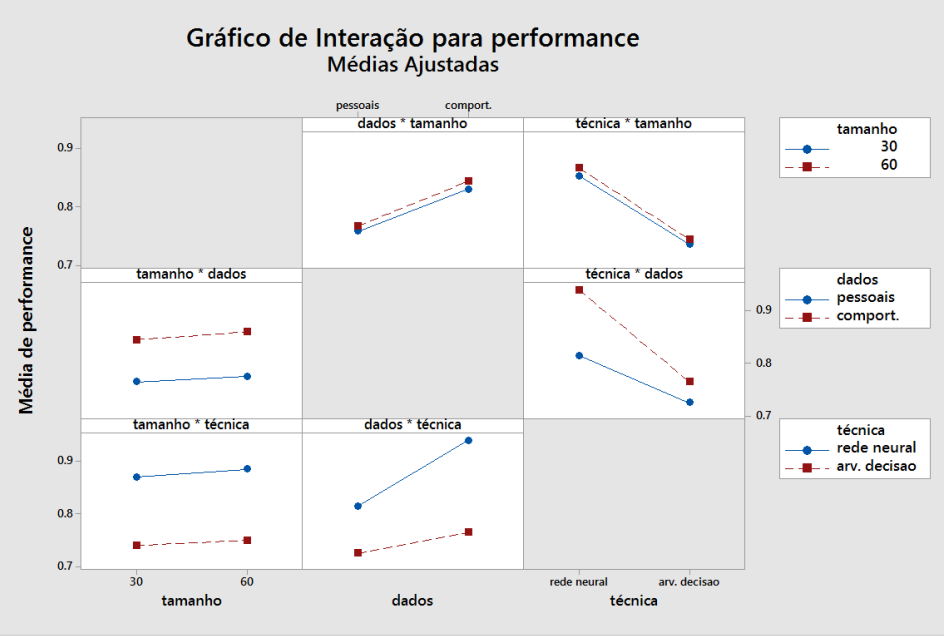

Figura 5-3: Gráficos de efeitos das e interação com apresentação das médias de performance ajustadas

O gráfico de efeitos principais demonstra a maior variação nas médias do fator Técnica de Modelagem e em seguida, com menor variação, para o fator Tipos de Dados. Não por acaso esses dois fatores apresentam maior contribuição na ANOVA (Tabela 5-2). O gráfico de interação mostra a relação mais expressiva entre os fatores Tipos de Dados e Técnica de Modelagem.

5.1.3.2 Análise de Regressão

Uma outra forma de analisar os efeitos dos fatores e das interações é através da análise de regressão. Nessa análise é definido um modelo de regressão linear para modelar a relação entre a variável resposta (variável dependente) e os fatores (ou variáveis independentes). Por exemplo, ao considerarmos um experimento com somente fatores A e B, o modelo de regressão linear é definido da seguinte forma.

$$Y = \beta_0 + \beta_1 X_i + \beta_2 X_j + \beta_{12} X_i X_j + \varepsilon$$

Em que:

β_0 é a média geral da resposta

X_i assume valor -1 ou 1, dependendo do nível do fator A.

X_j assume valor -1 ou 1, dependendo do nível do fator B.

As constantes desconhecidas β_j são denominadas parâmetros e ε representa o erro experimental, isto é, a variabilidade devido a fatores aleatórios não controlados no experimento. O método dos mínimos quadrados é o mais utilizado para estimar os parâmetros do modelo de regressão linear através da minimização da soma dos resíduos quadrados.

A derivação da equação e estimação dos parâmetros permite quantificar a força da relação entre Y e os parâmetros X_n , avaliar qual X_n pode não ter nenhuma relação com Y e para identificar quais subconjuntos do X_n contêm informações redundantes sobre Y. Em outras palavras, a análise de regressão permite compreender quais dentre as variáveis independentes estão relacionadas às variáveis dependentes e explicar as formas desse relacionamento. Essas características da regressão são fundamentais para a avaliação dos experimentos.

O resultado da aplicação da Análise de Regressão para o exemplo considerado (Tabela 5.1) é apresentado a seguir (Tabela 5-3). O sumário do modelo indica que o coeficiente de determinação (R^2)é 99,83%. Esse coeficiente indica que 99,83% da alocação da variação é explicada pelos fatores e suas interações.

```
Sumário do Modelo

        S      R2  R2(aj)   PRESQ  R2(pred)
0.0035355  99.98%  99.83%  0.0008    98.48%



Coeficientes Codificados

Termo                Coef        IC de 95%       Valor-T  Valor-P   VIF
Constante         0.81125  ( 0.79537,  0.82713)   649.00    0.001
tamanho           0.00625  (-0.00963,  0.02213)     5.00    0.126  1.00
dados             0.04125  ( 0.02537,  0.05713)    33.00    0.019  1.00
técnica          -0.06625  (-0.08213, -0.05037)   -53.00    0.012  1.00
tamanho*dados     0.00125  (-0.01463,  0.01713)     1.00    0.500  1.00
tamanho*técnica  -0.00125  (-0.01713,  0.01463)    -1.00    0.500  1.00
dados*técnica    -0.02125  (-0.03713, -0.00537)   -17.00    0.037  1.00

Equação de Regressão em Unidades Não codificadas

performance = 0.79250 + 0.000417 tamanho + 0.03750 dados - 0.06250 técnica
              + 0.000083 tamanho*dados - 0.000083 tamanho*técnica - 0.02125 dados*técnica
```

Tabela 5-3: Resultado da Análise de Regressão para o exemplo ilustrativo.

Os coeficientes da regressão identificam a direção, o tamanho e a significância estatística da relação entre um preditor e a resposta.

- O sinal dos coeficientes indica a direção da relação.

- Os coeficientes representam a mudança média na resposta para uma unidade de mudança no preditor mantendo os outros preditores no modelo constantes.
- O Valor-P para cada coeficiente testa a hipótese nula de que o coeficiente é igual a zero (sem efeito). Portanto, Valores-P baixos indicam que o preditor é uma adição significativa ao modelo.

O Fator de Inflação de Variância (VIF) é usado para descrever a multicolinearidade (correlação entre preditores) existente em uma análise de regressão. O VIF mede o quanto a variânvia de um coeficiente de regressão estimado aumenta se seus preditores estão correlacionados. A multicolinearidade é problemática porque pode aumentar a variância dos coeficientes de regressão tornando-os instáveis e difíceis de interpretar. É comum o uso das seguintes orientações:

- $VIF = 1$: Não correlacionados
- $1 < VIF < 5$: Moderadamente correlacionados
- $VIF > 5$: Altamente correlacionados

5.2 Plano Experimental

O planejamento, execução e avaliação do projeto experimental é executado através das etapas propostas por (Coleman & Montgomery 1993).

1) Caracterização do Problema
2) Escolha dos fatores de influência e níveis
3) Seleção das variáveis de resposta
4) Determinação de um modelo de planejamento de experimento
5) Condução do experimento
6) Análise dos dados
7) Conclusões e recomendações

As etapas 1 a 4 iniciais referem-se à etapa de planejamento dos experimentos e são detalhadas nesta seção. As etapas 5 e 6 pertencem à execução do experimento e estão descritas

no próximo capítulo. A etapa 7 referente às conclusões e recomendações está descrita no capítulo 7.

5.2.1 Caracterização do Problema

A condução dos experimentos visa identificar diretrizes relativas à construção de modelos preditivos que ajudarão desenvolvedores e pesquisadores a criar melhores soluções, já que hoje as decisões são ad hoc. As diretrizes auxiliarão na estruturação do conhecimento no domínio da predição de abandono de usuários na indústria de jogos free-to-play móveis ao responder às seguintes perguntas:

- Como essas características podem ser tratadas no processo de construção do modelo preditivo?
- Qual o peso dessas características, e seus possíveis tratamentos, na performance da previsão?
- A solução proposta é replicável para outros jogos dentro do mesmo domínio?

A proposta é que a avaliação dos resultados dos experimentos capture o conhecimento relevante para embasar as principais etapas da construção dos modelos preditivos na indústria de jogos móveis. Essa identificação dos melhores tratamentos e dos fatores com maior relevância auxiliará a responder esses questionamentos.

5.2.2 Escolha dos Fatores de Influência e Níveis

A seleção dos fatores e seus respectivos níveis é realizada com base nas escolhas propostas para a decisões-chaves vistas no capítulo anterior. Essas escolhas são as candidatas a diretrizes para a resolução das especificidades identificadas no domínio de jogos móveis. As decisões-chaves são transformadas nos fatores dos experimentos e as possíveis escolhas nos níveis dos fatores.

5.2.2.1 Tamanho da Janela

A janela de desempenho pode assumir diferentes configurações com relação ao tamanho e disposição da janela. Os respectivos níveis relativos ao tamanho da janela são apresentados abaixo.

Fator: Tamanho da Janela (Δt)	
Nível #1:	2 dias
Nível #2:	4 dias
Nível #3:	8 dias
Nível #4:	16 dias

Tabela 5-4: Apresentação dos níveis do fator Tamanho da Janela de Performance.

5.2.2.2 Disposição da Janela

Nós propomos os seguintes níveis para o fator disposição da janela.

Fator: Disposição da Janela	
Nível #1:	Única
Nível #2:	Superposta
Nível #3:	Disjunta

Tabela 5-5: Apresentação dos níveis do fator Disposição da Janela de Performance.

A escolha desses níveis permite o efeito do uso dessa janela na performance do modelo preditivo e também avaliar a interação existente entre esses níveis e os demais fatores avaliados no projeto experimental.

5.2.2.3 Tipos de Dados

Os tipos de dados propostos referem-se aos dados associados à análise RFE e também à análise de frequência. Essas candidatas são propostas como níveis de fatores do experimento.

Fator: Tipos de Dados	
Nível #1:	Análise RFE
Nível #2:	Análise RFM

Tabela 5-6: Apresentação dos níveis do fator Tipos de Dados.

5.2.2.4 Técnica de Modelagem

A etapa de modelagem do problema é alvo de parte significativa dos estudos identificados na revisão bibliográfica. As abordagens acadêmicas para construção de modelos preditivos costumam focar na apresentação de novas técnicas e comparação com abordagens

consideradas tradicionais. As pesquisas, entretanto, não investigam o impacto das técnicas na performance do modelo preditivo em comparação com as outras etapas do processo CRISP-DM.

Nós propomos a inclusão desse fator para análise do impacto e contribuição da etapa de modelagem na performance dos modelos preditivos. Os níveis propostos consistem em três principais técnicas de modelagem utilizadas na academia (Ngai et al. 2009) para avaliação do impacto dessas técnicas e relevância com relação à performance do modelo preditivo. Nesse cenário, nós propomos os seguintes níveis para o fator técnica de modelagem.

Fator: Técnica de Modelagem	
Nível #1:	**Regressão Logística** A técnica aplicada é baseada na implementação da Regressão Logística proposta por (Keerthi et al. 2005). Configuração: kernel = dot, C=1, épsilon=0,001, máximo número de iterações: 100000.
Nível #2:	**Árvore de Decisão** A construção da árvore de decisão usa o algoritmo C4.5, uma extensão do ID3, proposto por (Quinlan 1993). Configuração: critério = razão de ganho, máxima profundidade = 20, registros mínimos por folha = 2, registros mínimos para Split = 4.
Nível #3:	**Redes Neurais** O modelo preditivo é criado a partir de Redes Multi Layer Perceptron treinadas com Backpropagation. Configuração: camadas escondidas = 1, número de neurônios = 8, taxa de aprendizado = 0,3, momentum = 0,2, ciclos de treinamento = 500.

Tabela 5.10: Apresentação dos níveis do fator Técnica de Modelagem.

5.2.3 Seleção da Variável de Resposta

A variável de resposta do projeto experimental proposto consiste na performance do modelo preditivo a ser medida através da variável Área sob a Curva ROC (AUC) (Provost & Fawcett 2001). A AUC é um indicador de performance comumente usado na academia para problemas de classificação binária. Essa métrica é robusta com relação a bases de dados desbalanceadas e também ao custo de classificação incorreta, diferentemente da métrica precisão.

Além disso, a AUC permite evitar a avaliação do possível impacto na escolha do limiar para a decisão binária. Esse limiar é aplicado nas técnicas com saídas contínuas e depende de vários fatores como o custo associado aos dois tipos de erros de decisão, são eles: o erro Tipo I referente a situação em que um jogador *churner* é classificado como *non-churner* e o erro Tipo

II referente a quando um jogador *non-churner* é classificado como *churner*. A seleção da métrica AUC nos permite abstrair esse conceito.

Essa métrica é calculada a partir da Curva ROC por meio da integração numérica da curva ou ainda através do método dos trapézios. Essa curva é uma representação gráfica do desempenho de um classificador binário obtida ao desenhar um diagrama que representa a taxa de verdadeiros positivos (chamada de $sensibilidade$) em função da proporção de falsos positivos (chamada de $1 - especificidade$) para todos os possíveis pontos de corte (limiares) entre 0 e 1. Na curva ROC, um classificador perfeito corresponde a uma linha horizontal no topo do gráfico, porém esta dificilmente é alcançada. A AUC nesse cenário é igual a 1. Na prática, curvas consideradas boas estarão entre a linha diagonal (AUC=0,5) e a linha perfeita (AUC=1,0), onde quanto maior a distância da linha diagonal, melhor o sistema. A linha diagonal indica uma classificação aleatória.

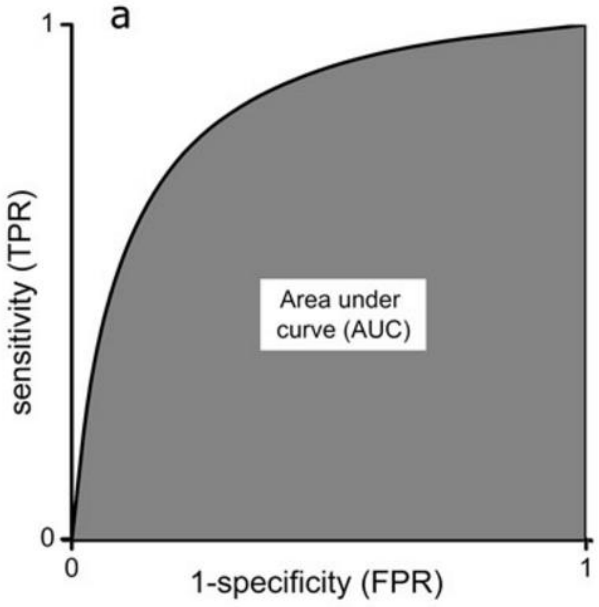

Figura 5-4: Curva ROC e área sob a Curva ROC (AUC).

A determinação da variável de resposta é realizada através da construção do modelo preditivo com base no respectivo tratamento sob análise e execução desse classificador para determinação da sua performance. O método selecionado para avaliação da capacidade de generalização do modelo, a partir do conjunto de dados, é a validação cruzada 10-fold. Esse método consiste em dividir o conjunto total de dados em k subconjuntos, onde $k = 10$, mutuamente exclusivos de mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste e os $k - 1$ restantes são utilizados para estimação dos parâmetros e calcula-se a performance do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste.

Para garantir a consistência nos resultados alcançados, todos os modelos construídos no projeto experimental usam a mesma amostra de dados. Além disso, os subconjuntos de dados aplicados no processo de validação cruzada também são mantidos estáticos para garantir que os modelos são construídos em condições de igualdade. Essas ações visam reduzir as fontes de variabilidade na condução dos experimentos e substituem o princípio de replicação.

5.2.4 Determinação do Modelo de Planejamento de Experimento

O planejamento de experimentos considerou o tipo de experimento fatorial por ser adequado para estudar, de forma eficiente e econômica, o efeito conjunto de vários fatores (variáveis independentes) sobre uma variável resposta de interesse (variável dependente).

Dentre os tipos de experimentos fatoriais existentes, a pesquisa selecionou o tipo Experimento Fatorial Completo por incluir todas as possíveis combinações entre níveis (valores para as variáveis) dos fatores de experimentos (variáveis independentes). Assim, em um experimento aleatório completo todos os tratamentos possíveis devem ser utilizados no experimento (ver Tabela 5-7), mas a ordem de execução é aleatória.

Fator	**Níveis**
#1 - Tamanho da Janela de Performance	[2,4,8,16] dias
#2 - Disposição da Janela de Performance	[ú$nica, superposta, disjunta$]
#3 - Tipos de Dados	$[RFE, RFM]$
#4 - Técnica de Modelagem	$[NN, DT, LR]$

Tabela 5-7: Projeto experimental e seus respectivos fatores e níveis.

O número total de tratamentos (N) é calculado a partir da análise combinatória dos fatores e seus níveis.

$$N = 4 \times 3 \times 2 \times 3 = 72$$

A aleatoriedade é essencial na condução dos experimentos dada a incerteza sobre todas as variáveis que de fato influenciam no experimento. Desta maneira é possível certificar de que um número maior de variáveis externas ao experimento influencia de maneira igual. A

aleatoriedade não evita complicações dentro do experimento, mas oferece alguma proteção contra o vício do experimento.

A construção do experimento aleatório completo envolve a execução das etapas listadas abaixo:

1) Enumeração de todas as combinações possíveis entre os níveis dos fatores, de 1 a N.
2) Geração de sequência de números aleatórios para a sequência de 1 até N.
3) Condução dos experimentos seguindo a sequência de números obtidos

Os experimentos são realizados de forma aleatória e devem produzir os resultados conforme a Tabela 5-8. Esses resultados são então utilizados na etapa de análise conforme métodos definidos nesse capítulo.

#	**Tamanho da Janela**	**Disposição da Janela**	**Tipos de Dados**	**Técnica de Modelagem**	**Performance**
1	2 dias	Única	RFE	Regressão Logística	0,7643
2	4 dias	Única	RFE	Redes Neurais	0,8113
...	...	...	...	...	...
72	16 dias	Superposta	RFM	Árvore de Decisão	0,5312

Tabela 5-8: Exemplo ilustrativo da tabela com todos os tratamentos e os respectivos resultados.

5.2.5 Comparação dos Resultados

A condução dos experimentos como apresentado até este momento assistirão na análise das escolhas propostas para as decisões-chaves com o propósito de validação desses candidatos a diretrizes no domínio de jogos móveis. Essa avaliação pretende avaliar o impacto das decisões-chaves e suas respectivas escolhas na performance da previsão.

A identificação das diretrizes avaliadas em somente uma base de dados, entretanto, não permite validar se as diretrizes propostas apresentam impacto similar também em outros jogos móveis. Com o objetivo de avaliar esse impacto em diferentes jogos, nós realizamos a aplicação dos experimentos em três bases de dados de jogos móveis reais. Os resultados obtidos para todos os 72 classificadores gerados para cada uma das bases são armazenados para avaliação da relação entre os resultados (Tabela 5-9).

Número do Experimento	Performance para Base de Dados #1	Performance para Base de Dados #2	Performance para Base de Dados #3
1	0,7643	0,7535	0,7441
2	0,8113	0,8011	0,7994
...	...	...	...
72	0,5312	0,5433	0,5126

Tabela 5-9: Exemplo ilustrativo dos resultados dos experimentos para as três bases de dados.

A avaliação da relação entre os resultados foi realizada através de coeficientes de correlação. Esses coeficientes medem o grau pelo qual duas variáveis tendem a mudar juntas. O coeficiente revela o grau e direção da relação. A comparação entre os resultados para os três jogos móveis foi realizada a partir da Correlação de momento de produto de Pearson (r) e da Correlação da ordem de posto de Spearman (ro).

Esses coeficientes são padronizados e apresentam resultados entre -1 e +1 em que a magnitude do coeficiente indica a força da relação. Para o coeficiente de correlação de Pearson ser +1, quando uma variável aumenta, as outras variáveis aumentam por uma quantidade consistente. Se a relação é que uma variável aumenta quando a outra aumenta, mas a quantidade não é consistente, o coeficiente de correlação de Pearson é positivo, mas menor que +1. Quando uma relação é aleatória ou inexistente, os dois coeficientes de correlação se aproximam de zero. As correlações entre as variáveis são consideradas fortes, de acordo com (Cohen 1988), nas situações previstas na Tabela 5-10.

Valor do Coeficiente	Força da Relação
0,1< | r | < 0,3	Correlação baixa / fraca
0,3< | r | < 0,5	Correlação média / moderada
| r | > 0,5	Correlação alta / forte

Tabela 5-10: Interpretação dos resultados da Correlação de Pearson, segundo (Cohen 1988).

A identificação de alta taxa de correlação entre as bases de dados demonstra que os tratamentos propostos para os experimentos apresentam impactos similares, positivos ou negativos, na performance. Além disso, a alta taxa de correlação fornece indícios de que resultados similares devem ser também encontrados em outras bases de jogos.

5.3 Conclusão e Observações

A proposta da abordagem orientada a execução de um projeto experimental permite avaliar os pontos críticos envolvidos na construção de um modelo preditivo para o problema de predição de abandono em jogos. A escolha do tipo de experimentação fatorial permite avaliar a relação entre as variáveis independentemente para extrair informação relevante sobre o impacto de cada um dos níveis propostos para os fatores e a partir dessa informação sistematizar o conhecimento no domínio de aplicação.

6 Execução dos Experimentos e Apresentação dos Resultados

Neste capítulo, nós apresentamos as atividades realizadas na execução do planejamento experimental. A execução inicia-se com a coleta dos dados comportamentais de jogos móveis reais. Esses dados são analisados, transformados e preparados para construção dos modelos resultantes da combinação fatorial entre os níveis dos fatores experimentais. Esses modelos são avaliados e suas performances utilizadas como base para execução dos experimentos.

Em seguida, nós apresentamos os resultados dos experimentos através da Análise de Variância e Análise de Regressão. Essas análises permitem identificar a importância dos fatores, e também da interação entre os fatores, na performance do modelo preditivo. As descobertas realizadas, e os seus respectivos significados, são amplamente discutidas para entendimento do impacto das especificidades de jogos, e seus possíveis tratamentos, na performance da previsão.

6.1 Apresentação das Bases de Dados

O plano experimental com todas as suas etapas é aplicado para a base de dados de 3 jogos móveis reais: 7 Seas, Dino Jump e Armies and Ants. Esses jogos móveis são apresentados nesta seção em conjunto com o processo de extração de suas respectivas bases de dados dos servidores das empresas desenvolvedoras dos jogos.

