PGCOMP - Programa de Pós-Graduação em Ciência da Computação
Universidade Federal da Bahia (UFBA)
Av. Adhemar de Barros, s/n - Ondina
Salvador, BA, Brasil, 40170-110

http://pgcomp.dcc.ufba.br
pgcomp@ufba.br

A problemática da evasão de jogadores em jogos grátis-para-jogar representa um desafio significativo na indústria de jogos eletrônicos. A crescente popularidade desses modelos de negócios, nos quais os jogadores podem acessar o jogo gratuitamente, coloca uma ênfase crucial na retenção desses usuários para garantir o sucesso financeiro e a sustentabilidade do jogo. Nesse cenário, a análise preditiva emerge como uma ferramenta essencial para antecipar e compreender os padrões de evasão.

O trabalho começou com um mapeamento sistemático de literatura no campo de modelos preditivos em game analytics, visando responder à principal questão de pesquisa. Como modelos preditivos estão sendo aplicados em game analytics?. A pesquisa foi conduzida com base em um protocolo que definiu os objetivos, questões de pesquisa e critérios de inclusão e exclusão.

Os principais resultados indicam que a pesquisa sobre modelos preditivos em game analytics tem crescido significativamente desde 2010, com uma variedade de técnicas de aprendizado de máquina sendo aplicadas. Além disso os objetos de predição mais investigados incluem a probabilidade de vitória, a predição de evasão e a perícia do jogador. Quanto às técnicas de pré-processamento, foram identificadas várias abordagens, como análise de componentes principais (PCA) e técnicas de raspagem da web (web scraping).

Focamos nossa pesquisa em predição de evasão, inicialmente pela definição de evasão e estabelecimento de datas de corte, com a consideração de múltiplas janelas de tempo para classificação dos jogadores como evadidos ou recorrentes. A análise abordou as ameaças à validade do trabalho, incluindo questões de definição de evasão, desequilíbrio de classes e o uso de técnicas como o SMOTE para balancear os dados.

Foram avaliados seis modelos de aprendizado de máquina, com ênfase em métricas como acurácia, precisão, recall e AUC (Area Under the Curve). A técnica de 10-fold cross validation foi aplicada para validar os modelos, proporcionando uma visão mais abrangente de seu desempenho. A análise da importância das features revelou quais características dos jogadores eram mais relevantes para a previsão da evasão, embora a interpretação dessas features tenha sido destacada como dependente do contexto do jogo.

Em última análise, o trabalho ofereceu insights promissores para a previsão de evasão de jogadores em jogos grátis-para-jogar, mas ressaltou a necessidade de abordagens cuidadosas e considerações contextuais para mitigar ameaças à validade e garantir a generalização dos modelos para diferentes conjuntos de dados e períodos no tempo.

Palavras-chave: Modelos Preditivos, Game Analytics, Aprendizado de Máquina, Jogos Eletrônicos, Predição de Evasão.

MSC | 176 | 2023

Predição da Evasão de jogadores em Jogo Grátis-Para-Jogar Utilizando Game Analytics.

Iury Maia De Almeida

Predição da Evasão de Jogadores em Jogo Grátis-Para-Jogar Utilizando Game Analytics.

Iury Maia de Almeida

Dissertação de Mestrado

Universidade Federal da Bahia
Programa de Pós-Graduação em
Ciência da Computação

Dezembro | 2023

UFBA

UNIVERSIDADE FEDERAL DA BAHIA

DISSERTAÇÃO DE MESTRADO

PREDIÇÃO DA EVASÃO DE JOGADORES EM JOGO GRÁTIS-PARA-JOGAR UTILIZANDO GAME ANALYTICS

IURY MAIA DE ALMEIDA

Programa de Pós-Graduação em Ciência da Computação

Salvador
20 de dezembro de 2023

PGCOMP-Msc-2023

IURY MAIA DE ALMEIDA

PREDIÇÃO DA EVASÃO DE JOGADORES EM JOGO GRÁTIS-PARA-JOGAR UTILIZANDO GAME ANALYTICS

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: RODRIGO ROCHA GOMES E SOUZA

Salvador
20 de dezembro de 2023

Ficha catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias Prof. Omar Catunda, SIBI – UFBA.

A447 Almeida, Iury Maia.

Predição da evasão de jogadores em jogo grátis-para-jogar utilizando game analytics./ Iury Maia de Almeida. – Salvador, 2023.

69 f.

Orientador: Prof. Dr. Rodrigo Rocha Gomes e Souza.

Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Computação, 2023.

1. Jogos Eletrônicos. 2. Jogadores. 3. Game Analytic. I. Souza, Rodrigo Rocha Gomes e. II. Universidade Federal da Bahia. III. Título.

CDU 004.42

“*Predição da Evasão de Jogadores em Jogo Grátis-Para-Jogar Utilizando Game Analytics*”

Iury Maia de Almeida

Dissertação apresentada ao Colegiado do Programa de Pós-Graduação em Ciência da Computação na Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Ciência da Computação.

Banca Examinadora

Rodrigo Rocha

Prof. Dr. Rodrigo Rocha Gomes e Souza (Orientador - PGCOMP)

Documento assinado digitalmente
gov.br TIAGO OLIVEIRA MOTTA
Data: 26/12/2023 12:05:52-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Tiago Oliveira Motta (UFRB

Documento assinado digitalmente
gov.br LYNN ROSALINA GAMA ALVES
Data: 22/12/2023 10:20:10-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Lynn Rosalina Gama Alves (UFBA)

AGRADECIMENTOS

É com grande emoção que dedico este momento para expressar meus sinceros agradecimentos a todas as pessoas que contribuíram para a realização deste sonho.

Primeiramente, gostaria de expressar minha profunda gratidão à minha família. Seu amor, apoio e incentivo foram a força motriz por trás de todas as minhas conquistas. Obrigado por estarem sempre ao meu lado, por acreditarem em mim e por me darem as bases necessárias para alcançar este marco em minha vida.

Aos meus amigos, que compartilharam comigo não apenas os momentos de estudo, mas também os de descontração e apoio mútuo, obrigado por serem uma fonte constante de motivação e alegria. Suas palavras de encorajamento e seus sorrisos foram essenciais para superar os desafios ao longo deste percurso.

Não posso deixar de expressar minha mais profunda gratidão ao Professor Rodrigo Rocha. Sua orientação, expertise e dedicação foram fundamentais para o sucesso deste projeto. Suas palavras sábias, críticas construtivas e incentivo constante foram verdadeiramente inspiradoras. Agradeço por compartilhar seu conhecimento e por acreditar em meu potencial, mesmo nos momentos mais desafiadores.

Sou profundamente grato por todo apoio, amor e inspiração que recebi ao longo do caminho. Que este seja apenas o início de muitas conquistas futuras, onde possamos continuar a crescer e aprender juntos.

Obrigado a todos.

Iury Maia de Almeida.

RESUMO

A problemática da evasão de jogadores em jogos grátis-para-jogar representa um desafio significativo na indústria de jogos eletrônicos. A crescente popularidade desses modelos de negócios, nos quais os jogadores podem acessar o jogo gratuitamente, coloca uma ênfase crucial na retenção desses usuários para garantir o sucesso financeiro e a sustentabilidade do jogo. Nesse cenário, a análise preditiva emerge como uma ferramenta essencial para antecipar e compreender os padrões de evasão.

O trabalho começou com um mapeamento sistemático de literatura no campo de modelos preditivos em *game analytics*, visando responder à principal questão de pesquisa, Como modelos preditivos estão sendo aplicados em *game analytics*?. A pesquisa foi conduzida com base em um protocolo que definiu os objetivos, questões de pesquisa e critérios de inclusão e exclusão.