6.1.1 Jogo Móvel #1: 7 Seas

O 7 Seas (Figura 6-1) é um jogo móvel casual disponível nos sistemas operacionais iOS e Android e lançado sob o modelo de negócios free-to-play em agosto de 2015 pela empresa

BigHut Games. O jogo possui mais de 200 mil downloads e uma base mensal de usuários ativos de aproximadamente 2 mil jogadores.

Figura 6-1: Jogo 7 Seas desenvolvido pela empresa Big Hut Games. Neste jogo casual do gênero quebra-cabeça (Match-3), o usuário é convidado a explorar os sete mares e encontrar os tesouros escondidos em centenas de missões.

As principais ações realizadas pelo usuário dentro do jogo 7 Seas são enviadas para o servidor do jogo através de uma mensagem no formato JSON. Este é um formato leve para intercâmbio de dados computacionais e devido a sua simplicidade tornou-se uma alternativa popular à linguagem XML. As mensagens com as informações sobre as ações realizadas pelo usuário dentro do jogo 7 Seas são enviadas para o servidor segundo o padrão da Tabela 6-1.

```
{
	"userID": "ABCD1-4321a879b185fcb9c6ca27abc5387e914"
	"sessionID": "4879bf37-8566-46ce-9f3b-bd18d6ac614e",
	"installTimestamp": "2015-10-17 15:03:45 ",
	"eventID": "specific event code – eg. gameStarted",
	"eventTimestamp": "yyyy-mm-dd hh:mm:ss.SSS",
	"eventParams": {
		"eventName": "specific event name",
		"eventLevel": "specific event level",
		"actionDescription": "action description ",
		"param1": "stringParam",
		"param2": true,
		"param3": 1234,
		"paramN": ["a", "b", "c"]
	}
}
```

Tabela 6-1: Formato da mensagem JSON enviada do jogo móvel para o servidor para armazenamento das principais ações realizadas pelos jogadores dentro do jogo 7 Seas.

As informações referentes à identificação do evento (*userID*, *sessionID*, *eventID* e *eventTimestamp*) são comuns a todas as mensagens. A caracterização do evento é realizada através dos parâmetros enviados na mensagem. Esses parâmetros especificam informações relevantes sobre o usuário (dados cadastrais), o dispositivo móvel usado pelo jogador, a evolução dentro do jogo (dados situacionais), a situação de término de uma partida, dentre outros aspectos. A lista completa de parâmetros é apresentada na Tabela 6-2.

eventName,eventLevel,gaUserStartDate,gaUserGender,gaUserAgeGroup,gaUserCountry,gaUserAcquisitionChannel,missionID,platform,virtualCurrencyAmount,virtualCurrencyName,hardwareVersion,userLanguage,timezoneOffset,userLocale,clientVersion,isFirstTime,collectInsertedTimestamp,transactionName,UILocation,gameOverBeforeSuccess,giftAccepted,giftName,isTutorial,keyBalance,mainEventID,missionMovesLeft,missionName,transactionType,transactionVector,isPayer,itemName,itemType,parentEventID,pushNotificationToken,operatingSystemVersion,productID,recipientID,senderID,grogBalance,UIAction,missionType,heartBalance,virtualCurrencyType,rewardName,deviceType,isSeeingEndLevel,operatingSystem,terminationReason,UIType,revenueValidated,UIName,productName,productType,productCategory,moreMovesUsed,sdkVersion,deviceName,itemAmount,uIScene,productAmount,missionMovesUsed,missionScore,coinBalance,tutorialChapterID,tutorialStepName,tutorialChapterName,tutorialStepID,acquisitionChannel,afAttrMessage,afAttrStatus,connectedUserID,socialPlatform,socialAlias

Tabela 6-2: Lista com todos os parâmetros relativos às ações realizadas pelo jogador.

O servidor do jogo, ao receber as mensagens, realiza a análise do conteúdo presente no arquivo JSON e armazena as ações do usuário como registros únicos dentro de um banco de dados relacional. Por efeitos práticos, especialmente com relação à velocidade e custo da realização de operações de INSERT na respectiva tabela de ações, a empresa desenvolvedora do jogo optou pela não utilização das regras de normalização e tampouco do conceito de chaves (primária e estrangeira). A tabela com os dados do jogo 7 Seas possui 7 colunas descritas abaixo:

- **event_id** – identificador único do evento.
- **user_id** – identificador único do usuário.
- **session_id** – identificador único da sessão do usuário.
- **event_date** – data de realização do evento (dia, mês e ano).
- **arrival_ts** – marca temporal, ou *timestamp*, da realização do evento (data e hora).
- **event** – descrição do tipo de evento.
- **install_ts** – marca temporal da instalação do aplicativo pelo usuário (data e hora).

Essa tabela com as ações dos usuários foi exportada do servidor e importada em um banco de dados local para nos familiarizarmos com os dados disponíveis através da avaliação

das estatísticas básicas da base. As estatísticas sobre os registros, armazenados no nível de granularidade de ação do usuário, estão descritas abaixo (Tabela 6-3).

Descrição	**Valores**
Período	292 dias (24/07/2015 a 10/05/2016)
Registros	37.543.669
Sessões de Jogo	642.663
Usuários	100.725
Tempo total de Jogo (todos os usuários)	26,661 anos
Tempo médio de Jogo (por usuário)	2,31 horas (139 minutos)

Tabela 6-3: Estatísticas básicas sobre a base de dados do jogo 7 Seas.

6.1.2 Jogo Móvel #2: Dino Jump

O Dino Jump: The Best Adventure (Figura 6-2) é um jogo móvel casual também lançado ao mercado sob o modelo de negócios free-to-play pela empresa BigHut Games. O jogo possui atualmente mais de 2 milhões de downloads e uma base mensal de usuários ativos de aproximadamente 20 mil jogadores. Nesse jogo, o usuário assume o papel de um dinossauro e deve ajudá-lo a escapar de um vulcão em erupção.

Figura 6-2: Imagens ilustrativas do jogo Dino Jump: The Best Adventure.

Os dados básicos de uso do Dino Jump, como dados comportamentais e histórico de compras, são armazenados em um servidor externo. Esses dados foram extraídos do servidor

do jogo, de maneira similar à realizada para o jogo 7 Seas, para realização da análise preliminar dos dados. A tabela com os dados do jogo Dino Jump possui 13 colunas descritas abaixo:

- **user_id** – identificador único do usuário.
- **session_id** – identificador único da sessão do usuário.
- **build** – identificador da versão de software do jogo.
- **event** – descrição detalhada (e textual) do evento.
- **category** – categoria do evento (design ou user).
- **country_code** – código do país de origem do usuário.
- **game_id –** identificador único do jogo.
- **arrival_ts** – marca temporal, ou *timestamp*, da realização do evento (data e hora).
- **install_ts** – marca temporal da instalação do aplicativo pelo usuário (data e hora).
- **platform** – identificador da plataforma (iOS ou Android).
- **device** – identificador do aparelho do usuário (ex: iPhone, iPad, iPod, ...).
- **os** – identificador da versão do sistema operacional do aparelho do usuário.
- **revenue –** informação da receita gerada no evento, no caso de compra de item virtual.

A Tabela 6-4 descreve a situação dos dados logo após a extração. Esses dados foram inseridos em uma base SQL local para realização das etapas de entendimento e pré-processamento dos dados.

Descrição	**Valores**
Período	90 dias (31/12/2014 a 31/03/2015)
Registros	8.520.816
Sessões de Jogo	405.252
Usuários	66.292
Tempo total de Jogo (todos os usuários)	657,37 anos
Tempo médio de Jogo (por usuário)	86,87 horas

Tabela 6-4: Estatísticas básicas sobre a base de dados do jogo Dino Jump.

6.1.3 Jogo Móvel #3: Armies and Ants

O Armies and Ants: Epic War Battle (Figura 6-3) é um jogo móvel *midcore* disponível nos sistemas operacionais iOs e Android e disponibilizado sob o modelo de negócios free-to-play pela empresa Oktagon. O jogo possui atualmente mais de 500 mil downloads e uma base mensal de usuários ativos de aproximadamente 16 mil jogadores. Nesse jogo, o usuário controla uma colônia de formigas com poderes especiais e combate outros jogadores em disputas por território e recursos naturais.

Figura 6-3: Imagem ilustrativa do jogo Armies and Ants.

Os dados básicos de uso do Armies and Ants, como dados comportamentais e histórico de compras, são armazenados em um servidor externo. Esses dados foram extraídos do servidor do jogo para realização da análise preliminar dos dados seguindo um processo similar ao aplicado para o jogo 7 Seas. As duas tabelas com os dados do jogo Armies and Ants possuem as colunas descritas abaixo:

- **Tabela #1 - Events**
  - **user_id** – identificador único do usuário.
  - **session_id** – identificador único da sessão do usuário.
  - **arrival_ts** – marca temporal, ou *timestamp*, da realização do evento (data e hora).
  - **event** – descrição detalhada (e textual) do evento.

- **install_ts** – marca temporal da instalação do aplicativo pelo usuário (data e hora).

- **Tabela #2: Players**
  - **name** – nome do usuário.
  - **first_login -** marca temporal do primeiro login do usuário (data e hora).
  - **first_login -** marca temporal do último login do usuário (data e hora).
  - **playtime** – tempo total de jogo do usuário, em segundos.

A Tabela 6-5 descreve a situação dos dados logo após a extração. Esses dados foram inseridos em uma base SQL local para realização das etapas de entendimento e pré-processamento dos dados.

Descrição	**Valores**
Período	492 dias (08/07/2014 a 12/11/2015)
Registros	7.735.124
Sessões de Jogo	3.867.562
Usuários	289.104
Tempo total de Jogo (todos os usuários)	71,89 anos
Tempo médio de Jogo (por usuário)	21,48 segundos

Tabela 6-5: Estatísticas básicas sobre a base de dados do jogo Armies and Ants.

6.2 Preparação dos Experimentos

A preparação dos experimentos consiste na execução das etapas preliminares do processo CRISP-DM para entendimento de negócios, entendimento de dados e análise e preparação dos dados. Os dados são coletados, analisados e selecionados para em seguida serem devidamente formatados para realização dos experimentos. Por motivos didáticos, a descrição do processo realizado na condução dos experimentos é realizada exclusivamente para o jogo 7 Seas, porém ações similares foram executadas nas bases dos jogos Dino Jump e Armies and Ants. As ações e descobertas diferentes realizadas nas demais bases serão pontuadas quando necessário.

6.2.1 Entendimento de Negócios e Dados

Essa etapa começa com as atividades para se familiarizar com os dados, identificar problemas de qualidade de dados, descobrir informações sobre os dados ou detectar subconjuntos interessantes para formar hipóteses de informações ocultas. Após a coleta dos dados do jogo 7 Seas, nós iniciamos a exploração e descrição dos dados para compreender a organização dos registros e comportamento básico dos usuários dentro do jogo. Primeiramente nós avaliamos a distribuição das sessões de jogo e das instalações sobre o período total de disponibilidade dos dados.

Os gráficos de distribuição demonstram que nos primeiros meses, entre julho e setembro de 2015, os números de sessões de jogo assim como a quantidade de instalações mantiveram-se com valores baixos. Esses meses correspondem ao período de pré-lançamento do jogo em que ações pontuais de marketing são realizadas para atração de usuários com o objetivo de testar e realizar melhorias pontuais no jogo.

O lançamento oficial, com destaque nas lojas de aplicativos, aconteceu de fato no final de outubro. É justamente por esse motivo que esse mês apresentou elevação significativa em ambas as medições. No mês seguinte, novembro de 2015, o jogo ainda está em destaque nas lojas de aplicativos e o número de instalações e sessões de jogo cresce significativamente. Os meses posteriores apresentam performance menor em comparação com novembro, com taxa média de sessões e downloads de 57.631 e 7.381, respectivamente (Figura 6-4).

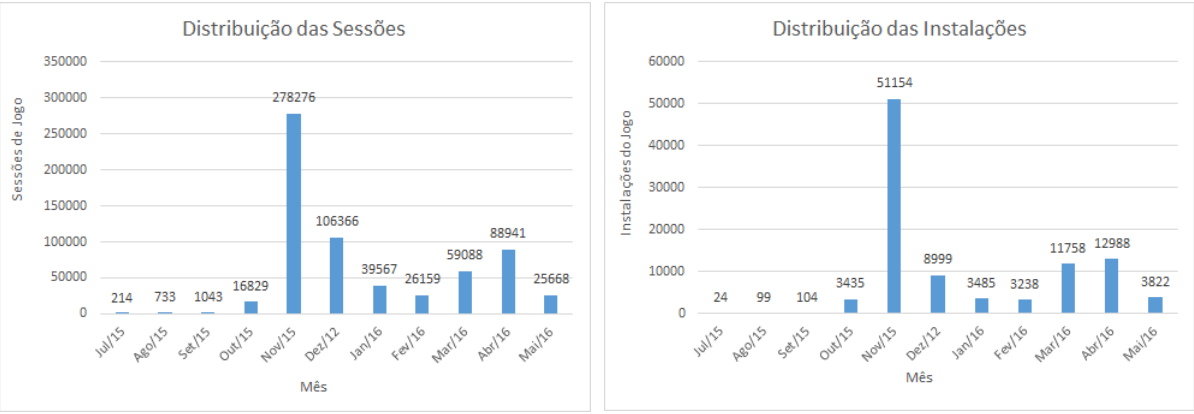

Figura 6-4: Distribuição das sessões e instalações do jogo durante o período de observação (julho de 2015 a maio de 2016). Esquerda) Distribuição das sessões de jogo. Direita) Distribuição das instalações do jogo.

O comportamento médio do usuário, com relação ao tempo de permanência dentro do jogo, também foi avaliado (Figura 6-5 – imagem da esquerda). É possível analisar que a distribuição dos jogadores se comporta como uma função *power-law* como previsto por (Lim

2012). É possível observar também que uma parcela significativa dos usuários abandona o jogo ainda nos primeiros dias. Esse é um comportamento comum na indústria de aplicativos, não somente em jogos, e a taxa de abandono nesses primeiros instantes é conhecida como *bounce rate*.

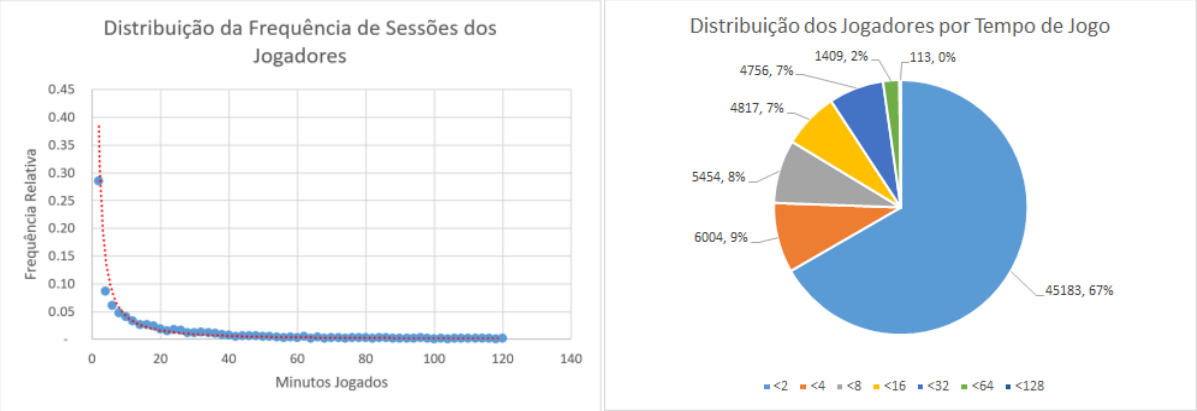

Figura 6-5: Distribuição dos jogadores por tempo e frequência de jogo durante o período de observação (julho de 2015 a maio de 2016). Esquerda) Distribuição dos jogadores pela frequência relativa e tempo total de jogo. Direita) Distribuição dos jogadores por tempo de jogo.

A análise dos dados também demonstrou a existência de *missing data* em uma grande quantidade de campos dos registros das ações. Os principais campos com ausência de dados são apresentados a seguir:

Categoria	**Campos**
Cadastral	`gaUserGender, gaUserAgeGroup, gaUserCountry, gaUserAcquisitionChannel, userLanguage, timezoneOffset, userLocale, connectedUserID, socialPlatform, socialAlias`
Missão	`missionMovesUsed, missionScore, missionMovesLeft, missionName, gameOverBeforeSuccess`
Gifts	`productID,recipientID,senderID, giftAccepted,giftName`
Tutorial	`tutorialChapterID, tutorialStepName, tutorialChapterName, tutorialStepID`
Transação	`transactionType, transactionVector, isPayer, itemName, itemType`

Tabela 6-6: Campos dos registros de ações com maior prevalência de missing data.

A categoria Cadastral consiste nos campos relacionados à demografia do usuário como idade, sexo, país, idioma além de informações sobre a conta de Facebook do usuário. Essas

informações não estão disponíveis na maioria das amostras pois os sistemas operacionais restringem acesso a informações pessoais e a taxa de usuários que realizam login no jogo através do Facebook também é baixa.

As demais categorias informadas (missão, gifts, tutorial e transação) também apresentam ausência de dados devido à natureza do evento enviado. Se o evento relata uma transação realizada na loja do jogo, como a aquisição de um item virtual, os demais campos relativos a tutorial, missão e gifts estarão vazios. E vice-versa.

Por outro lado, os campos de identificação da ação referente ao identificador do usuário, o identificador da sessão, a data de instalação do jogo e, por fim, a data de realização da respectiva ação não apresentam *missing data*.

6.2.2 Preparação dos Dados: Seleção e Limpeza

Essa etapa consiste na realização das atividades de construção e preparação final do conjunto de dados a partir dos dados brutos extraídos da base de dados. As tarefas de preparação de dados incluem a limpeza, seleção e transformação dos dados para uso na etapa seguinte de modelagem do problema.

Nós iniciamos as ações de preparação de dados através da realização das ações relativas à limpeza de dados para reduzir a quantidade de informações a serem manipuladas na base de dados e consequentemente abreviar o tempo de realização das ações de transformação dos dados. Após a realização da limpeza dos dados, a próxima etapa consiste na transformação do grão para permitir a construção, integração e formação dos dados. As ações são descritas em maiores detalhes abaixo.

6.2.2.1 Exclusão de registros desnecessários

A base de dados contém pouco mais de 9 meses de registros com relação ao comportamento dos usuários dentro do jogo 7 Seas. Esse período é maior do que o necessário para a condução dos experimentos com janelas de observação e resultado com tamanhos de 16 dias (máximo) e 30 dias, respectivamente. A seleção dos primeiros 46 dias, entre julho e agosto de 2016, implicaria no desperdício de uma grande quantidade de informações dada a distribuição de instalações e sessões demonstrada na Figura 6-4. Por esse motivo, nós optamos pela seleção do dia 01/12/015 como o ponto de observação e assim incluir todas as informações do mês de novembro na janela de observação. As informações posteriores à janela resultados foram excluídos.

```
DELETE FROM test.7seas
WHERE date > '2015-12-30';
```

6.2.2.2 Exclusão de campos desnecessários

Os registros das ações do usuário incluem uma série de informações inerentes ao jogo como pontuação, vidas, missões e presentes virtuais. E conforme previsto no plano experimental, nós decidimos pela remoção dos dados específicos do jogo para buscar soluções abrangentes e independentes das características e peculiaridades de cada um dos jogos. Os dados relacionados à sessão de jogo e compras foram mantidos e os demais dados removidos.

```
id, user_id, session_id, arrival_ts, install_ts, event, action
```

6.2.2.3 Transformação do Grão

Essa atividade consiste na definição do nível de granularidade desejado (usuário) para início das ações de pré-processamento para construção, integração e formatação dos dados. Os dados inicialmente obtidos na granularidade mensagem, ou ação do usuário, foram transformados para a granularidade usuário. A transformação resulta em uma nova tabela contendo somente as informações abaixo com o campo `user_id` como chave primária.

```
user_id, install_ts
```

Essa nova tabela possui 51.713 registros referentes ao número de usuários presentes no período considerado para análise diferentemente dos 25.939.325 registros existentes na tabela original referentes à quantidade de eventos.

6.2.3 Preparação dos Dados: Construção

Essa etapa é responsável pela transformação dos dados para construção dos atributos necessários para a criação de todos os classificadores com base em cada um dos tratamentos previstos no projeto experimental. A construção dos atributos (variáveis independentes) consiste na agregação dos dados do usuário pertencentes à janela de performance e depende da configuração da janela. A janela, por sua vez, varia conforme a combinação entre os fatores destacados na Tabela 6-7.

Fator	Níveis
#1 - Tamanho da Janela de Performance	[2,4,8,16] dias
#2 - Disposição da Janela de Performance	[única, $superposa, disjunta$]
#3 - Tipos de Dados	[RFE, RFM]

Tabela 6-7: Fatores com impacto na construção dos dados.

A abordagem utilizada na construção dos dados considerando as possíveis combinações entre os fatores é descrita em detalhes a seguir.

6.2.3.1 Configuração das Janelas de Performance

As combinações dos níveis de fatores 1 e 2, conforme apresentado na Tabela 6-7, impactam diretamente no espectro de dados a serem coletados de cada um dos usuários para construção dos dados. As combinações dos níveis de fatores (N) referentes à janela de performance estão descritas a seguir.

$$N = Fator\ \#1\ (4\ níveis)\ \times Fator\ \#2\ (3\ níveis) = 12$$

6.2.3.2 Tipos de Dados

Os níveis propostos para o fator tipo de dados estão relacionados à análise RFE e RFM. A RFE contempla a produção de 7 diferentes tipos de atributos derivados da sessão do usuário enquanto a RFM contempla a produção de 3 atributos adicionais relativos ao comportamento de compra. A descrição dos atributos foi realizada no capítulo anterior e a nomenclatura proposta é apresentada a seguir.

$$ow\_[supp|disj|unica]\_[rfe|rfm]\_[nome\ do\ atributo]\_[2|4|8|16]$$

Onde,

- ***ow*** — indica atributo construído a partir da janela de performance, ou janela de observação.
- ***[supp|disj|unica]*** — níveis do fator Disposição da Janela de Performance (*stat* = Estática e *stag* = Escalonada).

- ***[rfm|rfe]*** — níveis do fator tipos de dados (*rfe* = Análise RFE e *rfm* = Análise RFM).
- ***[nome do atributo]*** — nome do atributo associado ao nível indicado para tipos de dados. Os atributos podem assumir os valores previstos para níveis Análise RFE e Análise RFM apresentados em detalhes no capítulo 4.2.
- ***[2|4|8|16]*** — níveis do fator tamanho da janela de performance.

Para ilustrar a aplicação da nomenclatura, nós apresentamos abaixo exemplos de atributos construídos, sem a intenção de esgotar todas as possibilidades.

- *ow_supp_rfm_session_total_4* — Esse atributo foi construído tendo como base a janela Superposta de 4 dias de duração e considerando o número total de sessões executadas pelo usuário nesse período. Esse atributo será usado na construção dos modelos preditivos para todos os tratamentos com Tamanho de Janela = 4 dias, Tipo de Dados = Análise RFM e Disposição da Janela = Superposta.
- *ow_unica_rfe_purchase_frequency_16* — Esse atributo foi construído tendo como base a janela única de 16 dias de duração e considerando o atributo do tipo frequência de compras referente à última fatia de tempo das 8 partições consideradas. Esse atributo será usado na construção dos modelos preditivos para todos os tratamentos com Tamanho de Janela = 16 dias, Tipo de Dados = Análise RFE e Disposição da Janela = Única.

A nomenclatura proposta é responsável por dar nome a 132 atributos como calculado abaixo:

Com relação à **Tipos de Dados**: **11**
Com relação ao **Tamanho da Janela** (2, 4, 8, 16)**: 4**
Com relação à **Disposição da Janela** (única, superposta e disjunta): **3**
Total: $11 \times 4 \times 3 = 132$

6.2.4 Preparação dos Dados: Definição do Rótulo

A definição do rótulo é realizada com base na janela de resultado com prazo de 30 dias e a ausência de ações nesse período implica no rótulo *churner*. A regra descrita abaixo

exemplifica a definição da situação se o usuário abandonou o jogo (*churner*) ou não (*non-churner*).

```
IF quantidade de sessões entre 01/12/2015 e 30/12/2015 = 0
THEN usuário = churner.
ELSE usuário ≠ churner.
```

A definição do rótulo é responsável pela criação de mais um atributo a ser utilizado na aprendizagem supervisionada dos classificadores. Dessa maneira o total de atributos gerados é de 132 + 1 = **133.**

6.2.5 Preparação dos Dados: Formatação

Os atributos precisam ser devidamente preparados para uso no processo de construção do modelo preditivo. O principal passo executado aqui é apresentado a seguir.

- **Normalização**: O propósito da normalização é minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis. Nesse caso, os atributos tiveram seus valores normalizados para o mesmo intervalo. O método de normalização por desvio padrão foi utilizado com o objetivo de converter os dados em uma distribuição normal com $média = 0$ e $variância = 1$. A fórmula utilizada consiste em:

$$f(X) = (X - \mu)/\sigma$$

Onde X representa o vetor de atributos, μ representa a média e σ representa o desvio padrão. O intervalo de valores de uma distribuição normal não está dentro de [0,1], mas dentro da faixa $[-3, +3]$. Apesar da distribuição permitir valores dentro de uma faixa maior, ao considerarmos $[-3, +3]$ já capturamos 99,9% dos dados.

Após a execução dos passos para formatação, a base de dados está pronta para uso na etapa de modelagem (Tabela 6-8).

Antes	Depois
Registros: **37.543.669**	Registros: **51.713**
Usuários: **100.725**	Usuários: **51.713**
Atributos: **7**	Atributos: **133**

Tabela 6-8: Fatores com impacto na construção dos dados.

A base de dados é exportada para o formato CSV, acrônimo de *Comma-Separated Values*, para utilização na etapa de modelagem apresentada na próxima seção.

6.3 Condução dos Experimentos

A condução dos experimentos consiste na construção dos modelos preditivos para todas as combinações possíveis entre os fatores para avaliação da respectiva performance em cada um dos tratamentos existentes. O processo realizado na execução dos experimentos é ilustrado na Figura 6-6.

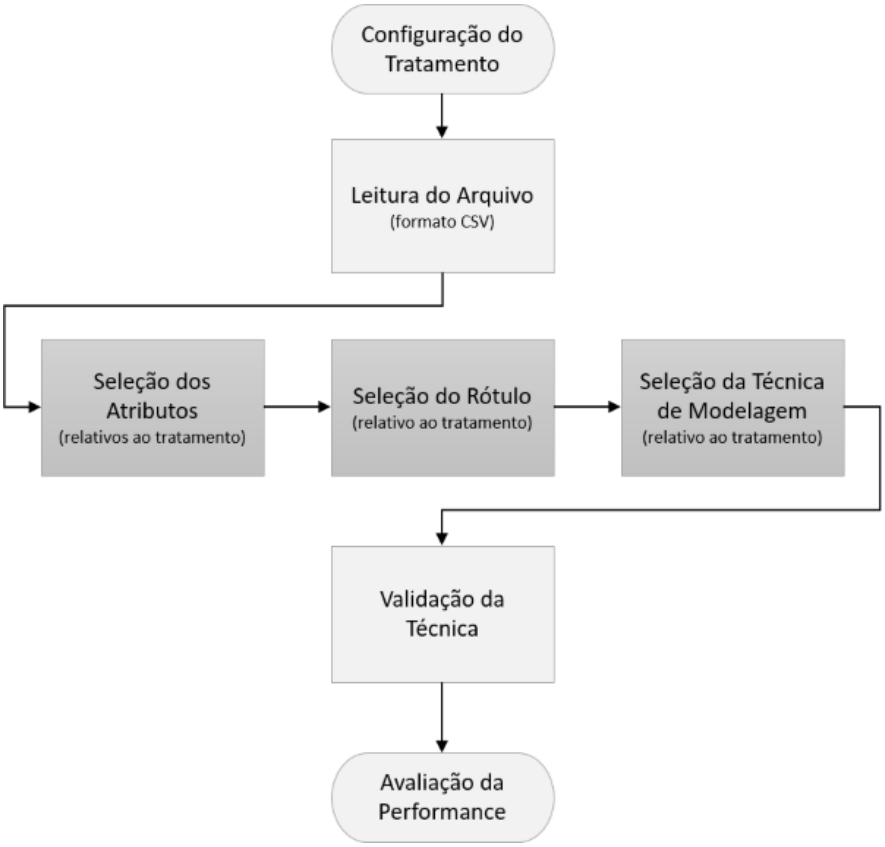

Figura 6-6: Representação gráfica do processo de condução de experimento relativo ao tratamento.

As etapas, e as respectivas ações realizadas, são descritas abaixo.

6.3.1 Configuração do Tratamento

A entrada do processo para condução de um experimento consiste na especificação do tratamento com a informação dos níveis de cada um dos fatores. A condução de todos os experimentos ($N = 72$) implica na realização desse processo para todos os tratamentos. Para efeito de ilustração consideremos o tratamento abaixo Tabela 6-9.

Tamanho da Janela	**Disposição da Janela**	**Tipos de Dados**	**Técnica de Modelagem**
8 dias	Única	Análise RFM	Regressão Logística

Tabela 6-9: Exemplo de tratamento.

6.3.2 Leitura do Arquivo CSV

As informações relativas ao entendimento e preparação dos dados foram armazenadas em um arquivo CSV. A leitura do arquivo proporciona acesso a todos os atributos ($N = 121$) armazenados no grão usuário.

6.3.3 Seleção dos Atributos

De acordo com a configuração do tratamento, os atributos relacionados são selecionados para modelagem da técnica e os demais são desconsiderados na construção do classificador. No exemplo da Tabela 6.7, somente os atributos referentes aos níveis de fatores considerados são mantidos para etapa de modelagem.

$$\#atributos = 1\ (tamanho) \times 1\ (disposição) \times 10\ (tipos\ de\ dados)$$
$$\#atributos = 10$$

Em outras palavras, do total de 121 atributos gerados são selecionados 10 deles relativos à janela de performance para permanência para as próximas etapas.

6.3.4 Seleção da Técnica de Modelagem

Após a seleção de todos os atributos a serem utilizados, a técnica de modelagem é também escolhida e configurada conforme explicitado no plano experimental. Para o exemplo

citado, a base de dados conta com 10 atributos formados e 1 rótulo. Esses dados são modelados através da técnica de Regressão Logística, conforme a configuração do tratamento.

6.3.5 Validação do Modelo

A validação consiste na aplicação de técnica para avaliar a capacidade de generalização de um modelo de classificação, a partir de um conjunto de dados. A técnica adotada é a validação cruzada. O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, utiliza-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo.

O método de particionamento dos dados escolhido é o k-*fold*, com $k = 10$. Esse método consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros e calcula-se a precisão do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. Ao final das k iterações calcula-se a precisão sobre os erros encontrados obtendo assim uma medida mais confiável sobre a capacidade do modelo de representar os dados.

6.3.6 Avaliação da Performance

Essa é a etapa final da modelagem de um tratamento. A avaliação da performance dos modelos construídos para cada um dos tratamentos é realizada através da Área sob a Curva ROC. Ao final da construção dos modelos para todos os tratamentos possíveis, a performance do modelo é armazenada para cada um dos tratamentos.

Após a condução dos experimentos os resultados são armazenados no formato abaixo com todas as combinações de tratamentos possíveis associadas às suas respectivas performances.

#	Tamanho da Janela	Disposição da Janela	Tipos de Dados	Técnica de Modelagem	Performance
1	2 dias	Superposta	RFE	Regressão Logística	0.7535
2	2 dias	Superposta	RFE	Redes Neurais	0.8163
...	...	...	...	...	...
72	16 dias	Escalonada	RFM	Árvore de Decisão	0.5312

Tabela 6-10: Exemplo ilustrativo da tabela com todos os tratamentos e os respectivos resultados dos modelos preditivos.

Esses dados são utilizados para avaliação dos experimentos e identificação do efeito de cada fator, assim como das interações entre os fatores. A descrição dos resultados é realizada na próxima seção.

6.3.7 Resumo da Execução

A execução foi realizada para as demais bases de dados dos jogos Dino Jump e Armies and Ants de maneira similar à realizada para o jogo 7 Seas explicitado nesta seção. Uma das particularidades identificadas consistiu basicamente na construção dos atributos relativos à análise RFM. Esses dois jogos móveis, apesar de terem sido lançados sob o modelo Freemium e realizarem a venda de itens virtuais, as ações relativas à compra desses itens não são armazenadas na base de dados.

Dessa forma, a preparação dos dados para esses jogos não incluiu a construção dos atributos monetários. O total de atributos construídos nessas bases de dados é de 97 diferentemente do 7 Seas que conta com 133 atributos.

Antes	**Depois**
Registros: 8.520.816	Registros: 54.237
Usuários: 66.292	Usuários: 54.237
Atributos: 7	Atributos: 97

Tabela 6-11: Resultado da transformação dos dados do jogo Dino Jump.

Antes	Depois
Registros: 7.735.124	Registros: 95.196
Usuários: 289.104	Usuários: 95.196
Atributos: 6	Atributos: 133

Tabela 6-12: Resultado da transformação dos dados do jogo Armies and Ants.

Além da mudança na quantidade de atributos gerados, a condução dos experimentos para essas duas bases resultou na construção de 36 classificadores, e não 72 como apresentado para a base do jogo 7 Seas. Essa alteração se deve basicamente à remoção do fator Tipos de Dados do plano experimental com a utilização exclusiva dos dados da análise RFE. O número total de tratamentos (N) é calculado a partir da análise combinatória dos fatores e seus níveis.

$$N = 4 \times 3 \times 1 \times 3 = 36$$

Outra particularidade identificada a partir da análise dos dados dos jogos foi a presença de ruídos nos dados extraídos do jogo Dino Jump. No Dino Jump a data das ações realizadas pelos usuários é registrada no momento em que a mensagem é recebida no servidor, porém nós identificamos que a mensagem enviada do jogo para o servidor pode ser armazenada para o envio posterior. Esse envio pode acontecer horas depois como pode acontecer semanas depois. Nos demais jogos avaliados (7 Seas e Armies and Ants), a data das ações é registrada no próprio cliente, mesmo que a ação seja armazenada no servidor posteriormente. Isso significa que a data real de realização da ação não é perdida. Esse ruído nos dados do Dino Jump impacta diretamente na construção dos atributos relativos à quantidade de sessões, frequência de sessões, duração das sessões e tempo de ausência.

6.4 Apresentação dos Resultados do Jogo 7 Seas

A avaliação dos experimentos é executada através das técnicas propostas de Análise de Variância (ANOVA) e Análise de Regressão. Os resultados são apresentados a seguir.

6.4.1 Análise de Variância

A ANOVA é aplicada para avaliação dos resultados do projeto experimental para avaliar se as diferenças amostrais observadas são reais (causadas por diferenças significativas nas populações observadas) ou casuais (decorrentes da mera variabilidade amostral). Essa análise parte do pressuposto, portanto, que o acaso só produz pequenos desvios sendo as grandes diferenças geradas por causas reais. Os resultados da ANOVA são apresentados na Tabela 6-13: Resultado da Análise de Variância.Tabela 6-13.

Análise de Variância

Fonte	GL	SQ Seq	Contrib.	QM (Aj.)	Valor F	Valor-P
Modelo	36	0.325082	99.97%	0.009030	3809.14	0.000
Linear	8	0.296566	91.20%	0.037071	15637.51	0.000
disposicao	2	0.013530	4.16%	0.006765	2853.66	0.000
dados	1	0.000015	0.00%	0.000015	6.23	0.017
tecnica	2	0.266904	82.08%	0.133452	56293.94	0.000
tamanho	3	0.016117	4.96%	0.005372	2266.21	0.000
Interações de 2 fatores	16	0.025399	7.81%	0.001587	669.64	0.000
disposicao*tecnica	4	0.018771	5.77%	0.004693	1979.50	0.000
disposicao*tamanho	6	0.003572	1.10%	0.000595	251.12	0.000
tecnica*tamanho	6	0.003057	0.94%	0.000509	214.91	0.000
Interações de 3 fatores	12	0.003117	0.96%	0.000260	109.57	0.000
disposicao*tecnica*tamanho	12	0.003117	0.96%	0.000260	109.57	0.000
Erro	35	0.000083	0.03%	0.000002		
Total	71	0.325165	100.00%			

Tabela 6-13: Resultado da Análise de Variância.

Na Tabela 6-13 foram mantidos somente os resultados estatisticamente significativos (Valor-P < 0,05) provocando a exclusão de interações não significativas entre 3 fatores. A ANOVA revela que a soma da contribuição individual dos fatores contribui com 91,20% da variação da performance. As interações entre 2 fatores, ou de segunda ordem, contribuem com 7,81% da variação da performance. As demais interações respondem por 0,96% da variação enquanto o erro calculado contribui com 0,03% da variação total com relação à performance do classificador.

A ANOVA também produz o gráfico de efeitos principais o qual representa a média de resposta de cada nível de fator conectado por uma linha (ver Figura 6-7). Esse gráfico permite examinar as diferenças entre as médias dos níveis de fatores.

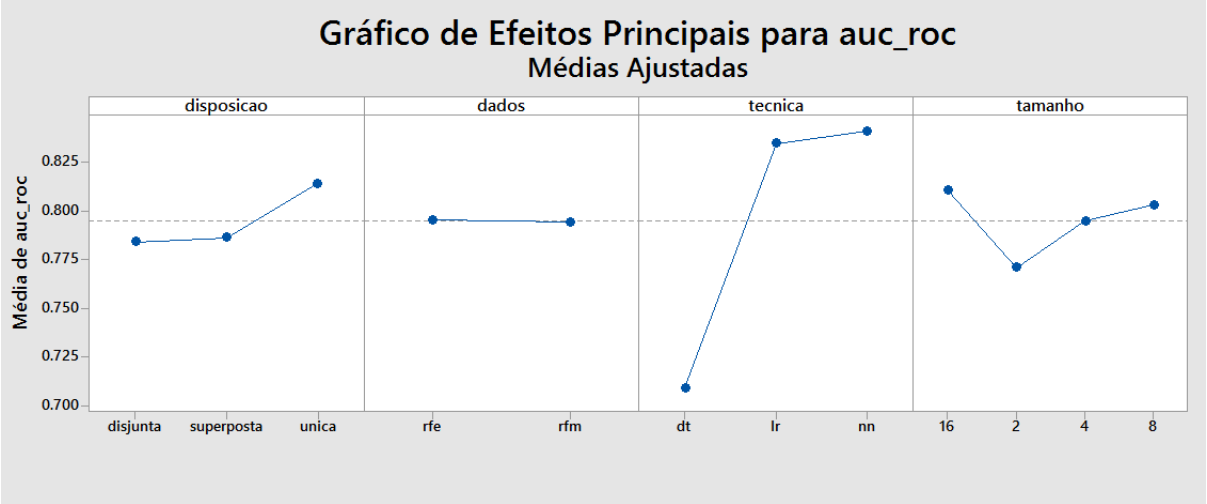

Figura 6-7: Gráfico de efeitos principais.

Os efeitos principais acontecem quando diferentes níveis de um fator afetam a resposta diferentemente. Os principais padrões responsáveis por indicar efeitos principais são:

- Quando a linha é horizontal (paralela ao eixo x), então não há efeito principal. Cada nível do fator afeta a resposta da mesma maneira e a média de resposta é a mesma em todos os níveis de fator.
- Quando a linha não é horizontal, então há um efeito principal. Níveis diferentes do fator afetam a resposta de forma diferente. Quanto mais íngreme a inclinação da linha, maior a magnitude do efeito principal.

O gráfico de efeitos principais demonstra, por exemplo, que a performance média para as técnicas de modelagem Regressão Logística (0,834) e Redes Neurais (0,841) apresentam a maior performance do que a Árvore de Decisão (0,708). Em termos de Tamanho de Janela, a performance média cresce em paralelo com o aumento do tamanho da janela, com a maior performance média para o nível 16 dias (0,810). A interação é apresentada pelo gráfico de interação apresentado na Figura 6-8 usado para visualizar as interações possíveis entre os fatores. Os padrões a serem procurados nesse gráfico são:

- Linhas paralelas em um gráfico de interação indicam que não há interação.
- Quanto maior a diferença de inclinação entre as linhas, maior o grau de interação.

É importante atentar que o gráfico de interação somente apresenta a interação se a mesma é estatisticamente significante.

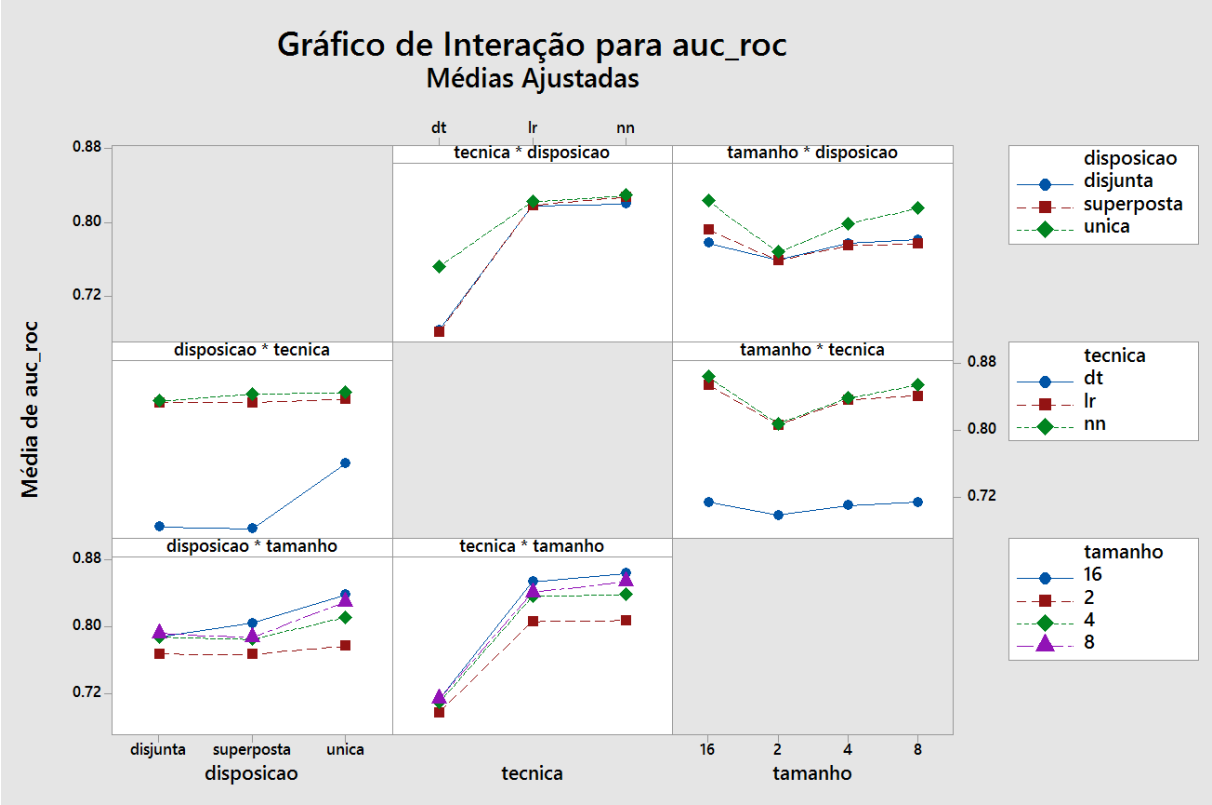

Figura 6-8: Gráfico de interação entre níveis de fatores.

O gráfico de interação demonstra, por exemplo, a interação mais expressiva entre os fatores Disposição da Janela e Técnica de Modelagem com contribuição de 5,77% na alocação de variação da média de performance. A análise gráfica da Figura 6-8, considerando a contribuição das interações segundo a ANOVA, permite identificar os pontos com padrões de interação. Esses pontos estão destacados na Figura 6-9.

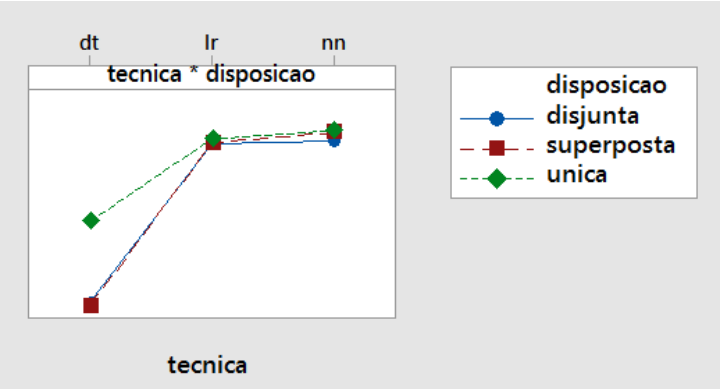

Figura 6-9: Padrões de interação entre os fatores com maior contribuição na alocação da variação.