Os principais resultados indicam que a pesquisa sobre modelos preditivos em *game analytics* tem crescido significativamente desde 2010, com uma variedade de técnicas de aprendizado de máquina sendo aplicadas. Além disso os objetos de predição mais investigados incluem a probabilidade de vitória, a predição de evasão e a perícia do jogador. Quanto às técnicas de pré-processamento, foram identificadas várias abordagens, como análise de componentes principais (PCA) e técnicas de raspagem da web (*web scraping*).

Focamos nossa pesquisa em predição de evasão, inicialmente pela definição de evasão e estabelecimento de datas de corte, com a consideração de múltiplas janelas de tempo para classificação dos jogadores como evadidos ou recorrentes. A análise abordou as ameaças à validade do trabalho, incluindo questões de definição de evasão, desequilíbrio de classes e o uso de técnicas como o SMOTE para balancear os dados.

Foram avaliados seis modelos de aprendizado de máquina, com ênfase em métricas como acurácia, precisão, *recall* e AUC (*Area Under the Curve*). A técnica de *10-fold cross validation* foi aplicada para validar os modelos, proporcionando uma visão mais abrangente de seu desempenho. A análise da importância das *features* revelou quais características dos jogadores eram mais relevantes para a previsão da evasão, embora a interpretação dessas *features* tenha sido destacada como dependente do contexto do jogo.

Em última análise, o trabalho ofereceu *insights* promissores para a previsão de evasão de jogadores em jogos grátis-para-jogar, mas ressaltou a necessidade de abordagens cuidadosas e considerações contextuais para mitigar ameaças à validade e garantir a generalização dos modelos para diferentes conjuntos de dados e períodos no tempo.

Palavras-chave: Modelos Preditivos, Game Analytics, Aprendizado de Máquina, Jogos Eletrônicos, Predição de Evasão

ABSTRACT

The issue of player churn in free-to-play games poses a significant challenge in the electronic gaming industry. The growing popularity of these business models, where players can access the game for free, places a crucial emphasis on retaining these users to ensure the financial success and sustainability of the game. In this scenario, predictive analysis emerges as an essential tool to anticipate and understand the patterns of player churn.

This study began with a systematic literature review in the field of predictive models in game analytics, aiming to answer the main research question: How are predictive models applied in game analytics? The research was conducted based on a protocol that defined the objectives, research questions, and inclusion and exclusion criteria.

The main findings indicate that research on predictive models in game analytics has grown significantly since 2010, with a variety of machine learning techniques being applied. Furthermore, the most investigated prediction targets include the probability of winning, churn prediction, and player expertise. Regarding preprocessing techniques, several approaches were identified, such as Principal Component Analysis (PCA) and web scraping techniques.

We focused our research on churn prediction, initially by defining churn and establishing cutoff dates, considering multiple time windows for classifying players as churned or recurrent. The analysis addressed the validity threats of the work, including churn definition issues, class imbalance, and the use of techniques like SMOTE to balance the data.

Six machine learning models were evaluated, with an emphasis on metrics like accuracy, precision, recall, and AUC (Area Under the Curve). The 10-fold cross-validation technique was applied to validate the models, providing a more comprehensive view of their performance. The analysis of feature importance revealed which player characteristics were most relevant for churn prediction, although the interpretation of these features was highlighted as context-dependent.

Ultimately, the work offered promising insights into the prediction of player churn in free-to-play games but emphasized the need for careful approaches and contextual considerations to mitigate validity threats and ensure the generalization of models to different datasets and time periods.

Keywords: Predictive Models, Game Analytics, Machine Learning, Electronic Games, Churn Prediction

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

INTRODUÇÃO

A indústria de jogos é uma das principais do entretenimento moderno; estima-se que em 2022 alcançou o número de 2,9 bilhões de jogadores mundialmente, possuindo uma projeção para crescimento de até 3,5 bilhões até 2025. O crescimento do número de jogadores é impulsionado por melhor infraestrutura da rede móvel e acesso a *smartphones*, tendo nos continentes Ásia, África e América Latina os principais mercados em crescimento. O mercado de jogos para dispositivos móveis gerou 50% de todo o lucro da indústria em 2022, correspondendo a 92,2 bilhões de dólares. A grande expansão desse mercado acontece principalmente graças à acessibilidade do modelo grátis-para-jogar[1].

O modelo grátis-para-jogar é aquele onde o jogador tem a oportunidade de começar a experiência de forma gratuita, e dentro desse jogo grátis existe a possibilidade de monetização da experiência em sua maioria através de micropagamentos, mas podendo abranger outras formas de arrecadação, como a publicidade (OLIVEIRA, 2022). Nesse modelo as empresas têm como objetivo minimizar as barreiras de entradas para o jogo com o objetivo de atrair a maior quantidade possível de jogadores (CASTRO; TSUZUKI, 2015). Logo, o modelo grátis-para-jogar é descrito como possuindo duas grandes vantagens. Primeiro, a possibilidade dos micropagamentos dentro do jogo serem flexíveis, atraindo diferentes tipos de jogadores que estão dispostos a gastar quantidades distintas de dinheiro (PAAVILAINEN et al., 2013). Além disso, a possibilidade da criação de uma grande comunidade de jogadores. Mesmo que muitos deles não comprem nada, a troca de informações e experiências implica em um aumento de visibilidade e atrai mais usuários (SHEN; WILLIAMS, 2011). Contudo este modelo gera um custo para o desenvolvedor, demandando uma grande parte dos esforços de desenvolvimento para que o jogador gratuito permaneça e ofereça algum tipo de retorno financeiro (OLIVEIRA, 2022). (SALEN; ZIMMERMAN, 2004) A dominância do modelo grátis-para-jogar desencadeia altas taxas de evasão, pois uma grande quantidade dos jogadores desinstalam os jogos ainda nos primeiros dias após a interação (DRACHEN et al., 2016), Assim para reter a maior quantidade de jogadores é essencial predizer quando um determinado jogador deixará o jogo. Logo, a habilidade de

[1] https://newzoo.com/products/reports/global-games-market-report

monitorar, analisar e predizer o comportamento dos jogadores é crucial para a criação de um negócio sustentável e para incentivar os jogadores a permanecer no jogo (DRACHEN et al., 2016; HADIJI et al., 2014). Além do mais, reter um consumidor em seu jogo custa cinco vezes menos do que investir para conseguir um novo (CASTRO; TSUZUKI, 2015). Nesse cenário são desenvolvidos modelos preditivos, que são sustentados por técnicas e ferramentas de *game analytics*, que é o processo de descobrir e comunicar padrões em dados (*analytics*), descrito e aplicado no contexto de desenvolvimento de *games* e pesquisa em *games* (DRACHEN; EL-NASR; CANOSSA, 2013). Comumente os modelos utilizam técnicas de aprendizado de máquina para catalogar e predizer quais jogadores deixarão o jogo e quando o farão.

Os modelos de predição de evasão são geralmente obtidos através de aprendizado supervisionado, que consiste em fornecer ao algoritmo um conjunto de dados de treinamento para os quais o rótulo da classe associada é conhecido. O resultado dos algoritmos pode ser um valor contínuo (regressão) ou predizer um rótulo de classe definido (classificação) (EL-NASR; DRACHEN; CANOSSA, 2016). Para a predição de evasão os modelos geralmente utilizam classificação, dividindo seus usuários entre dois rótulos, evadidos e recorrentes. O primeiro trabalho a formalmente definir evasão ancorando sua definição em game design foi (HADIJI et al., 2014).

Observando os trabalhos relacionados e os estudos desenvolvidos em predição de evasão, nosso estudo tem como objetivo principal desenvolver e avaliar um modelo preditivo genérico para evasão de usuários em jogos grátis-para-jogar. Para atingir o objetivo principal, o dividimos em objetivos secundários sendo estes o desenvolver um mapeamento sistemático sobre modelos preditivos em jogos para identificar tendências e os objetos preditivos utilizados na área de jogos. Identificar base de dados disponíveis para executar a pesquisa e coletar dados. Eleger métricas e atributos para selecionar as *features* do modelo a ser desenvolvido. Aplicar e avaliar o conjunto de *features* selecionadas em diversos algoritmos de aprendizagem de máquina. Por fim analisar os melhores modelos gerados.