6.4.2 Análise de Regressão

A análise de Regressão, como descrito na seção 5.1.3.2, gera uma equação para descrever a relação estatística entre um ou mais preditores (fatores e seus níveis) e a variável de resposta (performance). O resultado da Análise de Regressão é apresentado na Tabela 6-14.

Sumário do Modelo

S	R2	R2(aj)	PRESQ	R2(pred)
0.0015397	99.97%	99.95%	0.0003511	99.89%

Coeficientes

Termo	Coef	IC de 95%		Valor-T	Valor-P	VIF
Constante	0.794952	(0.794583,	0.795320)	4381.02	0.000	
disposicao						
disjunta	-0.010733	(-0.011254,	-0.010213)	-41.83	0.000	1.33
superposta	-0.008614	(-0.009135,	-0.008093)	-33.57	0.000	1.33
unica	0.019348	(0.018827,	0.019869)	75.40	0.000	*
dados						
rfe	0.000453	(0.000085,	0.000821)	2.50	0.017	1.00
rfm	-0.000453	(-0.000821,	-0.000085)	-2.50	0.017	*
tecnica						
dt	-0.086023	(-0.086544,	-0.085502)	-335.22	0.000	1.33
lr	0.039762	(0.039241,	0.040283)	154.95	0.000	1.33
nn	0.046261	(0.045740,	0.046782)	180.27	0.000	*
tamanho						
16	0.015779	(0.015141,	0.016417)	50.20	0.000	1.50
2	-0.024049	(-0.024687,	-0.023411)	-76.52	0.000	1.50
4	0.000021	(-0.000617,	0.000659)	0.07	0.947	1.50
8	0.008249	(0.007611,	0.008887)	26.25	0.000	*
disposicao*tecnica						
disjunta dt	-0.013733	(-0.014470,	-0.012996)	-37.84	0.000	1.78
disjunta lr	0.009249	(0.008512,	0.009986)	25.49	0.000	1.78
disjunta nn	0.004484	(0.003748,	0.005221)	12.36	0.000	*
superposta dt	-0.018323	(-0.019059,	-0.017586)	-50.49	0.000	1.78
superposta lr	0.007363	(0.006626,	0.008100)	20.29	0.000	1.78
superposta nn	0.010959	(0.010223,	0.011696)	30.20	0.000	*
unica dt	0.032056	(0.031319,	0.032792)	88.33	0.000	*
unica lr	-0.016612	(-0.017349,	-0.015875)	-45.77	0.000	*
unica nn	-0.015444	(-0.016180,	-0.014707)	-42.56	0.000	*
disposicao*tamanho						
disjunta 16	-0.011457	(-0.012359,	-0.010554)	-25.78	0.000	2.00
disjunta 2	0.007591	(0.006688,	0.008493)	17.08	0.000	2.00
disjunta 4	0.003889	(0.002987,	0.004792)	8.75	0.000	2.00
disjunta 8	-0.000024	(-0.000926,	0.000879)	-0.05	0.958	*
superposta 16	0.002802	(0.001900,	0.003705)	6.31	0.000	2.00
superposta 2	0.005402	(0.004500,	0.006305)	12.15	0.000	2.00
superposta 4	-0.001091	(-0.001993,	-0.000189)	-2.45	0.019	2.00
superposta 8	-0.007114	(-0.008016,	-0.006211)	-16.01	0.000	*
unica 16	0.008654	(0.007752,	0.009556)	19.47	0.000	*
unica 2	-0.012993	(-0.013895,	-0.012091)	-29.23	0.000	*
unica 4	-0.002798	(-0.003701,	-0.001896)	-6.30	0.000	*
unica 8	0.007137	(0.006235,	0.008040)	16.06	0.000	*
tecnica*tamanho						
dt 16	-0.010802	(-0.011704,	-0.009900)	-24.30	0.000	2.00
dt 2	0.013396	(0.012493,	0.014298)	30.14	0.000	2.00
dt 4	0.000677	(-0.000225,	0.001579)	1.52	0.137	2.00
dt 8	-0.003271	(-0.004173,	-0.002368)	-7.36	0.000	*

lr 16	0.003858	(0.002956, 0.004760)	8.68	0.000	2.00
lr 2	-0.004122	(-0.005025, -0.003220)	-9.27	0.000	2.00
lr 4	0.001859	(0.000956, 0.002761)	4.18	0.000	2.00
lr 8	-0.001594	(-0.002497, -0.000692)	-3.59	0.001	*
nn 16	0.006944	(0.006042, 0.007846)	15.62	0.000	*
nn 2	-0.009273	(-0.010176, -0.008371)	-20.86	0.000	*
nn 4	-0.002536	(-0.003438, -0.001633)	-5.70	0.000	*
nn 8	0.004865	(0.003963, 0.005767)	10.95	0.000	*

Tabela 6-14: Sumário do modelo de regressão e seus coeficientes.

Os resultados demonstram que a maioria dos preditores são significativos devido a seus Valores-P baixos com a principal exceção sendo o tamanho de janela 4 dias. O coeficiente de determinação, também chamado de *R squared* (R^2), estabelece que os preditores explicam 99,97% da variância da performance dos modelos de previsão de abandono.

A análise de regressão confirma, por exemplo, que as técnicas de Regressão Logística e Redes Neurais apresentam relação positiva com a variável resposta dado os coeficientes positivos de +0,0397 e +0,0462, respectivamente. Por outro lado, a técnica de Árvore de Decisão apresenta relação negativa dado sinal de seus coeficientes (-0,0860).

6.4.3 Validação do Modelo

A aplicação da ANOVA baseia-se em pressupostos básicos relativos aos efeitos e erros. Esses pressupostos precisam ser checados para validação do modelo. Os componentes básicos de um modelo válido, baseado em regressão, são:

$$Resposta = (Constante + Preditores) + Erro$$

Outra forma de representar esses componentes é:

$$Resposta = (Determinístico) + Estocástico$$

A parte determinística consiste na porção da resposta explicada pelas variáveis explicativas do modelo. O valor esperado da resposta é uma função de um conjunto de variáveis preditoras. Toda a informação explicativa / preditiva do modelo deve estar nessa parte. A parte estocástica representa os fatores aleatórios e imprevisíveis. Erro é a diferença entre o valor esperado e o valor observado. Juntando isso, as diferenças entre os valores esperado e observado

devem ser imprevisíveis. Em outras palavras, nenhuma das informações explicativas / preditivas deve estar no erro.

A ideia é que a parte determinística de modelo é tão boa em explicar (ou prever) a resposta que apenas a aleatoriedade inerente de qualquer fenômeno do mundo real permanece na porção de erro. Em outras palavras, se nós observarmos poder explicativo ou preditivo no erro isso significa que os preditores estão perdendo algumas das informações preditivas. Os gráficos de resíduos (Figura 6-10) nos permite avaliar esses pressupostos. Os resíduos representam os desvios das observações da média da amostra (= observado - previsto). No total foram gerados quatro gráficos:

- **Papel de Probabilidade:** Esse gráfico é usado para verificar o pressuposto de que os resíduos estão normalmente distribuídos. Nesse gráfico os resíduos devem ser aproximar o máximo possível da linha reta como visto na Figura 6-10 (a). A análise gráfica permite identificar que a distribuição normal parece ajustar-se bem aos dados da amostra.
- **Histograma dos resíduos:** Esse gráfico permite avaliar se a variância é normalmente distribuída. O gráfico também informa se os dados estão enviesados ou se existem *outliers* nos dados. A análise permite identificar que o gráfico segue uma curva em forma de sino indicando sua distribuição normal.
- **Resíduos Versus Valores Ajustados:** Esse gráfico é utilizado para verificar a suposição de que os resíduos têm uma variância constante. Portanto, os resíduos devem seguir um padrão simétrico e ter uma propagação constante em toda a faixa. Se algum comportamento sistemático for observado no gráfico, temos indícios de que alguma variável "extra" influenciou nos resultados do experimento, fato que viola uma das premissas básicas da ANOVA e compromete nossas conclusões.
- **Resíduos versus a ordem de coleta dos dados:** Este gráfico permite verificar a suposição de que os resíduos não estão correlacionados entre si. Da mesma forma que o gráfico anterior, nós devemos avaliar a presença de comportamento sistemáticos. A avaliação gráfica permite identificar a independência entre os resíduos.

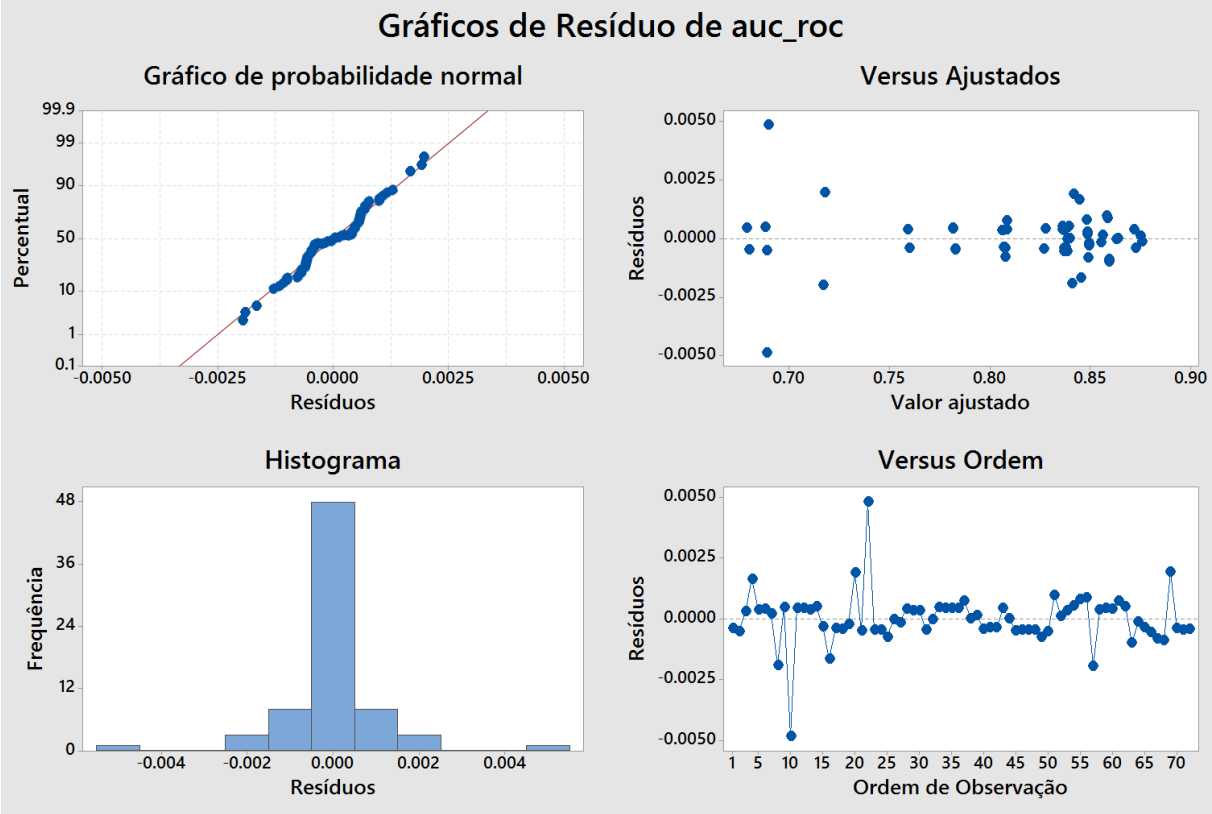

Figura 6-10: Gráficos de resíduos. A) no canto superior-esquerdo, o gráfico de Probabilidade Normal. B) no superior-direito, o gráfico de Resíduos Versus Valores Ajustados. C) no inferior-esquerdo, o Histograma de Resíduos. D) no inferior-direito, o gráfico de Resíduos Versus a Ordem de Coleta dos Dados.

A avaliação dos gráficos de resíduo determina que as pressuposições necessárias estão satisfeitas e, portanto, confirma a validade do modelo. Outro aspecto importante a ser avaliado é a presença de *outliers* nos dados. Um *outlier* é uma observação atipicamente grande ou pequena. *Outliers* podem ter um efeito desproporcional sobre os resultados estatísticos como a média causar interpretações errôneas.

Os gráficos de resíduos também permitem avaliar a presença, ou ausência, de outliers nos dados a partir da identificação de pontos distantes dos demais, especialmente em gráficos de dispersão. Em alguns casos, entretanto, é importante examinar mais de um tipo de gráfico porque outliers que aparecem em um gráfico podem não ser óbvios em outro gráfico. O gráfico *boxplots* é utilizado para identificação de *outliers*. Esse gráfico usa um asterisco (*) para identificar *outliers*. Esses *outliers* são observações pelo menos 1,5 vezes o intervalo entre quartis (Q3 - Q1) da borda da caixa.

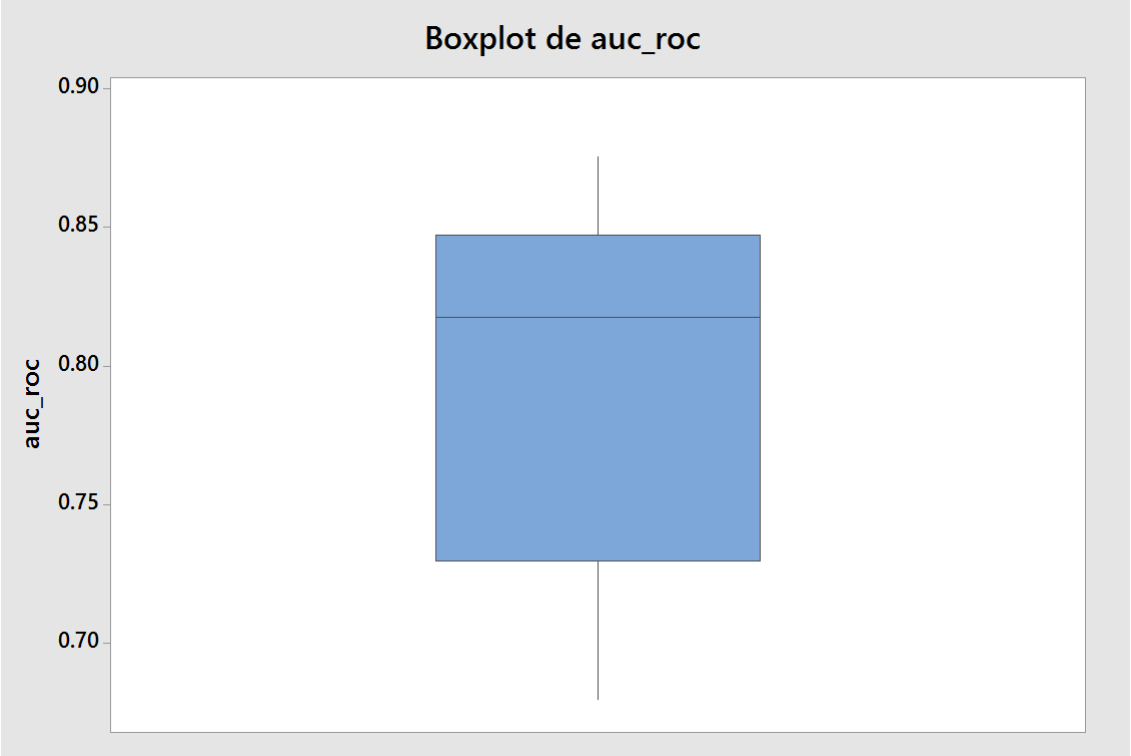

Figura 6-11: Gráfico de boxplot da distribuição para identificação de outliers.

A avaliação dos gráficos de resíduo e também do gráfico de *boxplots* permite afirmar que não há *outliers* nos dados que possam provocar efeitos desproporcionais e indesejados nos resultados estatísticos.

6.5 Apresentação dos Resultados dos Outros Jogos

Essa avaliação é realizada para identificação da relação entre os resultados encontrados para as três diferentes bases. A expectativa é identificar a relação positiva entre os resultados, ou seja, a descoberta de que as performances dos classificadores para previsão de abandono dos três jogos avaliados tendem a mudar juntos e na mesma direção. Antes de aplicar os coeficientes de correlação, nós construímos um gráfico com todas as performances alcançadas para permitir a avaliação gráfica dos resultados.

É importante notar que dada a restrição de dados monetários nas bases de dados dos jogos Dino Jump e Armies and Ants, a avaliação da correlação entre os dados foi realizada com base não em 72 experimentos, mas em 36 experimentos. O gráfico traça a performance dos classificadores (eixo Y) para cada um dos 36 experimentos (eixo X) como é possível ver na Figura 6-12. A identificação dos experimentos é realizada conforme a combinação dos níveis de fatores segundo a regra a seguir e na ordem apresentada:

$$experimento_x = (disjunta, superposta, unica) \times (rfe) \times (nn, lr, dt) \times (2,4,8,16)$$

Essa regra leva à produção dos experimentos de maneira ordenada. Para ilustrar os resultados nós apresentamos a seguir alguns dos tratamentos utilizados para produção dos experimentos. A numeração x é utilizada para identificar os tratamentos no gráfico de comparação da performance dos experimentos (Figura 6-12).

$$experimento_1 = \{disjunta, rfe, nn, 2\}$$
$$experimento_2 = \{disjunta, rfe, nn, 4\}$$
$$experimento_3 = \{disjunta, rfe, nn, 8\}$$
$$experimento_{10} = \{disjunta, rfe, dt, 4\}$$
$$experimento_{16} = \{superposta, rfe, nn, 16\}$$
$$experimento_{29} = \{unica, rfe, l2,2\}$$

A simples avaliação gráfica permite observar que há relação entre as séries de resultados dado que os valores de performance tendem a mudar de maneira proporcional. A avaliação gráfica também demonstra que a performance média para a base de dados do jogo Dino Jump é inferior às demais bases de dados. A provável explicação consiste no ruído identificado nos dados com relação ao aspecto temporal das informações.

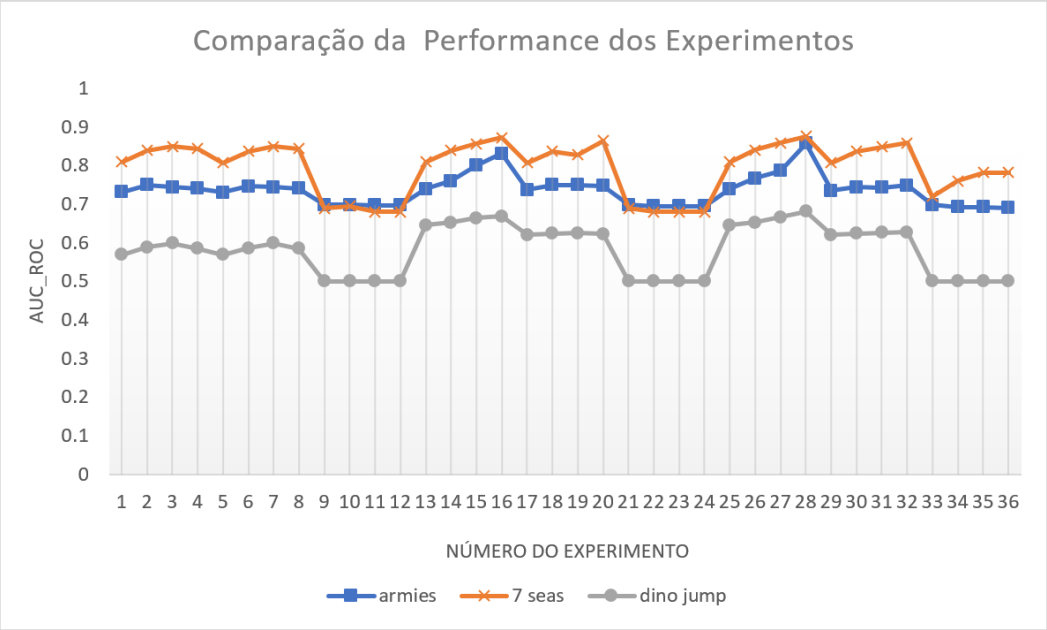

Figura 6-12: Comparação da performance dos experimentos para as três bases de dados.

Essa avaliação com base nas performances demonstra que as bases de dados apresentam correlação para as performances de tratamentos similares do projeto experimental. Essa informação é corroborada por meio da comparação dos efeitos principais nessas bases de dados (Figura 6-13). Esse gráfico demonstra que as médias das performances para os diferentes níveis dos experimentos apresentam correlação pois tendem à variar conjuntamente.

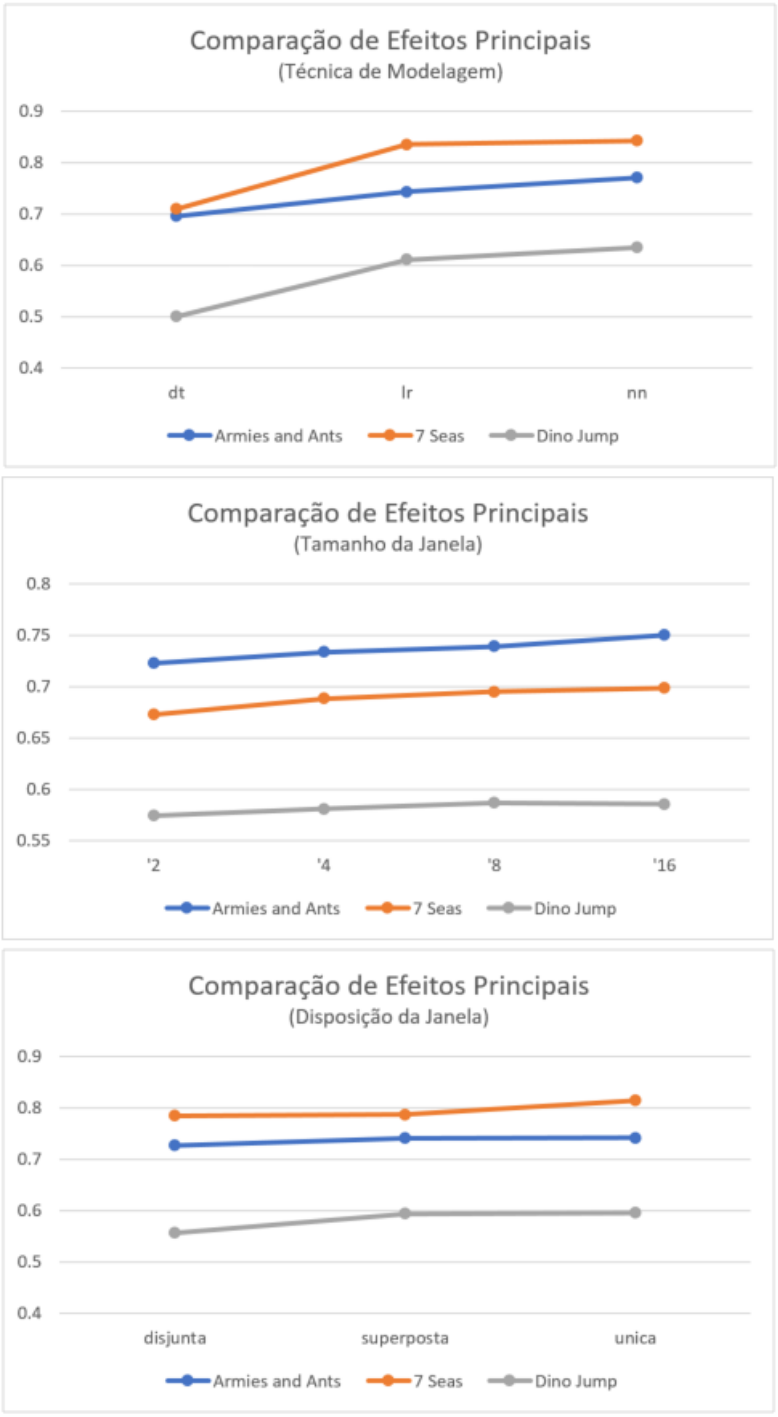

Figura 6-13: Comparação de Efeitos principais para os três principais fatores do experimento com contribuição significativa (>1%).

Para aprofundar a avaliação da relação entre os resultados, nós aplicamos o teste de correlação. O teste de correlação de Pearson demanda que os dados estejam normalmente distribuídos e valores extremos (*outliers*) podem afetar significativamente o resultado. Por esse motivo nós decidimos aplicar também a correlação Rô de Spearman dada a inexistência dessa premissa por se tratar de uma estatística não paramétrica. A correlação de Spearman funciona como o cálculo da correlação de Pearson não sobre os resultados diretos, mas sobre o ranking das performances. Os resultados das correlações são apresentados na Tabela 6-15.

Correlação de Pearson
(7seas, armies, dinojump)

	armies	7 seas
7 seas	0.788	
	0.000	
dino jump	0.881	0.870
	0.000	0.000

Conteúdo da Célula: Correlação de Pearson
Valor-p

Rô de Spearman
(7seas, armies, dinojump)

	armies	7 seas
7 seas	0.867	
	0.000	
dino jump	0.890	0.820
	0.000	0.000

Conteúdo da Célula: Rô de Spearman
Valor-p

Tabela 6-15: Resultado das correlações de Pearson (esquerda) e Rô de Spearman (direita).

De acordo com o Valor-P, todos os resultados são estatisticamente significativos. A análise dos resultados indica uma correlação muito forte entre os resultados alcançados para todas as bases. Essa forte correlação indica que a performance alcançada para os tratamentos apresenta relação positiva.

Em outras palavras, as performances dos classificadores nas diferentes bases de dados analisadas tendem a variar juntas e na mesma direção conforme a configuração do tratamento. Os tratamentos com alta performance (AUC acima da média) tendem a apresentar performance elevada em todos os jogos analisados. O contrário também é verdade, os tratamentos com baixa performance também tendem a apresentar performance inferior em todos os jogos. Esses resultados são discutidos em maior profundidade no próximo capítulo.

6.6 Conclusão e Observações

Essa seção apresentou em detalhes as bases de dados usadas nos experimentos e todo o processo executado para condução dos experimentos nos três jogos móveis reais com mais de 200 mil usuários. Os resultados alcançados nos experimentos realizados, 144 no total, foram avaliados à luz da análise de variância e análise de regressão. Além disso, a relação entre os

resultados foi avaliada através do teste de correlação. Na próxima seção nós realizamos a análise crítica dos resultados para identificação dos achados relevantes e discussão dos seus significados.

7 Análise dos Resultados

A análise dos resultados revela uma série de descobertas relevantes sobre a importância dos fatores analisados e os seus respectivos efeitos na performance de modelos para previsão de abandono em jogos móveis. Nesta seção nós discutimos os principais achados realizados com a condução dos experimentos e comparamos com as abordagens identificadas na revisão crítica do estado da arte. Esses achados servem como base para elaboração das diretrizes para construção de modelos preditivos de abandono em jogos móveis. As diretrizes são aplicadas na base de dados investigada para construção dos classificadores e comparação com as abordagens adotadas na revisão da literatura.

7.1 Achados Relevantes

A análise dos resultados demonstra que os fatores e as interações entre fatores (variáveis independentes) explicam 99,97% da variância da performance dos modelos de previsão de abandono (variável dependente ou variável resposta). Os fatores e as interações com contribuições estatisticamente significativas estão apresentados, em ordem decrescente de contribuição, na Tabela 7-1 e na Tabela 7-2, respectivamente. Os fatores respondem por 91,20% enquanto as interações entre fatores respondem por 8,77% da contribuição na alocação da variação com relação à performance do classificador.

O Técnica de Modelagem aparece como principal fator, com 82,08% de contribuição na alocação da variação. O Tamanho da Janela e a Disposição da Janela de performance aparecem com contribuição de 4,96 e 4,16%, respectivamente. Já o fator Tipos de Dados não apresenta contribuição (0,00%) na alocação da variação.

Fator	Contribuição	Valor-P
Técnica de Modelagem	82,08%	0,000
Tamanho da Janela de Performance	4,96%	0,000
Disposição da Janela de Performance	4,16%	0,000
Tipos de Dados	0,00%	0,017
Total	**91,20%**	

Tabela 7-1: Avaliação do resultado da ANOVA para os fatores do projeto experimental.

A interação entre os fatores é apresentada, em ordem de contribuição, na Tabela 7-2. A imagem revela que existe interações estatisticamente significativas e com contribuição relevante somente para duas das possíveis interações.

Interação	Contribuição	Valor-P
Disposição da Janela de Performance ∩ Técnica de Modelagem	5,77%	0,000
Disposição da Janela de Performance ∩ Tamanho da Janela de Performance	1,10%	0,000
Total	**6,87%**	

Tabela 7-2: Avaliação do resultado da ANOVA para as interações entre fatores relevantes (contribuição > 1%).

Com o propósito de avaliar em maiores detalhes os fatores de maior importância e contribuição relevante, nós realizamos a seguir a discussão individual dos fatores, dos seus níveis e das interações existentes com os demais fatores do projeto experimental. A ordem de discussão nas subseções seguintes é realizada em ordem decrescente de contribuição.

7.1.1 Técnica de Modelagem

O fator Técnica de Modelagem aparece como o principal fator com maior contribuição na alocação da variação (82,08%). Esse achado é uma surpresa. Vamos avaliar o porquê.

Durante a avaliação do estado da arte nós identificamos uma grande variação na performance de previsão de abandono. Os artigos avaliados através da métrica F1-score apresentam performance com média e desvio padrão de 70,7 e 23,1 enquanto os artigos

avaliados via AUC apresentam performance com média e desvio padrão de 0,809 e 0,072. A comparação direta entre as performances dos trabalhos não é possível dado que as pesquisas usam diferentes classificadores e também bases de dados distintas. Por outro lado, essa grande variação na performance fornece indícios de que a técnica de modelagem deveria ser um dos fatores com importante contribuição na alocação da variação.

Por outro lado, a modelagem é somente uma das etapas no processo de construção do modelo de previsão. As etapas anteriores relativas à seleção, pré-processamento e transformação dos dados consomem entre 50 e 80% do tempo total do projeto (Hall et al. 2011). E mais, essas etapas estão diretamente relacionadas à inclusão de conhecimento sobre o domínio de aplicação para construção de atributos (variáveis independentes). A expectativa era de que os demais fatores avaliados relacionados a essas atividades de processamento dos dados presentassem uma taxa de contribuição na alocação de variação mais expressiva. Os resultados são analisados a seguir.

7.1.1.1 Análise

Vamos avaliar mais a fundo os resultados encontrados para o fator Técnica de Modelagem e seus respectivos níveis. As técnicas de Redes Neurais e Regressão Logística apresentam as médias de performance 0,841 e 0,834, respectivamente, enquanto a Árvore de Decisão demonstra média de performance inferior, 0,708 (Figura 7-1). Esse achado era esperado e foi confirmado.

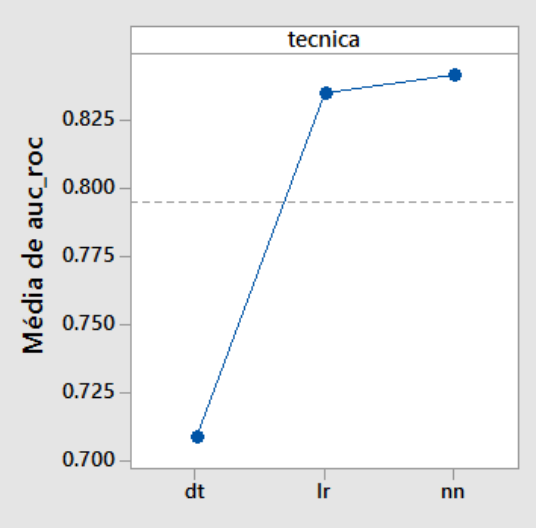

Figura 7-1: Gráfico de efeitos principais (Técnica de Modelagem).

A performance bastante inferior da Árvore de Decisão é responsável pelo aumento da variância desse fator. A avaliação direta da variância com os 3 níveis de fatores indica o valor de $5{,}60 \times 10^{-3}$ enquanto o cálculo da variância sem o nível Árvore de Decisão aponta uma redução expressiva para $2{,}44 \times 10^{-5}$. Para efeito de comparação, esse valor é menor do que a variância dos fatores Tamanho da Janela ($3{,}03 \times 10^{-4}$) e Disposição da Janela ($2{,}79 \times 10^{-4}$). Essa informação nos revela, portanto, que a exclusão do nível Árvore de Decisão implica na reordenação da importância contendo Tamanho da Janela e Disposição da Janela com maior contribuição da alocação da variação.

A baixa performance da técnica de Árvore de Decisão, por outro lado, confirma a informação de que as técnicas de Redes Neurais e Regressão Logística em geral apresentam performance mais elevada que a técnica de Árvore de Decisão (Dreiseitl & Machado 2002; Li et al. 2012). A discussão sobre esses achados é realizada a seguir.

7.1.1.2 Discussão

A avaliação dos dados demonstra que a técnica de Árvore de Decisão é a responsável pela elevada alocação da variação do fator Técnica de Modelagem. As técnicas de Redes Neurais e Regressão Logística, por outro lado, apresenta performance elevada e com valores médios similares.

$$Redes\ Neurais > Regressão\ Logística \gg Árvore\ de\ Decisão$$

A baixa performance da técnica de Árvore de Decisão é um achado relevante dos experimentos e avaliaremos mais a fundo esse aspecto. A baixa performance é provavelmente frutos de três aspectos: a limitação no tratamento de dados contínuos, a limitação da representação do espaço e a multicolinearidade dos dados. Nós vamos discutir cada um desses aspectos a seguir.

- Limitação no tratamento de dados contínuos

No processo de construção da árvore os dados são divididos de maneira otimizada entre cada nó da árvore, porém com perda de informação para variáveis contínuas. Essas variáveis são implicitamente discretizadas no processo de subdivisão promovendo a perda de informações ao longo do caminho. As árvores de decisão, assim como técnicas como KNN,

costumam apresentar performance inferior às técnicas de redes neurais e regressão logística (Dreiseitl & Machado 2002; Li et al. 2012).

- Limitação da representação do espaço

A provável explicação para a baixa performance da Árvore de Decisão é a limitação na representação do espaço definido pelos atributos em que cada folha corresponde a um hiper-retângulo onde a interseção destes é vazia e a união é todo o espaço. O resultado prático é que a árvore de decisão não promove a construção de uma superfície de decisão (hipersuperfície) para separação do espaço vetorial em dois conjuntos, um para cada classe. As imagens (Figura 7-2) demonstram o processo de divisão do espaço na qual a árvore de decisão subdivide o espaço em regiões cada vez menores enquanto a regressão logística divide o espaço em hiperplanos. É possível perceber que em problemas em que as classes não apresentam boa separação, a árvore de decisão torna-se mais suscetível ao *overfitting*, ou seja, quando o modelo se ajusta em demasia ao conjunto de dados com impacto na classificação de novas amostras. Esses fatos podem resultar em performances de classificação subótimas (Dreiseitl & Machado 2002).

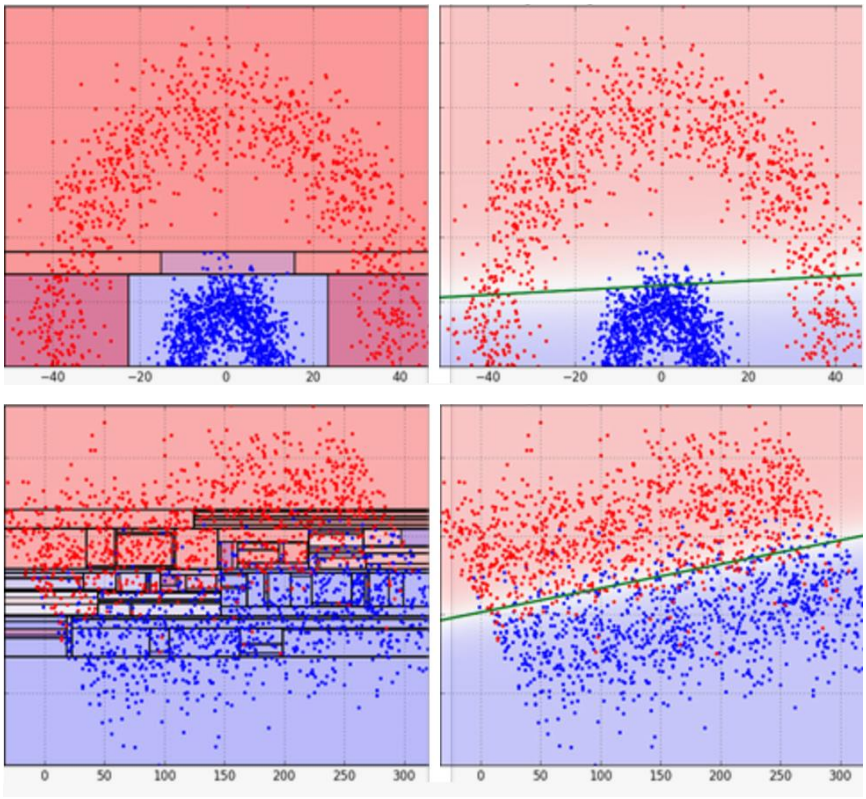

Figura 7-2: Ilustração do particionamento do espaço para as técnicas de Árvore de Decisão (esquerda) e Regressão Logística (direita) para dois exemplos distintos (cima e baixo).

- Multicolinearidade dos Dados

O problema de multicolinearidade dos dados acontece quando duas variáveis independentes explicam a mesma informação. As disposições de Superpostas e Disjuntas apresentam multicolinearidade em seus dados dado o processo utilizado para construção dos seus respectivos atributos. Essas disposições particionam a janela em intervalos menores e aplicam esses intervalos na construção dos dados. A janela superposta de 8 dias, por exemplo, inclui as informações da janela superposta de 4 dias.

Nesse cenário, a árvore de decisão, por utilizar uma abordagem gulosa (*greedy*), escolhe somente uma das variáveis independentes que explicam a mesma informação enquanto outros algoritmos costumam considerar ambas. Em geral, essa característica costuma ser útil para tratamento dos dados correlacionados, porém resultam na subutilização dos dados.

A avaliação da interação entre a técnica de modelagem e a disposição da janela apresenta indícios de que a multicolinearidade impacta na performance dado que justamente as disposições Superposta e Disjunta apresentam performance inferior (Figura 7-3).

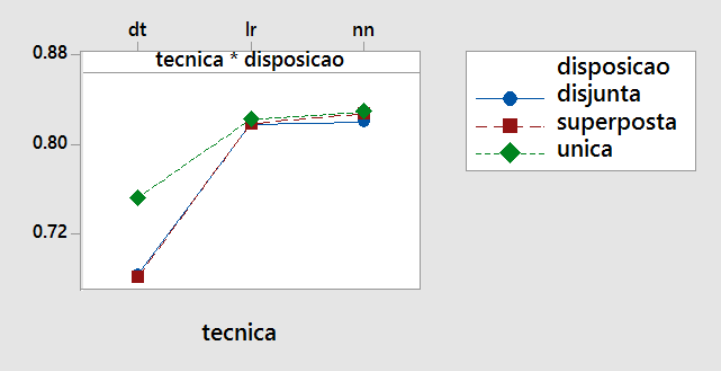

Figura 7-3: Gráfico de interação entre Técnica de Modelagem e Disposição da Janela.

7.1.1.3 Conclusões

A análise e discussão dos achados relevantes com relação às escolhas propostas para a decisão-chave relativa à Técnica de Modelagem revela informações importantes para a elaboração das diretrizes. As conclusões extraídas dessa análise são descritas a seguir:

1) Dado que o fator Técnica de Modelagem é responsável pela maior contribuição na alocação da variação, a recomendação geral é de que na construção de modelos preditivos sejam exploradas diferentes alternativas de técnicas de modelagem. É recomendado o uso das técnicas de redes neurais ou regressão logística em conjunto com outras alternativas dado que essas técnicas demonstraram performance elevada nos

experimentos. E também porque, conforme estudos, essas técnicas costumam apresentar performance elevada em comparação com outras técnicas.

2) Dado que a proposta do presente trabalho consiste na elaboração de diretrizes passíveis de serem aplicadas em jogos móveis nós optamos pela escolha de dados independentes da aplicação. A construção das variáveis independentes foi realizada basicamente a partir de dados relativos à sessão do usuário que permitem a fabricação de atributos como quantidade, duração, frequência, média, desvio padrão, dentre outros. Essas variáveis construídas a partir da mesma base apresentam maior chance de manifestar multicolinearidade e correlação entre si. Esses fatores apresentam pouco impacto sobre modelos baseados em regressão e associação além de técnicas como redes neurais, SVM e CART. Por outro lado, esses fatores tendem a apresentar impacto em técnicas como Árvore de Decisão (ID3 e C4.5) (Iniesta et al. 2016).

Essas conclusões são retomadas à frente para elaboração das diretrizes.

7.1.2 Tamanho da Janela de Performance

O fator Tamanho da Janela de Performance aparece como o segundo principal fator com maior contribuição na alocação da variação (4,96%), porém muito inferior ao fator Técnica de Modelagem. Esse achado é uma surpresa. Vamos avaliar o porquê.

A nossa expectativa era de que a etapa relativa à transformação dos dados a qual demanda a inserção de conhecimento do especialista e consome entre 50 e 80% do esforço da construção do classificador apresentasse uma importância maior na alocação da variação. Essa contribuição mais significativa também era aguardada dada a extensão adotada para o tamanho da janela com escolhas de 2, 4, 8 e 16 dias – em especial devido às janelas extremas, de 2 e 16 dias. Enquanto a janela de 16 dias captura informações sobre um período relevante em termos de ciclo de vida do usuário, a janela de 2 dias captura um volume de informações bastante reduzido. A expectativa era de que essa diferença conceitual estivesse refletida também na variação da performance dos modelos construídos a partir desses atributos. Os resultados são analisados a seguir.

7.1.2.1 Análise

Vamos avaliar mais a fundo os resultados encontrados para o fator Tamanho da Janela e seus respectivos níveis (Figura 7-4). Os tamanhos de janela 2, 4, 8 e 16 dias apresentam as médias de performance 0,770, 0,794, 0,803 e 0,811. A análise dos resultados demonstra que o crescimento no tamanho da janela implica no consequente aumento da performance média. Esse achado era esperado e confirmou-se. E mais, a análise indica que independentemente da disposição, tipo dos dados ou técnica de modelagem utilizada, o crescimento no tamanho da janela permite a extração de mais informações sobre o período recente do usuário.

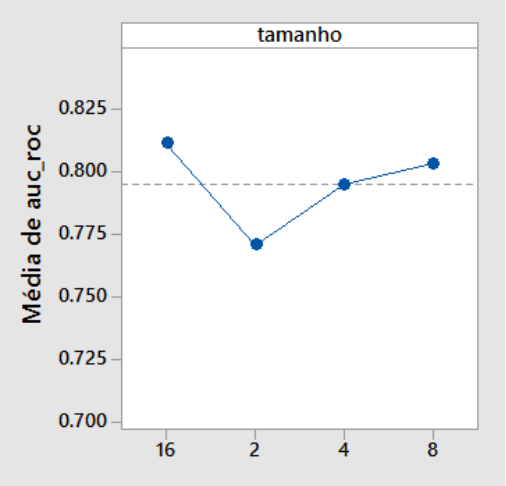

Figura 7-4: Gráfico de efeitos principais (Tamanho da Janela).

A análise da variância do Tamanho da Janela em comparação com a variância da Técnica de Modelagem após a exclusão da performance média da Árvore de Decisão indica que o fator Tamanho da Janela apresentaria maior variância. Com o propósito de isolar a avaliação do Tamanho da Janela, nós realizamos novo experimento com a técnica de Redes Neurais como único nível do fator. Esse novo experimento permite ainda remover o efeito negativo na performance provocada pela árvore de decisão. O resultado da ANOVA, assim como os gráficos de efeitos principais e de interação, são apresentados a seguir.