Nosso estudo desenvolveu uma nova definição de evasão que toma como base nas já existentes na literatura mas com um foco em detectar evasão de novos jogadores, além de melhor identificar o tempo no qual os mesmos evadirão. Realizamos um estudo onde catalogamos as *features* de vários modelos preditivos de evasão e identificamos 9 *features* genéricas comuns entre estes estudos que eram possíveis de serem derivadas. Com o conjunto de *features* definidos utilizamos 6 algoritmos distintos de aprendizagem de máquina e 4 variações de tempo aplicando ou não a técnica SMOTE para gerar 42 modelos de predição de evasão. Avaliamos os modelos e escolhemos os 4 melhores respectivos a cada janela de tempo para validação utilizando *10-fold cross validation*. Por fim sobre os 4 modelos finais analisamos o desempenhos das *features* com base em sua importância.

O restante deste trabalho está organizado da seguinte forma: o Capítulo 2 estabelece a fundamentação teórica, discutimos no Capítulo a definição de jogo, desenvolvimento de jogos, *game analytics* e predição de evasão. No Capítulo 3 é apresentado o mapeamento sistemático realizado onde explicamos o método de pesquisa, os resultados e ameaças a validade do estudo. No Capítulo 4 são apresentados os trabalhos relacionados quem englobam os modelos preditivos que este trabalho tem como referência. O Capítulo 5

discutimos a base de dados encontrada, o processo de limpeza e transformação de dados, a seleção das *features*, analisaremos os modelos gerados e a importância das *features*, por fim apresentamos a ameaças a validade dos resultados. O capitulo 6 é a conclusão e a sugestão para trabalhos futuros.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

O Capítulo 2 apresenta conceitos importantes para fundamentar a pesquisa, este Capítulo é estruturado em 5 seções. A Seção 2.1 é referente aos fundamentos que caracterizam um jogo do ponto de vista conceitual e social. A Seção 2.2 representa quais são os processos e etapas do desenvolvimento de jogos e como diferentes modelos de monetização impactam no processo de desenvolvimento. A Seção 2.3 é referente a introdução de conceitos sobre *game analytics*, explicando técnicas de analytics, coleta e transformações de dados são aplicadas a área de jogos. Por fim na Seção 2.4 modelos preditivos de evasão aplicados a jogos como eles são desenvolvidos e seu impacto em desenvolvimento de jogos.

2.1 JOGO

Jogos têm se tornado um tema mais recorrente na literatura, contudo eles se fazem presentes em nossa sociedade em um contexto cultural muito antes de se tornarem-se objetos de estudo. "De modo geral os jogos são considerados fenômenos transculturais que acompanham a humanidade desde os primórdios, sempre relacionado a algum tipo de aprendizado ou ganho cognitivo" (ANDRADE, 2013). Porém existem diferentes visões do o que é um jogo e seus limites, assim surgindo divergências sobre a definição do mesmo. Para Salen e Zimmerman (2004), "jogo é um software no qual um ou mais jogadores fazem decisões através de objetos controlados no jogo e seus recursos, com o intuito de atingir um objetivo". Essa é uma visão muito moderna do que é jogo provocada pela atual crescimento dos jogos eletrônicos. Como Hullett et al. (2012) comentam, "jogos estão progressivamente se tornando mais comuns na indústria de desenvolvimento de software". Além disso, El-Nasr, Drachen e Canossa (2013) tem uma ideia semelhante, definindo jogos como aplicações de software focadas na experiência de usuário; jogos não são pensados para serem mais eficientes e focados e sim desafiadores. Contudo tais definições excluem uma grande gama de jogos, como esportes, jogos de tabuleiro, brincadeiras de crianças, dentre outros não eletrônicos ou que não remetem apenas a software como jogos pervasivos ou ARGs (jogos de realidade alternativa do inglês - *alternate reality games*).

Compartilhamos da visão de Velloso (2017), onde entendemos que o objeto de estudo *video game* ou jogo eletrônico pertence a um conjunto maior de objetos denominados

jogos. Huizinga (1971) se destaca apresentando jogo como fenômeno cultural e não biológico, estudando-o em uma perspectiva histórica, apresentando alguns elementos inerentes a todos os jogos, como o conceito de início e fim, o espaço em sua definição de círculo mágico, onde comenta que "todo jogo se processa e existe no interior de um campo previamente delimitado, de maneira material ou imaginária, deliberada ou espontânea", além de apresentar o conceito de tensão, que se reflete na imprevisibilidade, regras que trazem ordem ao jogo e, por fim, ter uma ação voluntária. Caillois (1961) apresenta uma extensão do trabalho que Huizinga (1971) começou e fez um contraponto limitando a extensão de alguns conceitos considerados muito amplos, assim definindo 6 princípios inerentes que definem jogo:

- Livre: os jogadores não são obrigados a jogar; caso fossem, o jogo perderia seu atrativo e felicidade como qualidades divertidas;
- Separado: possui limites de espaço físico e temporal, definidos e fixados previamente;
- Imprevisível: o curso do jogo não pode ser determinado, nem estabelecer resultados antecipadamente, e com a possibilidade de inovações deixada à iniciativa do jogador;
- Improdutivo: não criar bens materiais ou riqueza, nenhum elemento de qualquer tipo; exceto pela troca de propriedades entre os jogadores, terminando em uma situação idêntica à predominante do início do jogo;
- Governado por regras: através de convenção suspender as leis comuns, e pelo momento estabelecer uma nova legislação, que sozinha é válida;
- Faz-de-conta: acompanhado pela percepção especial de uma segunda realidade ou uma irrealidade livre, contra a vida real. (CAILLOIS, 1961), P. 10. (Tradução nossa).

Jogos eletrônicos são um dos tipos de jogos e no desenvolvimento de jogos modernos existem diversas formas de comercialização; uma das mais comuns é a comprar-para-jogar (do inglês: *buy-to-play*) ou também conhecido como *premium*. Em contrapartida surgiu o modelo *freemium* (ou grátis-para-jogar, como é conhecido na indústria de jogos), referindo-se ao produto onde o componente principal é gratuito mas a receita, quando existe, é gerada através de produtos adicionais e serviços *premium* dentro do jogo (HAMARI; HANNER; KOIVISTO, 2017; ALHA et al., 2014). Além disso, dentro do modelo grátis-para-jogar existem ramificações onde os jogos oferecem outras formas de monetizar o jogo, seja ela através de propagandas, doações aos desenvolvedores ou apenas oferecer a experiência genuinamente gratuita sem nenhum tipo de monetização.

2.2 DESENVOLVIMENTO DE JOGOS

O desenvolvimento de jogos é bastante complexo, uma vez que a natureza multidisciplinar do processo de combinar som, arte, sistema de controles, inteligência artificial (AI) e fatores humanos também fazem o desenvolvimento de jogos diferente do desenvolvimento

de software tradicional (ALEEM; CAPRETZ; AHMED, 2016). Apesar de desenvolvedores de jogos poderem ajustar técnicas e processos tanto de engenharia de software quanto de produções artísticas, não existe consenso em como combinar métodos dessas duas áreas (ENGSTRÖM et al., 2018).

A engenharia de software do desenvolvimento de jogos ou GDSE (do inglês, *Game Development Software Engineering*) possui todas suas fases do ciclo de vida combinadas em três principais categorias: pré-produção, produção, e pós-produção (ALEEM; CAPRETZ; AHMED, 2016). A pré-produção é responsável por testar a viabilidade das mecânicas, incluindo possíveis estratégias de marketing e monetização; a produção é responsável pela documentação e implementação do jogo; e, por fim, a pós-produção tem foco nos testes e no marketing (ALEEM; CAPRETZ; AHMED, 2016).