Análise de Variância

Fonte	GL	SQ Seq	Contrib.	QM (Aj.)	Valor F	Valor-P
Modelo	15	0.012063	99.96%	0.000804	1417.92	0.000
Linear	6	0.011313	93.74%	0.001885	3324.32	0.000
disposicao	2	0.000478	3.96%	0.000239	421.70	0.000
dados	1	0.000004	0.04%	0.000004	7.54	0.025
tamanho	3	0.010830	89.75%	0.003610	6364.99	0.000
Interações de 2 fatores	9	0.000750	6.22%	0.000083	146.99	0.000
disposicao*tamanho	6	0.000743	6.16%	0.000124	218.39	0.000
dados*tamanho	3	0.000007	0.06%	0.000002	4.19	0.047
Erro	8	0.000005	0.04%	0.000001		
Total	23	0.012068	100.00%			

Tabela 7-3: Resultado da ANOVA para o projeto experimental com o nível Rede Neural mantido estático (sem variação).

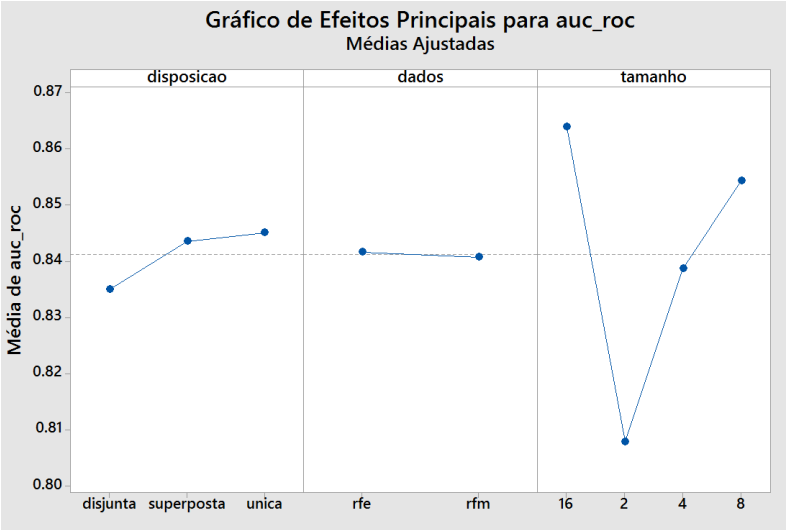

Figura 7-5: Gráfico de efeitos principais para o experimento com Redes Neurais como único classificador.

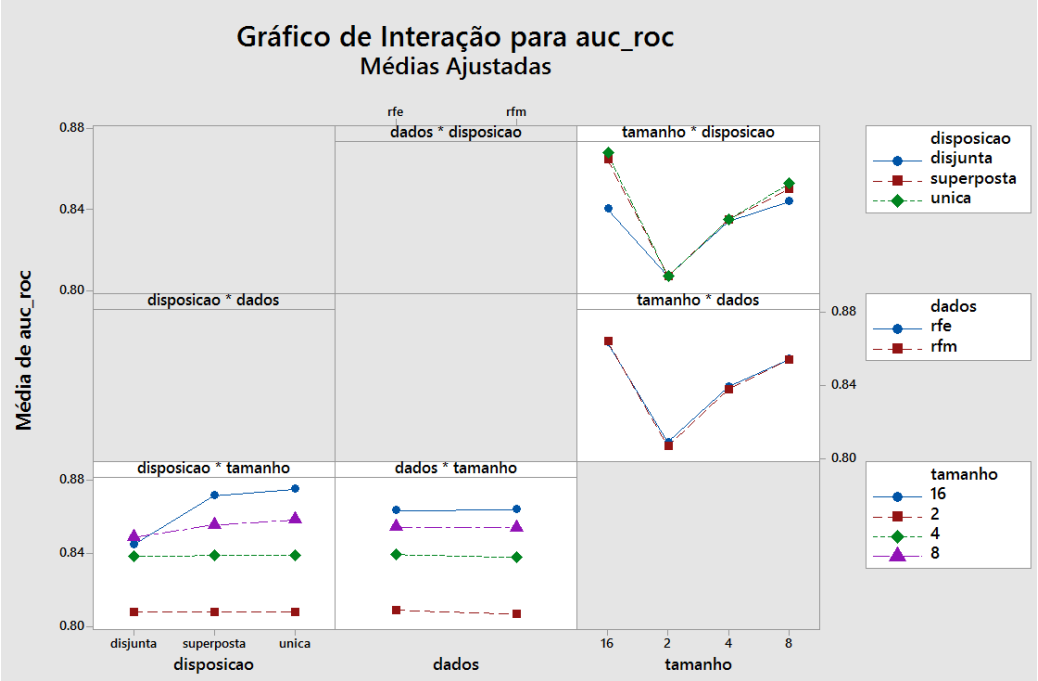

Figura 7-6: Gráfico de interação para o experimento com Redes Neurais como único classificador.

A discussão sobre esses achados e também sobre os resultados do novo experimento é realizada a seguir

7.1.2.2 Discussão

A análise confirma que após a escolha de uma técnica robusta, como redes neurais e regressão logística, o tamanho da janela é fator com maior influência na alocação da variação da performance. No novo experimento, a configuração do tamanho da janela responde por 89,75% da contribuição na variância seguido pela interação entre tamanho e disposição e o fator disposição com 6,22% e 3,96% de contribuição, respectivamente. A exclusão da regressão logística, e especialmente da árvore de decisão, implica no aumento da performance média para as janelas de tamanho 2, 4, 8 e 16 dias as quais apresentam performance média de 0,807, 0,838, 0,854 e 0,863.

A avaliação dessas performances, em conjunto com a análise gráfica (Figura 7-5), demonstra que a diferença entre a performance das janelas se aproxima de uma progressão geométrica decrescente de razão 0,5 com aumento de 3,84% (≈4%) entre os níveis 2 e 4, de 1,91% (≈2%) entre 4 e 8 e de 1,05% (≈1%) de aumento entre 8 e 16 dias. É provável, portanto, que o crescimento da janela para 32 dias represente um aumento pouco expressivo na performance do modelo preditivo.

A avaliação desse aumento de performance em comparação com as taxas de abandono médias da indústria (Figura 4-2) indica que a taxa de perda de usuários entre a primeira e a

segunda semana é maior do que o ganho gerado com o crescimento da performance. Enquanto o crescimento na performance de 8 para 16 dias é de 1,05%, o aumento na taxa de abandono de 7 para 14 dias é de 37,1%. Essa informação é calculada a partir do gráfico (Figura 4-2) em que 35% dos usuários estão retidos na primeira semana e somente 22% estão presentes na segunda semana.

A avaliação das interações através da ANOVA e do gráfico de interações (Figura 7-6) revela a interação entre Disposição e Tamanho com 6,16% de contribuição. A avaliação gráfica demonstra a performance inferior da janela Disjunta especificamente para os tamanhos de janela de 8 e 16 dias. O impacto na janela de 8 dias é sensível, mas perceptível. Já o impacto negativo na interação com a janela de 16 dias é mais expressivo. A performance média dessa combinação é inferior ao tamanho 8 dias.

A provável explicação para a baixa performance identificada nessa combinação é a configuração da Rede Neural, mais especificamente o número de neurônios na camada escondida. Para modelar relacionamento mais complexos a partir de um número maior de variáveis independentes, a rede neural requer uma maior quantidade de neurônios em suas camadas escondidas. Os experimentos realizados, entretanto, utilizam a mesma configuração para todos os tratamentos do experimento com somente uma camada escondida contendo 8 neurônios.

A definição do número de neurônios nas camadas escondidas em geral é realizada empiricamente dado que não se deve utilizar nem unidades demais, o que pode levar a rede a memorizar os dados de treinamento (*overfitting*), ao invés de extrair as características gerais que permitirão a generalização. E tampouco um número muito pequeno, que pode forçar a rede neural a gastar tempo em excesso tentando encontrar uma representação ótima. Há várias propostas para determinação da quantidade adequada de neurônios nas camadas escondidas de uma rede neural. A mais utilizada consiste na definição do número de neurônios em função da dimensão das camadas de entrada e saída da rede, conforme fórmula abaixo.

$$\#nós = \frac{\#entradas + \#saídas}{2} + 1$$

A aplicação dessa proposta para a janela Disjunta com 16 dias estipula a configuração da janela de performance com um valor entre $[28{,}41]$ neurônios, a depender dos tipos de dados usados (RFE ou RFM).

$$\#nós_{rfe} = \frac{8 \times 7 + 1}{2} + 1 \approx 28$$

$$\#nós_{rfm} = \frac{8 \times 10 + 1}{2} + 1 \approx 41$$

7.1.2.3 Conclusões

A análise e discussão dos achados relevantes com relação às escolhas propostas para a decisão-chave relativa ao Tamanho da Janela revela informações importantes para a elaboração das diretrizes. As conclusões extraídas dessa análise são descritas a seguir:

1) Essa expansão na quantidade de dados disponíveis fornece subsídios para o aumento no potencial discriminatório dos classificadores na atividade de separação dos usuários *churners* e *non-churners*. Os resultados indicam que as melhores performances são alcançadas com janelas próximas a 16 dias.

2) Por outro lado, a taxa de abandono na indústria é elevada nas primeiras semanas conforme dados da indústria (Figura 4-2) e, portanto, a recomendação é que será realizado um compromisso entre o aumento da performance referente ao tamanho da janela em comparação com a taxa de abandono no período. É preciso levar em consideração o ponto de operação a ser utilizado em cada situação. Porém, dada a elevação na taxa de abandono entre 7 e 14 dias de 37,1% a recomendação geral é que seja utilizada uma janela de performance igual ou menor do que 8 dias para construção de um classificador mais efetivo na retenção de jogadores.

3) Dada a diferença de performance identificada para a rede neural na interação com a disposição da janela e o tamanho da janela 16 dias, nós acreditamos que essa situação tenha ocorrido devido à uma configuração sub-ótima da rede neural. É de conhecimento geral que não há fórmula “mágica” para a configuração das redes neurais, porém é aconselhada a utilização das melhores práticas fundamentadas em conhecimento empírico (Haykin 1999).

Essas conclusões são retomadas à frente para elaboração das diretrizes.

7.1.3 Disposição da Janela de Performance

O fator Disposição da Janela de Performance aparece como o terceiro principal fator com maior contribuição na alocação da variação (4,16%), porém muito inferior ao fator Técnica de Modelagem. Esse achado também é, de certa forma, uma surpresa. Vamos avaliar o porquê.

A nossa expectativa era de que a aplicação de diferentes disposições de janelas com características mais informativas do que a janela Única fossem permitir a construção de classificadores com uma variância mais expressiva. As janelas Disjuntas e Superpostas permitem a inclusão de dados relativos à recência para prover informação em teoria mais relevante para a classificação dos usuários com maior propensão ao abandono. A contribuição na alocação da variância demonstrou-se baixa, porém nós imaginávamos que esse fator seria o terceiro ou quarto fator mais importante como previsto. Os resultados são analisados a seguir.

7.1.3.1 Análise

Vamos avaliar mais a fundo os resultados encontrados para esse fator e seus respectivos níveis. As disposições Única, Superposta e Disjunta apresentam as médias de performance 0,814, 0,786 e 0,784. A análise dos resultados demonstra que o crescimento no tamanho da janela implica no consequente aumento da performance média. A avaliação da Figura 7-3 apresenta a interação significativa entre a técnica de modelagem e a disposição. Essa interação é avaliada na seção 7.1.1.2 a qual revela que a árvore de decisão é responsável pela redução da performance para os níveis Superposta e Disjunta, dado que para as demais técnicas de modelagem a performance apresenta menor variância.

Para isolar essa interação e avaliar mais a fundo o impacto de cada um dos níveis desses fatores, nós vamos avaliar o experimento adicional realizado e apresentado na Tabela 7-3 com a manutenção somente dos tratamentos com Redes Neurais. É possível perceber nesse experimento que a disposição da janela representa ainda uma contribuição menos expressiva (3,96%) e também que a relação de importância entre os níveis de fatores não foi alterada (Figura 7-7).

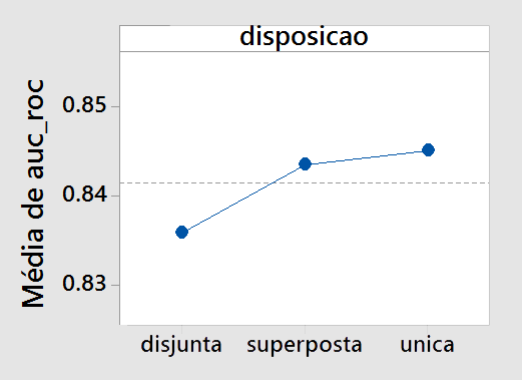

Figura 7-7: Gráfico de efeitos principais para o experimento com Redes Neurais (somente).

A reavaliação das médias de performance no experimento com a rede neural como técnica exclusiva aponta para uma performance média para as disposições Única, Superposta e Disjunta de 0,845, 0,844 e 0,836, respectivamente. Nesse segundo experimento, sem o impacto negativo da interação com a técnica de árvore de decisão, a performance dos níveis de fatores Única e Superposta tornam-se ainda mais próximos. Esse achado está dentro da expectativa dado que a disposição Única está contida dentro da Superposta.

A avaliação gráfica dos efeitos principais demonstra que a disposição Única apresenta a maior performance média em comparação com os demais níveis desse fator. A análise do gráfico de interações (Figura 7-7) também confirma que essa disposição apresenta média superior também nas interações com técnica de modelagem e tamanho da janela. A discussão sobre esses achados e também sobre os resultados do novo experimento é realizada a seguir.

7.1.3.2 Discussão

A análise confirma que após a escolha de uma técnica robusta, como redes neurais e regressão logística, e a seleção do tamanho da janela, de acordo com o *tradeoff* desejado, a disposição da janela é o terceiro fator com maior influência na alocação da variação da performance. A descoberta realizada é uma surpresa dado que a expectativa era de que as demais janelas apresentassem performance igual ou superior à disposição Única.

Na tentativa de compreender esse resultado, nós avaliamos as variáveis independentes construídas em cada uma dessas alternativas para buscar por pistas. Dentre várias análises realizadas, uma informação em particular desperta a atenção (Figura 7-8). As taxas elevadas de campos com valores zero nas variáveis independentes criadas pode ser um fator relacionado à performance. As janelas Superposta e Disjunta, justamente por considerarem partições menores do tempo dentro da janela de performance, produzem um número maior de valores zero. E são

também os níveis de fatores com menor performance. É preciso investigar experimentalmente essa possível explicação para avaliar o seu real impacto.

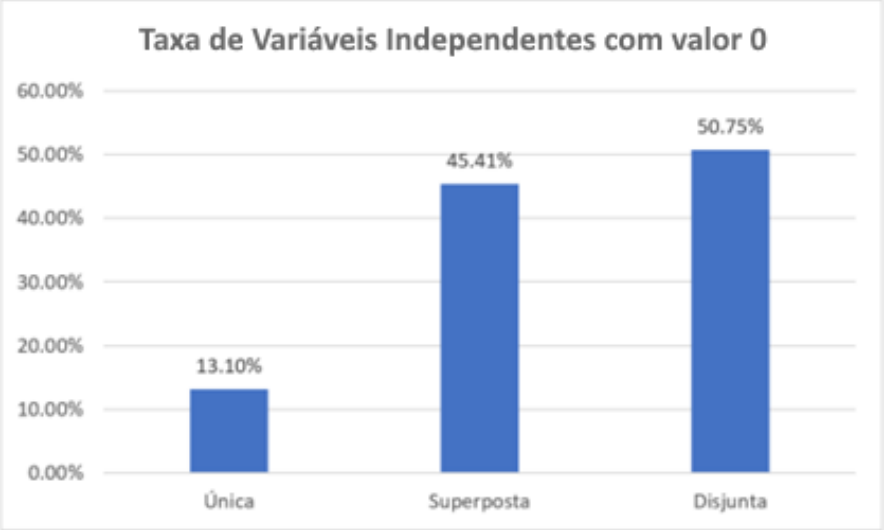

Figura 7-8: Avaliação da taxa de valores ausentes nos atributos (variáveis independentes) para cada uma das disposições.

7.1.3.3 Conclusões

A análise e discussão dos achados relevantes relativas à Disposição da Janela revela informações importantes para a elaboração das diretrizes. A conclusão extraída dessa análise é descrita a seguir:

1) Dada que a janela do tipo Única apresenta maior performance na avaliação dos efeitos principais e também na avaliação dos efeitos na interação com os demais fatores, a recomendação geral é que a janela Única seja aplicada na construção de classificadores da previsão nesse domínio. Essa janela apresenta benefícios ainda com relação à simplicidade na produção dos seus atributos que demanda menos esforço e tempo em comparação com os tipos Disjunta e Superposta.

Essas conclusões são retomadas à frente para elaboração das diretrizes.

7.1.4 Tipos de Dados

Os tipos de dados apresentam importância nula na alocação da variação por apresentarem 0,00% de contribuição, segundo a ANOVA. A análise de variância também não

demonstra nenhuma interação estatisticamente significativa dos Tipos de Dados com os demais fatores do experimento. Essa descoberta não é uma surpresa.

A nossa expectativa era que a expansão das informações comportamentais da análise RFE com a inclusão de dados monetários relativas ao histórico de compras para construção do tipo de dados RFM de fato não apresentasse contribuição significativa no aumento do poder discriminatório dos modelos preditivos de abandono. Em outros domínios estudados na revisão do estado da arte as variáveis monetárias relativas a histórico de compras e pagamentos do usuário se mostram úteis na classificação do usuário com relação à sua propensão ao abandono. Entretanto, em jogos móveis a baixa taxa média de conversão de usuários (2,2%) implica na construção de atributos relativos à monetização para uma parcela pequena da população e com provável impacto reduzido na performance da classificação. Uma abordagem possível, porém fora do escopo deste trabalho, seria a utilização de modelos distintos de previsão de abandono para os usuários que pagam daquees que não pagam. A discussão sobre esses achados e também sobre os resultados do novo experimento é realizada a seguir

7.1.4.1 Análise

O resultado dos experimentos não revela interação significativa com outros fatores a serem comentadas nesse trabalho. Por outro lado, a avaliação do poder discriminatório das variáveis dependentes usadas na pesquisa permite expandir a compreensão sobre quais atributos apresentam maior contribuição na classificação, segundo a técnica de Regressão Logística.

Coeficientes

Termo	Coef	Valor-P
ow_disj_rfe_session_freq_12	0.11810	0.00000
ow_disj_rfe_session_freq_8	0.08930	0.00200
ow_disj_rfe_session_freq_14	0.06620	0.00400
ow_supp_rfe_absence_total_2	0.02110	0.00000
ow_disj_rfe_absence_average_12	0.01598	0.00000
ow_disj_rfe_absence_average_4	0.01414	0.00000
ow_disj_rfe_absence_total_4	0.01122	0.00000
ow_supp_rfe_absence_total_16	0.01004	0.00000
ow_supp_rfe_absence_average_2	0.00941	0.04100
ow_disj_rfe_absence_average_16	0.00928	0.00000
ow_disj_rfe_absence_total_8	0.00773	0.00000
ow_disj_rfe_absence_average_8	0.00698	0.00700
ow_disj_rfe_absence_total_12	0.00680	0.00000

ow_disj_rfe_absence_average_10	0.00614	0.00900
ow_disj_rfe_absence_total_14	0.00600	0.00000
ow_disj_rfe_absence_total_6	0.00545	0.00500
ow_disj_rfe_absence_total_16	0.00441	0.00800
ow_supp_rfe_playtime_total_4	0.00001	0.00000
ow_disj_rfe_playtime_average_8	-0.00005	0.00200
ow_supp_rfe_playtime_average_4	-0.00006	0.00400
ow_supp_rfe_playtime_average_2	-0.00028	0.00000
ow_disj_rfm_last_purchase_12	-0.00517	0.01000
ow_supp_rfe_absence_average_16	-0.00719	0.00000
ow_supp_rfe_absence_total_4	-0.00929	0.00000
ow_disj_rfm_purchase_total_12	-0.01452	0.00400
ow_supp_rfe_session_freq_2	-0.41480	0.00000

Tabela 7-4: Avaliação dos coeficientes da regressão logística.

A avaliação dos coeficientes, por meio da sua magnitude e sinal, provê informações relevantes sobre a contribuição das variáveis estatisticamente significativas na previsão de abandono dos jogadores.

7.1.4.2 Discussão

A análise dos coeficientes da regressão revela que as variáveis com maior magnitude apresentam maior relevância na determinação da situação do usuário. O aumento dos valores dos atributos com coeficientes negativos tende a reduzir a probabilidade de abandono do usuário. Isso significa que o aumento na frequência de sessões nos 2 dias anteriores do ponto de observação tende a reduzir as chances de abandono do usuário.

O contrário também é verdadeiro, ou seja, a redução na magnitude dos valores dos atributos com coeficientes positivos tende a aumentar a probabilidade de abandono do jogador. Isso significa que a redução no tempo de ausência nos 2 dias anteriores ao ponto de observação contribui na redução da probabilidade de abandono. A regressão também achou relação entre o aumento da ausência total e média dos últimos 4 dias com o aumento da chance de abandono. As variáveis com magnitude do coeficiente acima de 0,01 são apresentados na Tabela 7-5.

Atributos com Coeficiente Positivo	Atributos com Coeficiente Negativo
ow_disj_rfe_session_freq_12	
ow_disj_rfe_session_freq_8	ow_supp_rfe_session_freq_2
ow_disj_rfe_session_freq_14	ow_disj_rfm_purchase_total_12
ow_supp_rfe_absence_total_2	
ow_disj_rfe_absence_average_12	
ow_disj_rfe_absence_average_4	
ow_disj_rfe_absence_total_4	
ow_supp_rfe_absence_total_16	

Tabela 7-5: Apresentação dos coeficientes de maior magnitude da regressão logística.

7.1.4.3 Conclusões

A análise e discussão dos achados relevantes relativas aos Tipos de Dados revela informações importantes para a elaboração das diretrizes. A conclusão extraída dessa análise é descrita a seguir:

1) Os tipos de dados relativos ao tempo de ausência, frequência de sessões e duração total de jogo estão presentes com maior frequência na lista de atributos estatisticamente significativos. A recomendação geral, portanto, é que esses tipos de atributos sejam explorados na construção de classificadores da previsão nesse domínio por meio da definição dos seus valores totais, médios, frequência, desvio padrão e variância, dentre outros. Esses tipos de dados estão conectados semanticamente por representarem o engajamento do usuário com o jogo.

Essas conclusões são retomadas à frente para elaboração das diretrizes.

7.1.5 Discussão sobre os Resultados para as Bases de Jogos

Com o objetivo de avaliar sobre possíveis indícios da capacidade de generalização das descobertas realizadas para aplicação em jogos dentro do mesmo domínio, nós realizamos os experimentos em 3 bases de dados de jogos reais. Os jogos utilizados (7 Seas, Dino Jump e Armies and Ants) são todos pertencentes ao mesmo domínio proposto, ou seja, consistem de jogos para dispositivos móveis lançados sobre o modelo de negócios Free-to-Play.

A avaliação desses indícios foi realizada por meio da análise de correlação da performance dos modelos preditivos construídos para os tratamentos propostos no projeto experimental. Os jogos Dino Jump e Armies and Ants não apresentam informações monetários e portando esses dados não foram considerados na análise, resultando na exclusão do nível de fator Análise RFM do fator Tipos de Dados. Dado o resultado dos projetos experimentais em que os tipos de dados contribuem com 0,00% na alocação da variação, nós estimamos que a perda pela ausência desses dados é pouco significativa.

A avaliação por meio da correlação de Pearson e de Spearman revelam que os resultados apresentam alta grau de correlação. A avaliação da correlação, realizada dois a dois, é sumarizada na Tabela 7-6.

Bases de Dados	**Correlação de Pearson**	**Rô de Spearman**
7 Seas x Dino Jump	0,870	0,820
7 Seas x Armies and Ants	0,788	0,867
Dino Jump x Armies and Ants	0,881	0,890

Tabela 7-6: Resultados da correlação entre os resultados experimentais para as 3 bases de dados.

Esse achado sugere que as escolhas avaliadas para as decisões-chaves na construção dos modelos preditivos são passíveis de serem utilizadas em jogos do mesmo domínio. A alta correlação não prova que as conclusões e diretrizes propostas com análise crítica do projeto experimental são aplicáveis a todos os jogos móveis, porém provê indícios de que as diretrizes possam ser aplicadas com sucesso em outros jogos móveis Free-to-Play.

7.2 Diretrizes

O planejamento, execução e análise do projeto experimental nos permitiu avaliar o impacto das possíveis escolhas para as decisões-chaves relativas ao processo de construção de modelos preditivos de abandono em jogos móveis. Nessa seção nós sintetizamos as descobertas realizadas na forma de diretrizes para compartilhar boas práticas a serem utilizadas em trabalhos no mesmo domínio de aplicação.

7.2.1 Decisão-Chave: Técnica de Modelagem

A escolha recomendada para essa decisão-chave consiste na seleção de técnicas de modelagem que promovam a construção de uma superfície de decisão (menos suscetíveis ao *overfitting*). Além disso, recomenda-se que sejam avaliadas diferentes técnicas para identificação daquela que melhor se adapta aos dados tendo na lista de alternativas as opções de Redes Neurais e Regressão Logística por apresentarem performance elevada nos 144 experimentos conduzidos nessa tese.

Essa recomendação em geral não é seguida pela maioria dos trabalhos avaliados que aplicam somente uma técnica de modelagem. Além disso, uma parcela significativa dos trabalhos modelou o problema através da técnica de árvore de decisão apontada em nosso estudo por potencialmente apresentar performance inferior.

7.2.2 Decisão-chave: Tamanho da Janela de Performance

A escolha recomendada para essa decisão-chave consiste na seleção de janelas com tamanho igual ou inferior a 8 dias. Apesar da performance superior atrelada à janela de 16 dias também avaliada experimentalmente, a aplicação de janelas com tamanho igual ou inferior a 8 dias permite identificar com uma antecedência maior os usuários com propensão ao abandono e atuar de maneira mais efetiva na retenção dos jogadores.

Com base na análise da evolução das taxas de abandono na indústria de jogos móveis, nós identificamos um aumento expressivo de 37,1% na taxa de abandono entre 7 e 14 dias. Dado que o aumento de performance de 8 para 16 dias é pouco expressivo, a recomendação é que a técnica seja aplicada o mais brevemente possível para permitir a aplicação de ações de retenção e fidelização do usuário para prevenção do abandono. Essa recomendação visa, portanto, não a identificação da performance mais efetiva, mas a construção de classificadores que permitam a descoberta de conhecimento acionável.

7.2.3 Decisão-chave: Disposição dos Dados

A escolha recomendada para essa decisão-chave consiste na seleção da disposição Única. A pesquisa indica que dentre as janelas mais populares na abordagem de Behavior Scoring, a janela Única é a responsável por apresentar performance superior na comparação dos e feitos principais e ainda na avaliação da interação com os demais fatores.

Esse tipo de janela implica em um processo de construção e preparação dos dados mais simples e menos custosa. A implicação prática é que o tempo consumido na construção de

janelas mais complexas como a Superposta e Disjunta pode ser economizado para aplicação das demais diretrizes.

7.2.4 Decisão-chave: Tipos de Dados

A escolha recomendada para essa decisão-chave consiste, em primeira instância, na preferência de uso de dados comportamentais independentes da aplicação e relacionados ao engajamento do usuário. Os atributos relativos ao tempo de ausência, frequência de sessões e duração total de jogo apresentaram maior poder discriminatório com base na análise de regressão e devem ser explorados no processo de construção de modelos preditivos.

A análise sugere ainda que a inclusão de dados monetários não confere maior poder discriminatório aos classificadores construídos a partir de dados comportamentais. Além disso, a pesquisa também indica que classificadores com performance elevada podem ser construídos a partir de dados independentes da aplicação.

Além das diretrizes relacionados às escolhas relativas às decisões chaves, a avaliação do estado da arte e das especificidades de jogos também permite propor recomendações e boas práticas mais gerais.

7.2.5 Outras Recomendações

As diretrizes propostas visam auxiliar no processo de construção de modelos preditivos e avançar o estado da arte no domínio de jogos móveis. Nessa direção, além das diretrizes nós elencamos algumas recomendações para auxiliar especialmente na comparação de modelos preditivos.

- Uso da métrica Área sob a Curva ROC

A avaliação crítica do estado da arte revela que uma parcela significativa dos trabalhos ainda usa a precisão como a principal métrica de performance. Essa métrica, entretanto, não é a mais apropriada como visto na seção 2.3.5 por ser tendenciosa com relação ao desequilíbrio na distribuição das classes (Fawcett & Provost 1997) e por não considerar o custo de classificação incorreta (Baesens et al. 2003; West 2000/9). É recomendado o uso da AUC por ser robusta com relação a esse tipo de problema e permitir abstrair a decisão sobre o melhor ponto de operação para a etapa final do CRISP-DM referente à implantação.

- Disponibilização da base de dados

A comparação entre pesquisas diferentes nesse domínio também dependem diretamente da base de dados utilizada a qual geralmente não é disponibilizada pelos autores. É recomendado que as bases de dados sejam compartilhadas para permitir em primeiro lugar a replicabilidade dos experimentos. Em segundo lugar, permitir a troca de experiências e comparação de soluções aplicadas sobre o mesmo conjunto de dados.

7.3 Aplicação das Diretrizes

A aplicação das diretrizes propostas no próprio jogo 7 Seas utilizado nos experimentos permite a construção de um classificador com performance acima da média identificada nos artigos avaliados. O uso das boas práticas propostas conduz à seleção do experimento com a configuração descrita a seguir.

- Técnica de Modelagem: Redes Neurais;
- Tamanho da Janela de Performance: 8 dias;
- Disposição da Janela de Performance: Única
- Tipos de Dados: Análise RFE

A avaliação do classificador para previsão de abandono é realizada com base na métrica área sob curva ROC para a qual é identificada o valor de 0,858. A performance acima da média desse classificador é um efeito colateral positivo da pesquisa realizada com foco no estudo das especificidades do domínio de jogos móveis para definição de diretrizes.

Para expandir a análise da aplicação das diretrizes, nós comparamos a performance desse classificador com a performance das pesquisas. A única restrição foi a realização da comparação para as bases de dados performance medida também por meio de AUC ROC. A avaliação está detalhada na Tabela 7-7.

Base de Dados	Técnica de Modelagem	Tamanho da Janela	Disposição da Janela	Tipos de Dados	AUC ROC
Monster World	Rede Neural	14 dias	Única + Disjunta	série temporal dos *logins*, saldo de moedas virtuais	0,930
7 Seas	Rede Neural	16 dias	Única	tempo de ausência, frequência de sessões e duração total de jogo	0,876
7 Seas	Rede Neural	8 dias	Única	tempo de ausência, frequência de sessões e duração total de jogo	0,858
Diamond Dash	Rede Neural	14 dias	Única + Disjunta	série temporal das partidas, precisão dos movimentos, convites enviados, dias no jogo, última compra, dias desde a última compra	0,815
Heavy Metal Machine	kNN	7 dias	Disjunta	série temporal das sessões de jogo	0,791
APB: Reloaded	kNN	15 dias	Disjunta	série temporal das sessões de jogo	0,759
RF Online	kNN	30 dias	Disjunta	série temporal das sessões de jogo	0,750

Tabela 7-7: Análise comparativa da performance do classificador gerado a partir das diretrizes com o estado da arte.

A avaliação compara os classificadores construídos para as bases de dados do jogo 7 Seas, os jogos Monster World e Diamond Dash (Runge et al. 2014), e também os jogos Heavy Metal Machine, APB:Reloaded e RF Online (Castro & Tsuzuki 2015). Os classificadores com melhor performance dessa comparação (vide Monster World 7 Seas e Diamond Dash) usam Redes Neurais, apresentam tamanho de janela entre 8 e 16 dias e aplicam a disposição de janela Única. E mesmo com relação aos dados é possível perceber que parte dos atributos selecionados para os jogos Monster World e Diamond Dash (Runge et al. 2014) estão relacionados ao engajamento do usuário como frequência de *logins* e sessões e compras. Essa avaliação

novamente provê indícios de que as diretrizes possam ser aplicadas com sucesso em outros jogos móveis Free-to-Play, porém não é possível comparar diretamente resultados em bases de dados diferentes.

Essa análise da Tabela 7-7 associada a análise de correlação da Tabela 7-6 fornecem indícios de que as diretrizes são passíveis de serem aplicadas em outros jogos móveis.

8 Conclusão e Trabalhos Futuros

A previsão de abandono em jogos móveis é uma atividade essencial para antecipar a intenção do jogador de descontinuar a relação com jogo, permitindo às empresas desenvolvedoras de jogos aplicar ações proativas de retenção e fidelização. A construção de modelos preditivos de abandono de usuários de jogos móveis é dependente do domínio e das suas especificidades. Em jogos móveis, a pesquisa nessa área é ainda incipiente. Os poucos trabalhos identificados na revisão bibliográfica tomam muitas decisões ad hoc, aplicam metodologias genéricas baseadas na mineração de dados tradicional orientada a dados, não fazem uma discussão profunda sobre as especificidades do domínio de jogos e seus possíveis impactos na construção e na performance do modelo preditivo. Enfim, não há ainda na literatura claras diretrizes ou práticas comuns para a construção de modelos preditivos de abandono em jogos móveis.

8.1 Objetivos e Contribuições

Para avançar no estado da arte, por meio da aplicação da abordagem D3M e BS já aplicada com sucesso em outras áreas de aplicação, nós aprofundamos a discussão sobre as especificidades da área de abandono em jogos móveis. Em seguida, identificamos decisões-chave a serem tomadas no processo de construção do modelo preditivo, assim como as principais escolhas possíveis para tais decisões. Enfim, realizamos vários experimentos a fim de avaliar cada uma dessas escolhas em termos de desempenho da predição.

Os experimentos, realizados em 3 bases distintas de jogos contemplando 201.146 jogadores, puderam revelar algumas diretrizes importantes que serão úteis a pesquisadores e desenvolvedores quando forem construir modelos preditivos para abandono em jogos móveis. Assim, pudemos cumprir o objetivo central desta tese.

A primeira contribuição do presente trabalho consiste na avaliação das especificidades de jogos para identificação das características únicas desse domínio com impacto no processo de construção dos modelos preditivos de abandono de jogadores. Essa avaliação indica os desafios relacionados às decisões-chaves do processo e ao mesmo tempo propõe possíveis escolhas candidatas a diretrizes no domínio de jogos móveis.

A segunda e principal contribuição do trabalho consiste nas diretrizes propostas com base na avaliação, interpretação e discussão dos resultados experimentais. As diretrizes foram avaliadas a partir da avaliação da correlação dos resultados entre as três diferentes bases de dados usados nos experimentos. A avaliação demonstra forte correlação nos resultados fornecendo indícios da capacidade de generalização das diretrizes em jogos móveis.

8.2 Limitações

A presente tese apresenta uma série de limitações dentre as quais se destacam as seguintes. O número reduzido de jogos móveis considerados na condução e avaliação dos experimentos consiste em uma das principais limitações da tese. A utilização de somente 3 bases de dados não torna possível afirmar categoricamente que os achados são passíveis de generalização para todos o domínio de jogos móveis.

A impossibilidade de comparação efetiva dos resultados encontrados com a performance encontrada na revisão da literatura também é uma das limitações importantes. O uso de bases de dados diferentes nas pesquisas aliada à ausência de descrição detalhada de todos os passos executados na construção dos classificadores e também a aplicação de métricas de performance distintas inviabilizam a comparação dos resultados alcançados em trabalhos distintos. Essa limitação impõe barreiras à pesquisa de melhores soluções para avanço no estado da arte.

8.3 Trabalhos Futuros

Como trabalhos futuros nós planejamos expandir o trabalho realizado nesta tese para investigar os aspectos listados abaixo com o propósito de avançar no estado da arte.

- Realização dos experimentos em um número maior de jogos móveis para reduzir a limitação apresentada.

- Implementação das técnicas aplicadas nos artigos avaliados no estado da arte para permitir a comparação com os classificadores construídos a partir das diretrizes propostas.
- Avaliação aprofundada das possíveis categorias de atributos e seus respectivos impactos na performance de modelos preditivos. Essa é uma área ampla e demanda um estudo específico e direcionado.
- Avaliação do impacto da configuração das partições construídas através das disposições de janelas de performance do tipo Superposta e Disjunta. Nesta tese nós aplicamos a configuração com elemento inicial 2, quociente 2 (Superposta) e razão 2 (Disjunta), porém o impacto dessa decisão pode ser avaliado com maior profundidade.
- Expansão dos experimentos para áreas correlatas com especificidades similares ao domínio de jogos móveis como aplicativos móveis e jogos Massivos Multiplayer Online também disponibilizados sobre o modelo free-to-play. Nós acreditamos que as diretrizes e resultados colaterais (classificador com performance elevada) alcançados nesta tese podem ser aplicados também nesses domínios, porém é preciso investigar e validar essa hipótese.
- Uso dos dados de consumo (monetários) para segmentação dos usuários e avaliação da melhoria na previsão de abandono.
- Aplicação de métodos de otimização para identificação das melhores configurações dos parâmetros das técnicas de modelagem.

Referências

Abdou, H.A., 2009. Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert systems with applications*, 36(9), p.11402–11417.

Adomavicius, G. & Tuzhilin, A., 2001. Expert-Driven Validation of Rule-Based User Models in Personalization Applications. *Data mining and knowledge discovery*, 5(1-2), p.33–58.

Advisors, W., 2015. 5 Reasons why Customer Retention is better than Customer Acquisition | Wheelhouse Advisors. Available at: http://www.wheelhouseadvisors.net/5-reasons-why-customer-retention-is-better-than-customer-acquisition/ [Acessado janeiro 30, 2017].

Albuquerque, M. T. C. F., Ramalho, G.L., Corruble, V., Santos, A.L.S., Freitas, F. 2014. Helping Developers to Look Deeper inside Game Sessions. *XIII Brazilian Symposium on Games and Digital Entertainment, pp: 31-40.*

ACM Digital Library. Available at: http://dl.acm.org/ [Acessado janeiro 11, 2017a].

App Annie Special Report: A Look at the Growth of Google Play. *App Annie Content*. Available at: https://www.appannie.com/insights/google-io-special-report-launch-2014/ [Acessado janeiro 8, 2017b].

CiteSeerX. Available at: http://citeseerx.ist.psu.edu/ [Acessado janeiro 11, 2017c].

Engineering Village - First choice for serious engineering research. Available at: http://www.engineeringvillage2.org [Acessado janeiro 11, 2017d].

IEEE Xplore Digital Library. Available at: http://ieee.org/ieeexplore [Acessado janeiro 11, 2017e].

ScienceDirect.com | Science, health and medical journals, full text articles and books. Available at: http://www.sciencedirect.com [Acessado janeiro 11, 2017f].

Scopus. Available at: https://www.scopus.com/ [Acessado janeiro 11, 2017g].

Tablet ultrapassa vendas de desktop e notebook pela 1ª vez no Brasil. (2016)*Tecnologia e Games*. Available at: http://g1.globo.com/tecnologia/noticia/2014/03/tablet-ultrapassa-vendas-de-desktop-e-notebook-e-pela-1-vez-no-brasil.html [Acessado janeiro 11, 2017].

The Global Games Market 2016 | Per Region & Segment | Newzoo. (2016) *Newzoo*. Available at: https://newzoo.com/insights/articles/global-games-market-reaches-99-6-billion-2016-mobile-generating-37/ [Acessado janeiro 7, 2017].

AppBrain, 2017. Number of available Android applications - AppBrain. Available at: https://www.appbrain.com/stats/number-of-android-apps [Acessado janeiro 8, 2017].

Au, W.-H. & Chan, K.C.C., 2003. Mining fuzzy association rules in a bank-account database. *IEEE Transactions on Fuzzy Systems*, 11(2), p.238–248.

Au, W.-H., Chan, K.C.C. & Yao, X., 2003. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, 7(6), p.532–545.

Baesens, B. et al., 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *The Journal of the Operational Research Society*, 54(6), p.627–635.

Bae, S.M., Ha, S.H. & Park, S.C., 2005/1. A web-based system for analyzing the voices of call center customers in the service industry. *Expert systems with applications*, 28(1), p.29–41.

Balena, F. & Dimauro, G., 2005. *Practical guidelines and best practices for Microsoft Visual Basic and Visual C# developers*,

Bauckhage, C. et al., 2012. How players lose interest in playing a game: An empirical study based on distributions of total playing times. In *2012 IEEE Conference on Computational Intelligence and Games (CIG)*. p. 139–146.

Bauer, H.H., Grether, M. & Leach, M., 2002/2. Building customer relations over the Internet. *Industrial Marketing Management*, 31(2), p.155–163.

Bellazzi, R. & Zupan, B., 2008. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2), p.81–97.

Berson, A. & Smith, S.J., 2002. *Building Data Mining Applications for CRM*, New York, NY, USA: McGraw-Hill, Inc.

Borbora, Z. et al., 2011. Churn Prediction in MMORPGs Using Player Motivation Theories and an Ensemble Approach. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. p. 157–164.

Borbora, Z.H. & Srivastava, J., 2012. User Behavior Modelling Approach for Churn Prediction in Online Games. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. p. 51–60.

Bowcock, J., 2008. Apple - Press Info - iPhone App Store Downloads Top 10 Million in First Weekend. Available at: http://www.apple.com/pr/library/2008/07/14iPhone-App-Store-Downloads-Top-10-Million-in-First-Weekend.html [Acessado janeiro 8, 2017].

Boyles, J.L., Smith, A. & Madden, M., 2012. Privacy and data management on mobile devices. *Pew Internet & American Life Project*, 4. Available at:

http://www.pewinternet.org/2012/09/05/privacy-and-data-management-on-mobile-devices/.

Brasil, G., 2015. Pesquisa Game Brasil 2016. Available at: http://www.pesquisagamebrasil.com.br/pesquisa-2016.

Buckinx, W. & Van den Poel, D., 2005. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European journal of operational research*, 164(1), p.252–268.

Cailean, I., 2014. With Rising Cost of User Acquisition, App Marketers Must Be Smart. *Trademob*. Available at: http://www.trademob.com/rising-user-acquisition-costs-app-marketers-need-smarter-2/ [Acessado janeiro 8, 2017].

Cao, L. 2008. Domain Driven Data Mining (D3M). *2008 IEEE International Conference on Data Mining Workshops*, *Pisa.* pp. 74-76.

doi: 10.1109/ICDMW.2008.98

Carmichael, S., 2013. What it means to be a "whale" — and why social gamers are just gamers. *VentureBeat*. Available at: http://venturebeat.com/2013/03/14/whales-and-why-social-gamers-are-just-gamers/ [Acessado janeiro 11, 2017].

Carson, E.D., 2015. 64% of mobile game revenue is from 0.23% of players. *Investor's Business Daily*. Available at: http://www.investors.com/news/technology/click/mobile-games-64-percent-revenue-from-0-23-players/ [Acessado janeiro 11, 2017].

Cassab, H. & MacLachlan, D.L., 2006. Interaction fluency: a customer performance measure of multichannel service. *International Journal of Productivity and Performance Management*, 55(7), p.555–568.

Castanedo, F. et al., 2014. Using deep learning to predict customer churn in a mobile telecommunication network. Available at: http://wiseathena.com/pdf/wa_dl.pdf.

Castro, E.G. & Tsuzuki, M.S.G., 2015. Churn Prediction in Online Games Using Players' Login Records: A Frequency Analysis Approach. *IEEE Transactions on Computational Intelligence in AI and Games*, 7(3), p.255–265.

Chambers, C. et al., 2010. Characterizing Online Games. *IEEE/ACM Transactions on Networking*, 18(3), p.899–910.

Chang, S.E., Changchien, S.W. & Huang, R.-H., 2006/5. Assessing users' product-specific knowledge for personalization in electronic commerce. *Expert systems with applications*, 30(4), p.682–693.

Chen, M.-C., Chiu, A.-L. & Chang, H.-H., 2005. Mining changes in customer behavior in retail marketing. *Expert systems with applications*, 28(4), p.773–781.

Chen, Z.-Y., Fan, Z.-P. & Sun, M., 2012. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European journal of operational research*, 223(2), p.461–472.

Chiang, D.-A. et al., 2003. Goal-oriented sequential pattern for network banking churn analysis. *Expert systems with applications*, 25(3), p.293–302.

Chu, B.-H., Tsai, M.-S. & Ho, C.-S., 2007. Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems*, 20(8), p.703–718.

Cohen, J., 1988. Statistical power analysis for the behavioural sciences. Hillside. *NJ: Lawrence Earlbaum Associates*.

Coleman, D.E. & Montgomery, D.C., 1993. A Systematic Approach to Planning for a Designed Industrial Experiment. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences*, 35(1), p.1–12.