A fase de pós-produção é marcada por dois principais pontos, sendo eles os testes beta que consistem em avaliar o jogo como um todo utilizando testadores externos à equipe de desenvolvimento. Geralmente nessas sessões é permitido aos testadores jogarem do início ao fim ou em alguns casos apenas alguns cenários (ALEEM; CAPRETZ; AHMED, 2016). Aliado aos testes existe o controle de qualidade, que é um processo de validação para assegurar a qualidade do jogo (ALEEM; CAPRETZ; AHMED, 2016); geralmente, mas não exclusivamente, são avaliados cinco princípios: funcional: indica se todas as funcionalidades estão implementadas corretamente; completude interna: indica se as mecânicas, ramificações e condições foram corretamente endereçadas; balanceamento: indica se o jogo não está nem muito difícil nem muito fácil: diversão, indica se o jogo é envolvente; e, por fim, acessibilidade: indica se o jogo é fácil de se entender (RAMADAN; WIDYANI, 2013).

Após o lançamento do jogo temos a última fase de desenvolvimento que tem se tornando muito importante com os design de jogos modernos o pós lançamento ou manutenção do jogo publicado. Geralmente está fase é composta pôr a utilização de ciclos de atualização denominado *patches*. Um *patch* pode corrigir pequenos *bugs*, introduzir novas funcionalidades ou serem grandes os suficientes para caracterizar uma versão inteiramente diferente do jogo comumente denominado DLC (*downloadable content*) ou Expansões (FARRIER et al., 2012). A depender do modelo de monetização escolhido para a produção do jogo a fase de pós lançamento vai ser a mais longa e custosa do desenvolvimento, como exemplo no modelo grátis-para-jogar onde o foco do desenvolvimento muda pois o objetivo não é mais vender o melhor jogo possível, mas, vender conteúdos no jogo (HAMARI et al., 2017). Dessa forma os desenvolvedores tendem a criar um ciclo de atualizações mais constante, uma vez que estas atualizações são oportunidades de monetização e ferramentas para os jogadores interagirem com o jogo e amigos assim aumentando engajamento (OLIVEIRA, 2022).

2.3 GAME ANALYTICS

Game analytics não é uma área totalmente independente; ela possui raízes em diversas outras, como métodos de inspeção de usabilidade, *business intelligence*, estatística, mineração de dados e muitas outras. Logo, *game analytics* é um domínio específico de *analytics*, descrito e aplicado no contexto de desenvolvimento de *games* e pesquisa em

games (DRACHEN; EL-NASR; CANOSSA, 2013). Nessa área torna-se recorrente a menção de 3 importantes conceitos, sendo eles telemetria, métricas e *analytics* no contexto de desenvolvimento de jogos.

Telemetria em jogos é a habilidade de coletar os dados em um cliente e transmiti-los para um servidor. Os dados de telemetria permitem aos *designers* investigar o comportamento dos jogadores de uma grande população e, algumas vezes, de todos os jogadores (GAGNÉ; EL-NASR; SHAW, 2011). Utilizar telemetria ajuda a ter uma melhor representatividade da população de jogadores, já que não é analisada apenas uma amostra da população (GAGNÉ; EL-NASR; SHAW, 2011).

Métricas no contexto de desenvolvimento de jogos são dados transformados em atributos que descrevem diversas propriedades dos jogadores, como "tempo total jogado" ou "quantidade diária de usuários" (DRACHEN; EL-NASR; CANOSSA, 2013). Logo, as métricas são dados numéricos obtidos da interação entre o jogador e o software, porém refletindo um comportamento específico do jogador (TYCHSEN, 2008). Além disso, à medida que jogos se tornam mais populares e complexos, o teste tradicional não é mais capaz de cobrir todos os possíveis estados do jogo. Isso faz com que coleções de métricas sejam uma das melhores ferramentas para entender os jogadores (HULLETT et al., 2012).

Analytics é o processo de descobrir e comunicar padrões em dados, para resolver problemas em negócios ou desenvolver previsões para apoiar decisões de gerenciamento empresariais, direcionar ações ou melhorar desempenho. Logo, *game analytics* é *analytics* aplicado ao desenvolvimento de jogos (DRACHEN; EL-NASR; CANOSSA, 2013). *Analytics* é normalmente dependente de modelagem computacional. Existem vários ramos ou domínios de *analytics*, por exemplo, análise de marketing, análise de risco, análise de rede e *game analytics*. É importante mencionar que *analytics* não e a mesma coisa que análise de dados, *analytics* é um termo global no qual abrange toda a metodologia de encontrar e comunicar padrões em dados, enquanto análise de dados é referente a aplicação pontual em um grupo de dados (DRACHEN; EL-NASR; CANOSSA, 2013).

2.4 PREDIÇÃO DA EVASÃO DE JOGADORES

Na literatura existem diversos trabalhos que tentam definir e propõem soluções para predizer a evasão de jogadores, como em (LEE et al., 2016; PERIÁÑEZ et al., 2016; DRACHEN et al., 2016; WEBER et al., 2011; TAMASSIA et al., 2016; BORBORA; SRIVASTAVA, 2012; MILOŠEVIĆ; ŽIVIĆ; ANDJELKOVIĆ, 2017; RUNGE et al., 2014; HADIJI et al., 2014; CASTRO; TSUZUKI, 2015), além de trabalhos relacionados em e-commerces, telecomunicações, dentre outras áreas. Porém para esta pesquisa a definição de predição de evasão é a proposta por Hadiji et al. (2014), pois esse foi o primeiro trabalho a definir formalmente o conceito de predição de evasão em jogos grátis-para-jogar alinhados aos padrões da indústria.

Hadiji et al. (2014) tratam a predição de evasão como um problema de classificação onde cada jogador pode ser considerado evadido (positivo) ou recorrente (negativo). Esses estados são atribuídos através de um determinado ponto no tempo; (HADIJI et al., 2014). Referem-se a este marco como *cutoff date* (em português, dia de corte). O primeiro caso, tipo P1, é mais direto da definição onde é considerado todo e qualquer jogador sem

nenhuma sessão após o dia de corte como um evadido. A Figura 1 apresenta à esquerda um exemplo do problema P1 onde o eixo X representa os dias, o eixo Y exemplos de jogadores e os pontos pretos sessões de um determinado jogador (HADIJI et al., 2014). Argumentam que esta definição é muito rígida e não é reflete aplicações do mundo real uma vez que não leva em conta o possível retorno dos jogadores.

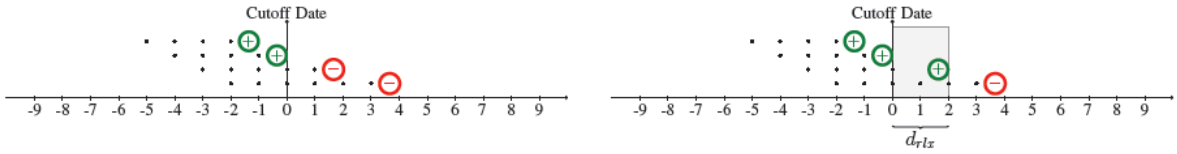

Figura 1: à esquerda o problema tipo P1 onde o dia de corte é definido sem dias de relaxamento e à direita a definição do problemas tipo P2 com 2 dias de relaxamento Fonte: (HADIJI et al., 2014)

O problema de tipo P2 tem o intuito de refletir melhor a situação em um cenário real, onde os jogadores que tiveram baixo número de sessões ou dias após a data de corte já serão considerados como evadidos. O restante dos dias jogados pelo precisão cair dentro da janela de tempo definida pela variável d - rls (dias de relaxamento). Hadiji et al. (2014) argumenta que esses dias podem refletir a diferença entre jogadores engajados e os que não estão mais. Além do mais, do ponto de vista da indústria, os jogadores que estão dentro dos dias de relaxamento iriam provavelmente sair; agir cedo nesses jogadores pode os incentivar a continuar no jogo.

Lee et al. (2016) definiu a evasão de jogadores em dois casos de acordo com o período de tempo no qual os jogadores possuem atividade. O modelo 1 que consiste em selecionar uma data de corte então identificar todos os usuários ativos dentro do período de duas semanas antes a data selecionada e categorizá-los caso possuam algum tipo de interação na semana seguinte a data de corte. O modelo 2 segue de forma semelhante ao modelo 1 a única diferente são os períodos de tempo estudos sendo estes de 4 semanas para a coleta dos usuários ativos e de duas semanas para a categorização de evasão. Todos os usuários que não estavam ativos dentro do período de seleção dos usuários são removidos pois já são considerados evadidos.

Uma importante definição de evasão foi proposta por Drachen et al. (2016), com o intuito de capturar o comportamento dos jogadores assim que começam a interagir com o jogo e descobrir de forma rápida se estes iram continuar jogando ou não a data de corte que antes era fixa nas definições anteriores agora é com base na data do primeiro acesso de cada jogador como demonstra a figura 2. Sendo assim ele experimentou vários períodos diferentes de predição, o primeiro com base no final da primeira sessão do jogador, o segundo com base no primeiro dia de jogo e o terceiro com base na primeira semana de jogo. Respectivamente para cada um dos períodos de predição ele aplicou técnicas distintas, para a primeira sessão foi aplicado regras de decisão com base em heurísticas, para o primeiro dia de jogo foi aplicado diferentes classificadores previamente utilizados para predição de evasão e para o final da primeira semana foi uma estratégia combinando

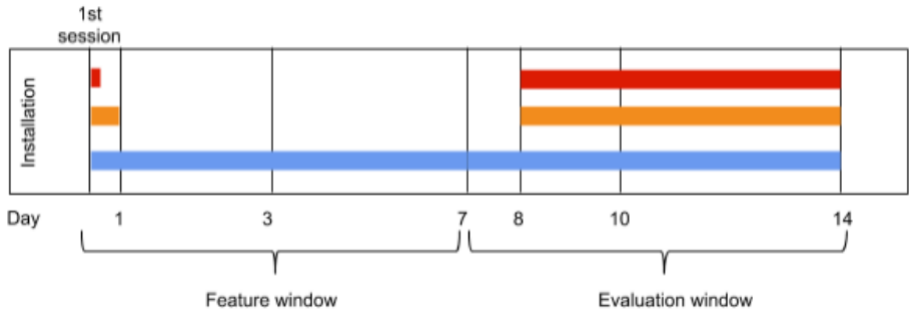

Figura 2: Fonte: (DRACHEN et al., 2016)

os resultados de 3 classificados.

Capítulo 3

MODELOS PREDITIVOS EM GAME ANALYTICS - UM MAPEAMENTO SISTEMÁTICO

Este capítulo é dedicado a descrever os procedimentos, métodos e resultados relacionados ao mapeamento sistemático de literatura sobre modelos preditivos em *game analytics*, respondendo à principal questão de pesquisa e suas subquestões.

Um mapeamento sistemático ajuda a evidenciar um domínio de conhecimento mesmo com altos níveis de granularidade. Ele identifica grupos de evidências e áreas não exploradas dentro de um tópico de pesquisa, ajudando futuras pesquisas a identificar áreas onde é necessário um conjunto maior de estudos primários (KITCHENHAM STUART CHARTERS, 2007).

O capítulo está estruturado da seguinte forma: A Seção 2.3 apresenta uma introdução a *game analytics* e a Seção 3.2, aos modelos preditivos. A Seção 3.3 apresenta o método de pesquisa; a Seção 3.4 apresenta os resultados das questões de pesquisa e a Seção 3.5, a análise dos dados coletados. Na Seção 3.6 são apresentadas as ameaças à validade do estudo.

3.1 MÉTODO DE PESQUISA

Essa Seção apresenta o planejamento do mapeamento sistemático. O estudo foi desenvolvido seguindo como orientação o processo sistemático proposto por Petersen et al. (2008), que inclui as diretrizes a seguir. O fluxograma do método de pesquisa utilizado está demonstrado na imagem 3.

- Definição do protocolo: O protocolo é um artefato que define os principais objetivos da pesquisa; a intenção é padronizar e tornar o trabalho como um todo replicável futuramente;
- Definição das questões de pesquisa: Durante a extração de dados esse documento é utilizado para agregar e auxiliar na leitura e identificação dos dados importantes para responder as perguntas definidas no protocolo
- Formulário de coleta de dados: Documento desenvolvido para guiar a coleta de dados durante a leitura dos estudos primários.

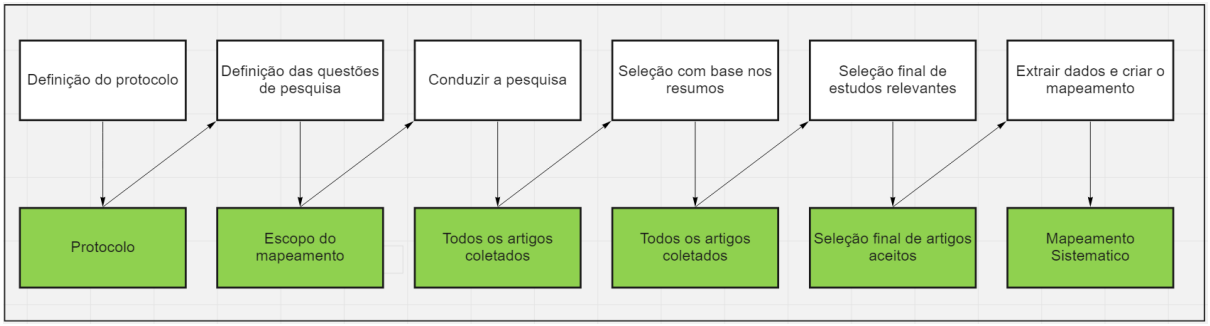

Figura 3: Processo do método de pesquisa utilizado

3.1.1 Definição de protocolo

Para nortear a condução do mapeamento, foi elaborado um protocolo de pesquisa. Nesse estudo, o foco foi identificar modelos preditivos na área de *game analytics*, especificamente os objetos de predição, as seleções de recursos e jogos utilizados. Foram incluídos estudos publicados até 2018. O ano de início não foi definido para ter uma ideia melhor de quando os estudos dessa área começaram a ser publicados. O protocolo utilizado foi estruturado utilizando o PICOC proposta por)BARBARA2007STUART. O PICOC é definido por cinco categorias, população é onde a evidencia é coletada, intervenção onde é definido o escopo do tópico a ser estudado, comparação que é aplicada a intervenção para comparar diferentes estudos, saídas o resultado esperado do estudo e o contexto que é onde os estudos estão inseridos.

- **População**: Estudos na área de *game analytics*.
- **Intervenção**: Estudos em *game analytics* que desenvolvem algum tipo de modelo preditivo.
- **Comparação**: Não aplicável uma vez que não estamos fazendo comparação entre estudos.
- **Saídas**: conjunto de estudos sobre modelos preditivos em *game analytics*
- **Contexto**: modelos preditivos (dados utilizados, técnicas, ferramentas, métricas, processos).

3.1.2 Definições das questões de pesquisa

Para definir os escopo do mapeamento sistemático foi definida uma pergunta principal e um conjunto de questões secundárias. A pergunta principal foi: **"Como modelos preditivos estão sendo aplicados em *game analytics*?"** Por modelos preditivos consideramos técnicas de aprendizagem de máquina que após treinadas são capazes de identificar e ou replicar determinados padrões.

As sub-questões(SQ) a seguir foram desenvolvidas para auxiliar na resposta da questão principal.

SQ1: "Quais são os principais eventos e ou simpósios onde são publicados modelos preditivos dentro de *game analytics*?" O objetivo dessa pergunta é monitorar os principais eventos de publicação dessa área, para identificar trabalhos semelhantes.