Collins, M., 2012. What's a Typical Subscription Commerce Retention Rate? *Matt Collins*. Available at: http://www.mattcollins.net/2012/07/whats-a-typical-subscription-e-commerce-retention-rate [Acessado janeiro 28, 2017].

Consulting, G., 2013. SaaS Churn Rate - What's Acceptable? *Customer Success-driven Growth*. Available at: http://sixteenventures.com/saas-churn-rate [Acessado janeiro 28, 2017].

Cooper HM. Synthetizing research: a guide for literature reviews. *3rd ed. Thousand Oaks: Sage*; 1998

Costello, S., 2016. Charting The Explosive Growth of the App Store. *Lifewire*. Available at: https://www.lifewire.com/how-many-apps-in-app-store-2000252 [Acessado janeiro 8, 2017].

Coussement, K. & De Bock, K.W., 2013/9. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of business research*, 66(9), p.1629–1636.

Coussement, K. & Van den Poel, D., 2008/1. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34(1), p.313–327.

Cox, L.A., 2002. Data Mining and Causal Modeling of Customer Behaviors. *Telecommunication Systems*, 21(2-4), p.349–381.

CyberAgent, 2016. Survey of Market Trends for Mobile Games in Four Major European Countries (U.K., Germany, France, and Italy). Available at: https://www.cyberagent.co.jp/en/newsinfo/press/detail/id=12026 [Acessado janeiro 8, 2017].

Datta, P. et al., 2000. Automated Cellular Modeling and Prediction on a Large Scale. *Artificial Intelligence Review*, 14(6), p.485–502.

David, B.A., 2015. Cost Per Install of Mobile Apps: The Current Trend. *appn2o*. Available at: https://www.appn2o.com/cost-per-install-of-mobile-apps-the-current-trend/ [Acessado janeiro 8, 2017].

Demiriz, A., 2004. Enhancing Product Recommender Systems on Sparse Binary Data. *Data mining and knowledge discovery*, 9(2), p.147–170.

Dhamanwar, M.A.R & Murab, S.A. 2016. A review paper on data driven data mining. *International Journal of Research In Science & Engineering*. e-ISSN: 2394-8299 Volume: 2 Special Issue: 1.

Douglas, S. et al., 2005. Mining customer care dialogs for "daily news". *IEEE transactions on audio, speech, and language processing*, 13(5), p.652–660.

Draganov, D. 2014. Freemium Mobile Games: Design & Monetization. Published by Dimitar Draganov. Jul 28, 2014.

Dreiseitl, S. & Ohno-Machado, L. 2002. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform*. 35 (5-6), pp. 352-359.

Gilles, C., 2002. The story behind successful CRM. Available at: http://www.bain.com/publications/articles/the-story-behind-successful-crm.aspx [Acessado janeiro 4, 2017h].

Edge, R., 2013. *Predicting Player Churn in Multiplayer Games using Goal-Weighted Empowerment*, Citeseer. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.401.9126&rep=rep1&type=pdf .

Fawcett, T. & Provost, F., 1997. Adaptive Fraud Detection. *Data mining and knowledge discovery*, 1(3), p.291–316.

Feng, W.-C., Brandt, D. & Saha, D., 2007. A Long-term Study of a Popular MMORPG. In *Proceedings of the 6th ACM SIGCOMM Workshop on Network and System Support for Games*. NetGames '07. New York, NY, USA: ACM, p. 19–24.

Ferreira, A.B. de H., 1994. Dicionário Aurélio Eletrônico. v. 1.4. *Rio de Janeiro: Nova Fronteira. [Links]*.

Fisher Ronald, A., 1935. The design of experiments. *London: Oliver and Boyd*.

Fox, K.F.A. & Kotler, P., 1994. Marketing estratégico para instituições educacionais. *São Paulo: Atlas*.

Fuller, D., 2016. Mobile Games Represent 85% Of All App Revenue In 2015 | Androidheadlines.com. *AndroidHeadlines.com* |. Available at: http://www.androidheadlines.com/2016/02/mobile-games-represent-85-of-all-app-revenue-in-2015.html [Acessado janeiro 8, 2017].

Grubb, J., 2016. Game of War's paying players spent an average of $550 on its in-app purchases in 2015. *VentureBeat*. Available at: http://venturebeat.com/2016/04/01/game-of-wars-paying-players-spent-an-average-of-550-on-its-in-app-purchases-in-2015/ [Acessado janeiro 11, 2017].

Hadiji, F. et al., 2014. Predicting player churn in the wild. In *2014 IEEE Conference on Computational Intelligence and Games*. p. 1–8.

Hall, M., Witten, I. & Frank, E., 2011. Data mining: Practical machine learning tools and techniques. *Kaufmann, Burlington*.

Ha, S.H., 2006. Digital content recommender on the Internet. *IEEE intelligent systems*, 21(2), p.70–77.

Ha, S.H., Bae, S.M. & Park, S.C., 2002. Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. *Computers & Industrial Engineering*, 43(4), p.801–820.

Hashmi, N., Butt, N. & Iqbal, M., 2013. Customer Churn Prediction in Telecommunication: A Decade Review and Classification. *International Journal of Computer Science*, 10(5), p.271–282.

Haykin, S. 1999. Neural Networks: A Comprehensive Foundation. *Prentice Hall.*

Hof, R., 2012. Google Research: No Mobile Site = Lost Customers. Available at: http://www.forbes.com/sites/roberthof/2012/09/25/google-research-no-mobile-site-lost-customers [Acessado janeiro 29, 2017].

Ho, T.K., 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), p.832–844.

Hsieh, N.-C., 2004. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert systems with applications*, 27(4), p.623–633.

Hung, S.-Y., Yen, D.C. & Wang, H.-Y., 2006. Applying data mining to telecom churn management. *Expert systems with applications*, 31(3), p.515–524.

Hwang, H., Jung, T. & Suh, E., 2004/2. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert systems with applications*, 26(2), p.181–188.

Hwong, C., 2016. A day in the life of the average mobile gamer. *Verto Analytics*. Available at: http://www.vertoanalytics.com/average-mobile-game-day/ [Acessado janeiro 19, 2017].

IFPI, 2015. IFPI - Digital Music Report 2015. Available at: http://www.ifpi.org/news/Global-digital-music-revenues-match-physical-format-sales-for-first-time [Acessado janeiro 7, 2017].

IFPI, 2016. IFPI - The recording industry worldwide. Available at: http://www.ifpi.org/global-statistics.php [Acessado janeiro 7, 2017].

Ingham, T., 2015. Global record industry income drops below $15bn for first time in decades - Music Business Worldwide. *Music Business Worldwide*. Available at: http://www.musicbusinessworldwide.com/global-record-industry-income-drops-below-15bn-for-first-time-in-history/ [Acessado janeiro 7, 2017].

Iniesta, R. Stahl, D. & McGuffin, P. 2016. Machine learning, statistical learning and the future of biological research in psychiatry. Psychological Medicine (2016), 46, 2455–2465. Cambridge University Press.

Jameson, 2016. Freemium Pricing: Can it Work for My App? Available at: https://www.dogtownmedia.com/freemium-pricing-can-it-work-for-my-app/ [Acessado janeiro 8, 2017].

jeannt, 2016. Analyzing Customer Churn using Machine Learning. Available at: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-azure-ml-customer-churn-scenario [Acessado janeiro 28, 2017].

Jenamani, M., Mohapatra, P.K.J. & Ghose, S., 2003. A stochastic model of e-customer behavior. *Electronic commerce research and applications*, 2(1), p.81–94.

Jensen, A. & la Cour-Harbo, A., 2001. *Ripples in Mathematics: The Discrete Wavelet Transform*, Springer Science & Business Media.

Jiang, T. & Tuzhilin, A., 2006. Segmenting Customers from Population to Individuals: Does 1-to-1 Keep Your Customers Forever? *IEEE transactions on knowledge and data engineering*, 18(10), p.1297–1311.

Jiao, J. (roger), Zhang, Y. & Helander, M., 2006/5. A Kansei mining system for affective design. *Expert systems with applications*, 30(4), p.658–673.

Jonker, J.-J., Piersma, N. & Van den Poel, D., 2004/8. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. *Expert systems with applications*, 27(2), p.159–168.

Kawale, J., Pal, A. & Srivastava, J., 2009. Churn Prediction in MMORPGs: A Social Influence Based Approach. In *2009 International Conference on Computational Science and Engineering*. p. 423–428.

Keerthi, S.S. et al., 2005. A Fast Dual Algorithm for Kernel Logistic Regression. *Machine learning*, 61(1-3), p.151–165.

Kennedy, K. et al., 2013. A window of opportunity: Assessing behavioural scoring. *Expert systems with applications*, 40(4), p.1372–1380.

Kim, J.K., Song, H. S., Kim, T. S., Kim, H.K., 2005. Detecting the change of customer behavior based on decision tree analysis. *Expert Systems*, 22(4), p.193–205.

Kimura, H., 2014. Cost Per Install (CPI) Of Mobile Apps Continues To Rise. Available at: https://sensortower.com/blog/cost-per-install-cpi-of-mobile-apps-continues-to-rise [Acessado janeiro 8, 2017].

Kim, W.C. & Mauborgne, R., 2005. Blue ocean strategy: How to Create Uncontested Market Space and Make the Competition Irrelevant Harvard Business School Press. *Boston, MA*.

Kim, Y., 2006. Toward a successful CRM: variable selection, sampling, and ensemble. *Decision support systems*, 41(2), p.542–553.

Kim, Y.-H. & Moon, B.-R., 2006. Multicampaign assignment problem. *IEEE transactions on knowledge and data engineering*, 18(3), p.405–414.

Kinnear, T.C., Taylor, J.R. & Kresge, S.S., 1991. *Marketing research: an applied approach*, McGraw-Hill New York, NY.

Kitchenham, B.A. et al., 2002. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, 28(8), p.721–734.

Koh, H.C. & Gerry, C.K.L., 2002. Data mining and customer relationship marketing in the banking industry. *Singapore Management Review*, 24(2), p.1.

Kuo, R.J., Liao, J.L. & Tu, C., 2005/8. Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce. *Decision support systems*, 40(2), p.355–374.

Landis, J.R. & Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), p.159–174.

Larivière, B. & Van den Poel, D., 2005. Investigating the post-complaint period by means of survival analysis. *Expert systems with applications*, 29(3), p.667–677.

Larivière, B. & Van den Poel, D., 2005/8. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert systems with applications*, 29(2), p.472–484.

Lee, C.-H., Kim, Y.-H. & Rhee, P.-K., 2001. Web personalization expert with combining collaborative filtering and association rule mining technique. *Expert systems with applications*, 21(3), p.131–137.

Lee, D., 2013. App Store "full of zombies" claim on Apple anniversary - BBC News. *BBC News*. Available at: http://www.bbc.com/news/technology-23240971 [Acessado janeiro 8, 2017].

Lejeune, M.A.P.M., 2001. Measuring the impact of data mining on churn management. *Internet Research*, 11(5), p.375–387.

Li CP1, Zhi XY, Ma J, Cui Z, Zhu ZL, Zhang C, Hu LP. 2012. Performance comparison between Logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. *Chin Med J*. 125(5):851-7.

Liao, S.-H. & Chen, Y.-J., 2004. Mining customer knowledge for electronic catalog marketing. *Expert systems with applications*, 27(4), p.521–532.

Lim, N. 2012. "Freemium Games Are Not Normal." June 26. Available at: http://www.gamasutra.com/blogs/NickLim/20120626/173051/Freemium_games_are_not_normal.php [Acessado janeiro 3, 2017].

Liu, D.-R. & Shih, Y.-Y., 2005/3. Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management*, 42(3), p.387–400.

Lovell, N., 2010. How to publish a game. *Gamesbrief. London*.

Lovell, N., 2011. Retention rate, churn and duration - Gamesbrief. *Gamesbrief*. Available at: http://www.gamesbrief.com/2011/11/retention-rate-churn-and-duration/ [Acessado janeiro 21, 2017].

MacKay, D.J.C., 1996. Bayesian Methods for Backpropagation Networks. In P. E. Domany, P. D. J. L. van Hemmen, & P. K. Schulten, orgs. *Models of Neural Networks III*. Physics of Neural Networks. Springer New York, p. 211–254.

Madden, G., Savage, S.J. & Coble-Neal, G., 1999/7. Subscriber churn in the Australian ISP market. *Information Economics and Policy*, 11(2), p.195–207.

Marbán, O., Mariscal, G., & Segovia, J. (2009). A Data Mining & Knowledge Discovery Process Model, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), InTech, DOI: 10.5772/6438.

Marques, A.I., García, V. & Sanchez, J.S., 2012. A literature review on the application of evolutionary computing to credit scoring. *The Journal of the Operational Research Society*, 64(9), p.1384–1399.

McClintock, P., 2016. Global 2015 Box Office: Revenue Hits Record $38 Billion-Plus. *The Hollywood Reporter*. Available at: http://www.hollywoodreporter.com/news/global-2015-box-office-revenue-851749 [Acessado janeiro 7, 2017].

McNab, H. & Wynn, A., 2000. *Principles and practice of consumer credit risk management*, CIB Publishing.

van der Meulen, R. & Rivera, J., 2014. Gartner Says Less Than 0.01 Percent of Consumer Mobile Apps Will Be Considered a Financial Success by Their Developers Through 2018. *http://www.gartner.com/*. Available at: http://www.gartner.com/newsroom/id/2648515 [Acessado janeiro 8, 2017].

Minotti, M., 2016. Video games will become a $99.6B industry this year as mobile overtakes consoles and PCs. *VentureBeat*. Available at: http://venturebeat.com/2016/04/21/video-games-will-become-a-99-6b-industry-this-year-as-mobile-overtakes-consoles-and-pcs/ [Acessado janeiro 7, 2017].

MobiPhoneSpec, 2017. MobiPhoneSpec - Cell Phone Screen Resolution, Sorted by Size. Available at: http://mobiphonespec.com/cellphone-screen-resolution-by-size.php [Acessado janeiro 8, 2017].

Mozer, M.C. et al., 2000. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 11(3), p.690–696.

Neslin, S. et al., 2004. Defection detection: improving predictive accuracy of customer churn models. *Tuck School of Business, Dartmouth College*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.489.5495&rep=rep1&type=pdf

Newzoo, 2016. Newzoo - Top Countries by Game Revenues. *Newzoo*. Available at: https://newzoo.com/insights/rankings/top-100-countries-by-game-revenues/ [Acessado janeiro 7, 2017].

Ngai, E.W.T., Xiu, L. & Chau, D.C.K., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2, Part 2), p.2592–2602.

Ng, K. & Liu, H., 2000. Customer Retention via Data Mining. *Artificial Intelligence Review*, 14(6), p.569–590.

Nie, G. et al., 2009. Finding the Hidden Pattern of Credit Card Holder's Churn: A Case of China. In G. Allen et al., orgs. *Computational Science – ICCS 2009*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 561–569.

Oreski, S., Oreski, D. & Oreski, G., 2012. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert systems with applications*, 39(16), p.12605–12617.

Payne, A. & Frow, P., 2006. Customer Relationship Management: from Strategy to Implementation. *Journal of Marketing Management*, 22(1-2), p.135–168.

Pearson, D., 2015. Mobile marketing costs continue to rise exponentially. *GamesIndustry.biz*. Available at: http://www.gamesindustry.biz/articles/2015-05-29-mobile-marketing-costs-continue-to-rise-exponentially [Acessado janeiro 8, 2017].

Peyton, J., 2014. What's the Average Bounce Rate for a Website? *gorocketfuel.com*. Available at: http://www.gorocketfuel.com/the-rocket-blog/whats-the-average-bounce-rate-in-google-analytics/ [Acessado janeiro 10, 2017].

Pfeifer, P.E. & Carraway, R.L., 2000. Modeling customer relationships as Markov chains. *Journal of interactive marketing*, 14(2), p.43.

Provost, F. & Fawcett, T., 2001. Robust classification for imprecise environments. Machine Learning, 42 (3) (2001), pp. 203–231.

Quinlan, J.R., 1993. *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Randolph, J. 2009. A Guide to Writing the Dissertation Literature Review. *Practical Assessment, Research & Evaluation, v14 n13 Jun 2009*

Rosset, S. et al., 2002. Customer Lifetime Value Modeling and Its Use for Customer Retention Planning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. New York, NY, USA: ACM, p. 332–340.

Runge, J. et al., 2014. Churn prediction for high-value players in casual social games. In *2014 IEEE Conference on Computational Intelligence and Games*. p. 1–8.

Ruta, D., Adl, C. & Nauck, D., 2009. New Churn Prediction Strategies in the Telecom Industry. In *Intelligent Data Analysis: Developing New Methodologies Through Pattern Discovery and Recovery*. IGI Global, p. 218–235.

Savetratanakaree, K. et al., 2014. Departure Prediction of Online Game Players. In *Advanced Materials Research*. Trans Tech Publications, p. 1370–1374.

Shen, Z., 2013. *Knowledge Discovery in High Throughput Screening: Towards Creation of a Data Mining Techniques Selection Guideline*. Available at: http://dspace.library.uu.nl/handle/1874/274957.

Sinclair, B., 2014. Only 2.2% of free-to-play users ever pay - Report. *GamesIndustry.biz*. Available at: http://www.gamesindustry.biz/articles/2014-04-09-only-2-2-percent-of-free-to-play-users-ever-pay-report [Acessado em janeiro 19, 2017].

Slater, S.F. & Narver, J.C., 2000. Intelligence generation and superior customer value. *Journal of the Academy of Marketing Science*, 28(1), p.120.

Smith, K.A., Willis, R.J. & Brooks, M., 2000. An analysis of customer retention and insurance claim patterns using data mining: a case study. *The Journal of the Operational Research Society*, 51(5), p.532–541.

Sonders, M., 2016. New mobile game statistics every game publisher should know in 2016. *SurveyMonkey Intelligence*. Available at: https://www.surveymonkey.com/business/intelligence/mobile-game-statistics/ [Acessado janeiro 19, 2017].

Stillwagon A., 2014. Did You Know: A 5% Increase in Retention Increases Profits by Up to 95%. Available at: https://smallbiztrends.com/2014/09/increase-in-customer-retention-increases-profits.html [Acessado em: Janeiro 11, 2016]

Takahashi, D., 2013. Facebook touts why its social network makes mobile games better. *VentureBeat*. Available at: http://venturebeat.com/2013/01/24/facebook-touts-why-its-social-network-makes-mobile-games-better/ [Acessado janeiro 19, 2017].

Takahashi, D., 2016. Mobile games hit $34.8B in 2015, taking 85% of all app revenues. *VentureBeat*. Available at: http://venturebeat.com/2016/02/10/mobile-games-hit-34-8b-in-2015-taking-85-of-all-app-revenues/ [Acessado janeiro 8, 2017].

TapJoy, 2016. Top Personas of Free-to-Play Mobile Games. Disponível em: http://www.visualistan.com/2016/04/the-top-personas-of-free-to-play-mobile-gamers-and-how-to-treat-them.html [Acessado em 10/01/2017]

Trujillo, J., 2006. A report on the first international workshop on best practices of UML: (BP-UML'05). *ACM SIGMOD Record*, 35(3), p.48–50.

Valadares, J., 2011. Free-to-play Revenue Overtakes Premium Revenue in the App Store. *Flurry),[Online] July*, 7. Available at: http://flurrymobile.tumblr.com/post/113367742230/free-to-play-revenue-overtakes-premium-revenue-in.

Valadares, S., 2014. 80% dos Apps da Apple Store nunca foram baixados. Pequenas Empresas e Grandes Negócios. Disponível em: http://revistapegn.globo.com/Colunistas/Silvia-Valadares/noticia/2014/07/80-dos-apps-da-apple-store-nunca-foram-baixados.html [Acessado em 08/01/2017].

Van Den Poel, D., 2003. *Predicting Mail-Order Repeat Buying: Which Variables Matter?*, Ghent University, Faculty of Economics and Business Administration. Available at: http://ideas.repec.org/p/rug/rugwps/03-191.html [Acessado janeiro 3, 2017].

Van den Poel, D. & Buckinx, W., 2005. Predicting online-purchasing behaviour. *European journal of operational research*, 166(2), p.557–575.

Van Dreunen, J., 2011. A business history of video games: Revenue models from 1980 to today. *The Game Behind The Video Game: Business, Regulation, and Society in the Gaming Industry. New Brunswick, New Jersey, USA*.

Verhoef, P.C. et al., 2003/3. The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision support systems*, 34(4), p.471–481.

Verhoef, P.C. & Donkers, B., 2001. Predicting customer potential value an application in the insurance industry. *Decision support systems*, 32(2), p.189–199.

Victor, S. Y. Lo, 2002. The true lift model: a novel approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter, 4(2), p. 78-86.*

Walz, A., 2015. The Data Behind Customer Acquisition and Retention for F2P Mobile Games. *apptentive.com*. Available at: https://www.apptentive.com/blog/2015/04/09/the-data-behind-customer-acquisition-and-retention-for-f2p-mobile-games/ [Acessado janeiro 11, 2017].

Wang, G. et al., 2010. Predicting Credit Card Holder Churn in Banks of China Using Data Mining and MCDM. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. p. 215–218.

Wang, Y.-F., Chuang, Y.-L., Hsu, M.-H., Keh, H.-C., 2004. A personalized recommender system for the cosmetic business. *Expert systems with applications*, 26(3), p.427–434.

Wei, C.-P. & Chiu, I.-T., 2002. Turning telecommunications call details to churn prediction: a data mining approach. *Expert systems with applications*, 23(2), p.103–112.

West, D., 2000/9. Neural network credit scoring models. *Computers & operations research*, 27(11–12), p.1131–1152.

Wirth, R. 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. P. 29–39.

Xu, Y. et al., 2002. Adopting customer relationship management technology. *Industrial Management & Data Systems*, 102(8), p.442–452.

Zhao, H., Sinha, A.P. & Ge, W., 2009/3. Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert systems with applications*, 36(2, Part 2), p.2633–2644.

Zhao, Y. et al., 2005. Customer Churn Prediction Using Improved One-Class Support Vector Machine. In X. Li, S. Wang, & Z. Y. Dong, orgs. Advanced Data Mining and Applications. Lecture Notes in Computer Science. International Conference on Advanced Data Mining and Applications. Springer Berlin Heidelberg, p. 300–306.