SQ2: "Quais são as técnicas de aprendizagem de máquinas aplicadas aos modelos preditivos em *game analytics*? por exemplo algoritmos (SVM, redes neurais, naive bayes...) e técnica (classificação supervisionada, regressão linear)" Essa pergunta visa identificar quais são as técnicas mais usadas para o desenvolvimento de modelos preditivos em *games*.

SQ3: "Quais são os jogos mencionados e quais trabalhos possuem algum tipo de vínculo com iniciativas privadas?" essa pergunta tem como objetivo identificar qual era o nível de interação da pesquisa com empresas de jogos privadas. Para melhor identificar essa pergunta dividimos os jogos trabalhados em categorias: jogos sérios, demonstrações, jogos convencionais e jogos *mobile*.

SQ4: "Quais são os objetos de predição de modelos preditivos mais investigadas em *game analytics*? o objeto da predição por exemplo retenção, comportamento do jogador, latência de rede... analisar ao longo do tempo" O objetivo dessa pergunta é identificar qual é o objeto de predição mais comum em jogos e os menos comuns.

SQ5: "Quais técnicas de pré-processamento são utilizadas? ex.: seleção de *features*, normalização de *features*." Essa pergunta visa identificar técnicas usadas no pré-processamento de dados ou seleção de *features*.

3.1.3 Estratégia de busca dos estudos primários

Nessa Seção apresentaremos a implementação do mapeamento sistemático. Isso inclui as estratégias de busca, critérios de inclusão e exclusão, além do processo de seleção.

Para realizar as buscas, selecionamos um grupo de palavras-chave com o intuito de capturar uma grande porção dos artigos da área. As palavras selecionadas foram, *data mining*, *game analytics*, *game data analysis*, *game telemetry*, *game metrics*, *machine learning*, *prediction*"

Uma vez com o conjunto das palavras-chave selecionadas, para realizar as buscas nas bases de dados de forma mais satisfatória foi feita uma combinação que melhor refletisse o grupo de estudo ao atual. Para chegar na combinação adequada foram feitas diversas combinações e com base no número de resultados, observando se o grupo de artigos realmente era pertencente ao objeto de pesquisa. Sendo definido da seguinte forma:

(game OR gaming OR player) AND ("predicting"OR "classification"OR "regression") AND ("machine learning"OR "data mining"OR "neural network")

Utilizamos a *string* realizando as devidas alterações para cada uma das bases de dados selecionadas. Cada base tinha uma forma possivelmente diferente de implementação dos conectores AND ou OR, o que gerava algum tipo de adaptação. Para a realização da pesquisa utilizamos as bases de dados correspondentes: ACM, IEEE, Springer, Science Direct, Wiley Online Library, dblp.

Tabela 1: Critérios de inclusão e exclusão

Critério	Inclusão(I)/Exclusão(E)
Estudos que relacionam jogos a modelos preditivos	I
Ultima versão publicada do artigo	I
Artigos publicados até 2018	I
Artigos em inglês	I
Modelos preditivo não relacionados a jogos	E
Artigos duplicados	E
Artigos secundários e terciários	E
Literatura cinza	E
Modelos preditivos baseados em surveys	E

3.1.4 Critérios de seleção dos estudos

Após a seleção das bases de dados e a extração dos artigos, foi necessário estabelecer critérios de inclusão e exclusão dos estudos. Os critérios levam em conta diversos aspectos dos estudos e servem como um filtro para selecionar o ano de publicação, linguagem, tipo e tema dos artigos. Pretendemos focar apenas nos estudos que estão dentro de *game analytics*, que trabalharam com algum tipo de modelo preditivo e que estão associados diretamente com dados de telemetria coletados de um jogo. A Tabela 1 apresenta os critérios definidos.

3.1.5 Procedimento de seleção dos estudos

Utilizando as strings descritas na Tabela 2 identificamos inicialmente 1.065 dos quais 48 eram artigos duplicados resultando em 1017. Após a coleta dos trabalhos procedemos para a leitura dos resumos e títulos de todos os artigos selecionados aplicando as regras de inclusão e exclusão descritas na Tabela 1 assim restaram 167 trabalhos. Durante a etapa final todos os 167 artigos foram lidos em sua completude e aplicando as regas da Tabela 1 finalizamos com 79 trabalhos, durante a leitura dos artigos também foi executada a extração dos dados relevantes para responder as questões de pesquisa.

3.1.6 Estratégia de extração e classificação dos dados

O esquema de classificação dos estudos e dos dados dos estudos primários foi subdivido em 5 tipos de classificação, com o intuito de responder cada uma das questões de pesquisa descritas na Seção 3.1.2

- **Classificação dos locais de publicação**. Com o intuito de responder à questão **SQ1**, classificamos onde cada artigo final foi publicado. Os artigos selecionados foram classificados dentro dos periódicos, *workshops*, simpósios e demais meios de publicação relevantes.

Tabela 2: Descrição das strings de busca

Repositório	String	Artigos
ACM	(('game' 'gaming' 'player') AND ('prediction' 'predictive' 'predict' 'predicting' 'classification' 'regression') AND ('machine learning' 'data mining' 'neural network')) OR recordAbstract:(('game' 'gaming' 'player') AND ('prediction' 'predictive' 'predict' 'predicting' 'classification' 'regression') AND ('machine learning' 'data mining' 'neural network')) OR keywords.author.keyword:(('game' 'gaming' 'player') AND ('prediction' 'predictive' 'predict' 'predicting' 'classification' 'regression') AND ('machine learning' 'data mining' 'neural network'))	561
IEEE	(game OR gaming OR player) AND ("predicting"OR "classification"OR "regression") AND ("machine learning"OR "data mining"OR "neural network")	308
SPRINGER	'(Game AND OR AND Gaming AND OR AND player) AND (prediction AND OR AND predictive AND OR AND predict AND OR AND predicting AND OR AND classification AND OR AND regression) AND ("machine learning"AND OR AND "data mining"AND OR AND "neural network") '	48
Science Direct	(game OR gaming OR player) AND ("predicting"OR "classification"OR "regression") AND ("machine learning"OR "data mining"OR "neural network")	98
Wiley Online Library	(game OR gaming OR player) AND (prediction* OR predictive* OR predict* OR predicting* OR classification OR regression) AND ("machine learning"OR "data mining"OR "neural network")	20
DBLP	(Game | Player | Gaming) (prediction | predictive | predict | predicting | classification | regression) ("neural networks")	30

- **Classificação das técnicas usadas**. Durante a leitura, cada artigo foi classificado de acordo com as técnicas utilizadas, entre naïve Bayes, SVM, árvores de decisão, redes neurais, dentre outras. Essa classificação nos ajuda a responder à pergunta **SQ2**.
- **Classificação dos tipos de jogos**. A classificação foi aplicada à medida que a leitura dos artigos foi realizada. As categorias têm como intuito ajudar a responder à pergunta **SQ3** e para isso categorizamos nas seguintes classes: jogos convencionais, prova de conceito e jogo sério.
- **Classificação dos objetos de predição**. Para responder à **SQ4**, classificamos os objetos de predição de cada artigo. Para a realização dessa classificação as categorias foram criadas em conjunto com a leitura dos trabalhos; quando um novo trabalho que possuía um objeto preditivo que não se encaixava em um anterior, uma nova categoria era adicionada.
- **Classificação das técnicas de pré-processamento**. Para responder à **SQ5**, classificamos as técnicas de pré-processamento utilizadas para trabalhar os dados utilizados no modelos de cada artigo. Contudo não encontramos artigos descrevendo os métodos de pré-processamento para conseguirmos classificá-los.

3.2 RESULTADOS

Na última etapa do processo, os dados foram extraídos e interpretados para responder cada umas das subquestões de pesquisa SQ1 a SQ5.

SQ1: "Quais são os principais eventos e ou simpósios onde são publicados modelos preditivos dentro de *game analytics*?"

79 artigos foram analisados em sua completude e foi gerado um histograma das datas de publicação na imagem 4, é possível observar que a publicação mais antiga encontrada é de 1997 e é observável que a partir do ano de 2010 as publicações relacionadas a modelos preditivos em jogos aumentaram drasticamente.

Dividimos a classificações dos meios de publicação em três categorias conferências, *Journals* e *workshops*, destas categorias 62 artigos foram publicados em conferências, 12 artigos foram em *journals* e 5 foram em *workshops*. No gráfico da imagem 5 é evidenciado as 4 conferências com maior número de publicações, para melhor clareza do gráfico geramos com apenas conferências que possuíam 2 ou mais artigos publicados. A IcoICT (*International Conference on Information and Communication Technology*) e a FDG (*International Conference on the Foundations of Digital Games*) com 2 artigos publicados. A FedCSIS(*conference on computer science and intelligence systems*) com 4 artigos publicados e por fim a CIG*Conference on Computational Intelligence and Games*) a conferência com 10 publicações possuindo o maior número de publicações em nosso estudo.

Foram identificados 7 *journals* distintos, tendo *Transaction on computacional inteligence and AI in Games* o *jornal* com maior número de publicações totalizando 4. Dentre as 5 publicações em *workshops* foram identificados 3 *workshops* distintos tendo o SIGCOMM(*Workshop on Network and system support for games*) o *workshop* com mais

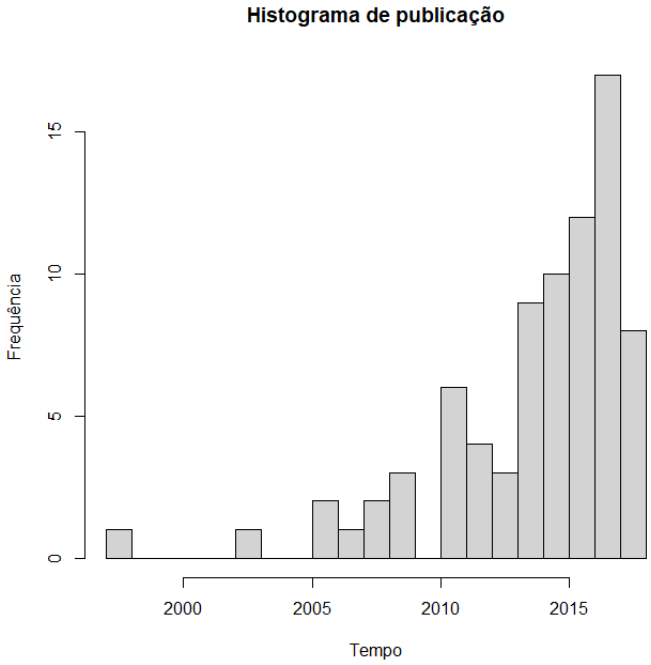

Figura 4: Histograma das publicações

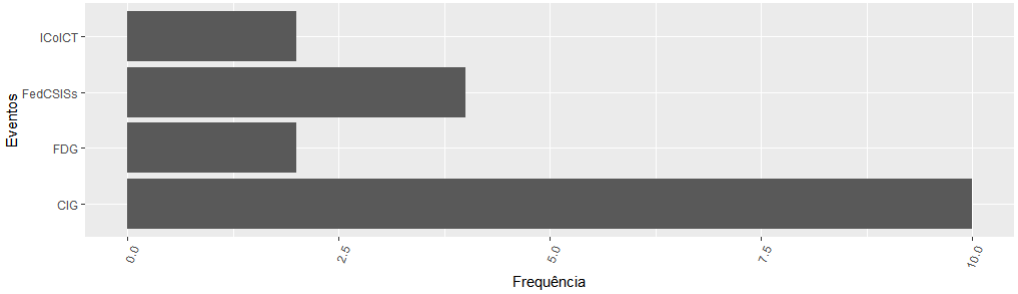

Figura 5: Número de publicação por conferência

publicações sendo 3 no total.

SQ2: "Quais são as técnicas de aprendizagem de máquinas aplicadas aos modelos preditivos em *game analytics*?

Foram identificados 67 técnicas diferentes utilizadas para gerar os modelos preditivos em nossa pesquisa. Sendo destas técnicas redes neurais (*Neural Network*) foi a mais recorrente aparecendo em 33 estudos (41.7%) seguida de árvore de decisão com 21 (21.5%) e SVM também com 17 (21.5%) ocorrências as demais técnicas e suas frequências podem ser observadas na imagem 6. No gráfico da imagem 6 selecionamos apenas as técnicas com 5 ou mais ocorrências para simplificar a visualização.

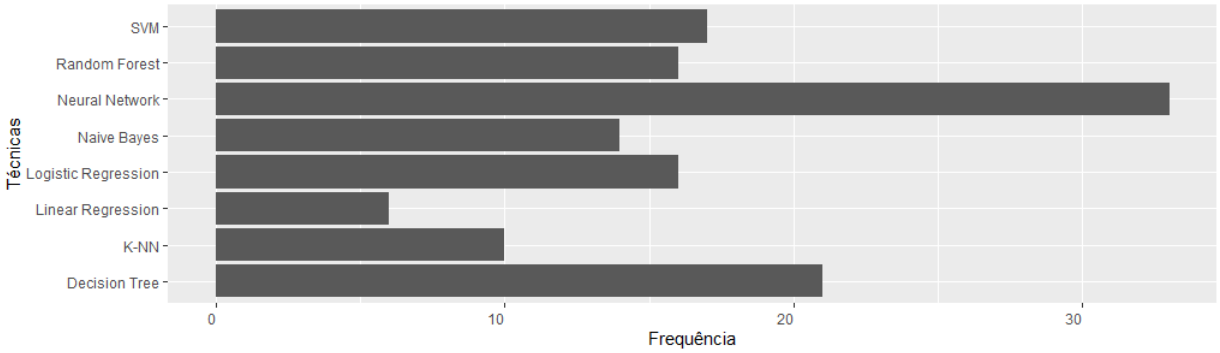

Figura 6: Recorrência de diferentes Técnicas de aprendizagem de máquina.

SQ3: "Quais são os jogos mencionados e quais trabalhos possuem algum tipo de vínculo com iniciativas privadas?"

Foram identificados 58 jogos distintos mencionados nos estudos o gráfico da imagem 7 reflete os mais recorrentes, escolhemos apresentar no gráfico apenas os jogos que se repetiam duas ou mais vezes sendo *World of Warcraft* e Dota os jogos com maior frequência 6 respectivamente.

Para responder a segunda parte da pergunta e estabelecer se algum artigo teve algum vínculo com alguma empresa durante a produção dos estudos, tentamos identificar se os artigos mencionavam as empresas aos quais tinha recebido algum tipo de auxílio. Contudo esse método se provou improdutivo uma vez que pouquíssimos artigos tinham algum tipo de vínculo diretamente mencionados. Partindo desse ponto de vista, classificados os jogos encontrados em 3 categorias: jogo convencional, jogo sério e prova de conceito. Foram identificados 68 trabalhos que estudaram Jogos convencionais, 6 provas de conceito e 5 relacionados a jogos sérios.

SQ4: "Quais são as abordagens de modelos preditivos mais investigadas em *game analytics*?

Durante o processo da pesquisa catalogamos os objetos preditivos de todos os estudos, contudo devido as especificidades e definições de cada estudo tornaram a categorização desses objetos muito granular. Para melhor o agrupamento os estudos passamos a analisar não apenas o objeto preditivo mas qual o objetivo da predição, desta forma dentre os 79 trabalhos encontramos 33 objetivos de predição diferentes.

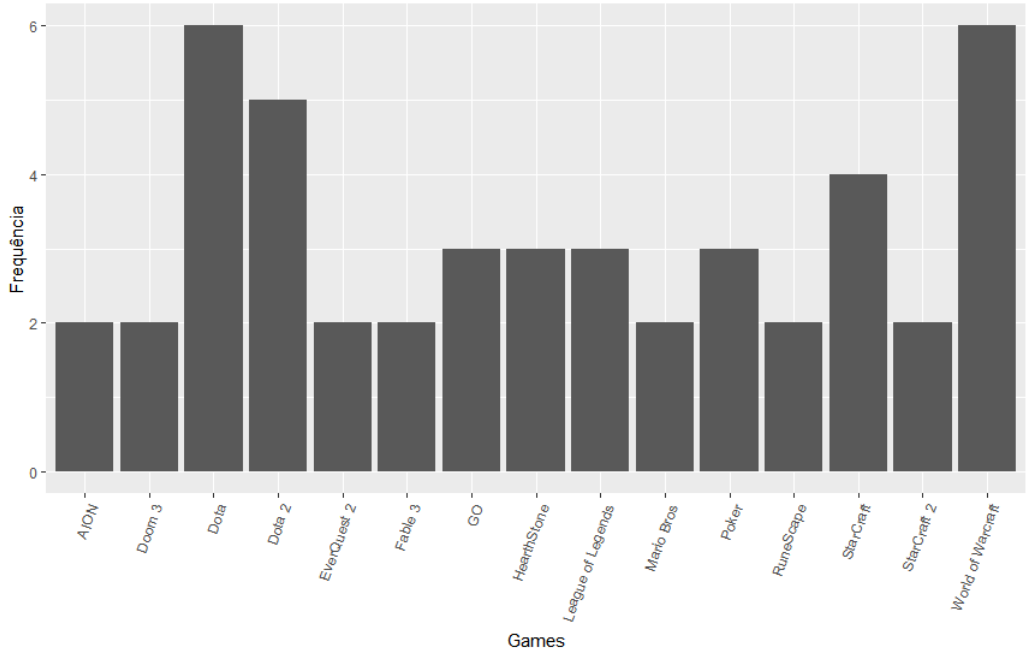

Figura 7: Recorrência de diferentes Jogos identificados.

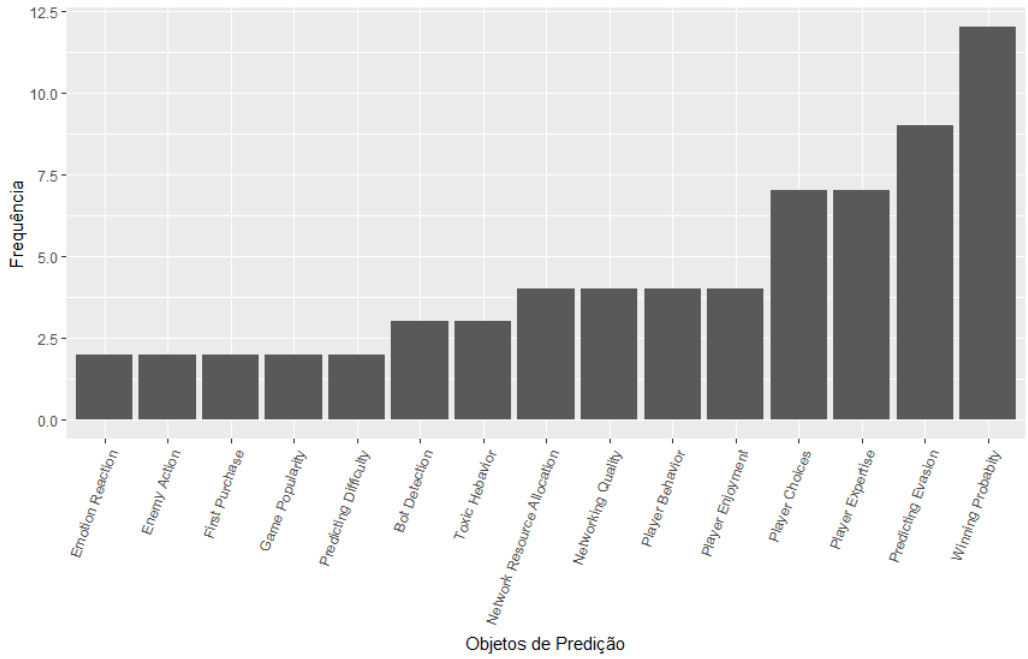

Figura 8: Frequência dos Objetos de predição.

A imagem 8 demonstra os objetos preditivos com duas ou mais ocorrências. Sendo probabilidade de vitória o mais frequente com 12 trabalhos encontrados, predição de evasão com 9 trabalhos e perícia do jogador com 7. É Válido analisar que alguns trabalhos não se limitaram a apenas um tipo de objeto tendo múltiplos tipos de predição em suas pesquisas.

SQ5: "Quais técnicas de pré-processamento são utilizadas? ex.: seleção de *features*, normalização de *features*."

Identificar a resposta para essa pergunta foi muito complexa uma vez que boa parte dos artigos não detalha as técnicas utilizadas no pré-processamento dos dados ou definição de *features*. Contudo identificamos 11 trabalhos que dentre estes detalharam 7 técnicas de pré-processamento distintas. No estudo realizado por Koulieris et al. (2016) com foco em *gaze predition* foi utilizada uma técnica com abordagem probabilística usando interferência Bayesiana para correlacionar com maior precisão os objetos em cena com o olhar.

Sobre tratamento de imagem Li e Chen (2006) utilizaram uma técnica nomeada de esquema de interesse, este é baseado em um sistema que correlaciona as interações de objetos em um cenário com as intenções dos jogadores, definindo que suas ações futuras são influenciadas pelos objetos próximos. Para remover bordas de imagens do data set, Erdem e Halici (2016) aplicaram uma técnica para removê-las utilizando uma borda de uma única cor a qual é removida automaticamente por uma combinação da transformação de Hough e o detector de bordas Canny. Detectores de unidades de ação foram utilizados por Vinkemeier, Valstar e Gratch (2018) com o intuito de auxiliar na detecção de expressões faciais, uma vez com as unidades de ação categorizadas a informação de cada frame era sumarizada e utilizada para predizer se o jogador irar dar *fold* ou não.

Análise de componente principal (PCA) é uma técnica usada para reduzir a dimensões de um conjunto de *features* do data set utilizada por Lim e Harrell (2013). A técnica consiste em decompor os dados em vários componentes, cada um definido como uma combinação linear do conjunto original de *features* usando coeficientes. Raspagem de rede (*web scraping*) é a ferramenta mais utilizada na pesquisa com 5 pesquisas utilizando algumas ferramentas para conseguir dados de páginas web. A *synthetic minority over-sampling technique* (SMOTE) foi a última técnica de encontrada na pesquisa, ela é utilizada em conjuntos de dados com classes não balanceadas com o intuito de criar sinteticamente novos *data points* a fim de balancear o data set sendo utilizada no trabalho (LEE et al., 2016).

3.3 AMEAÇAS À VALIDADE

Nesse estudo foi Identificado as seguintes ameaças à validade:

Palavras Chaves: O conjunto de palavras chaves usados na elaboração das strings de pesquisa pode não representar a completude do domínio que tentamos estudar. É possível que a adição, remoção ou modificação da estrutura da strings gere melhores resultados de busca retornando uma visão mais precisa do estado da arte. Contudo por ser uma área expansão, definições, palavras, técnicas e outras características são muito voláteis tornando a mitigação desse passo muito difícil.

Base de Dados: Apesar de termos feito a pesquisa em uma boa quantidade de base de dados, deixamos várias outras fora do estudo como a Springer link e o Google Acadêmico. Devido ao acesso dificultado de muitos artigos da Springer link acabamos por não utilizar a base de dados. O Google Acadêmico por ser um indexador, retornou uma quantidade de resultados muito alta, além de possuir uma grande quantidade de duplicadas impossibilitado a realização do estudo.

O Pesquisador: O estudo foi realizado em sua grande parte por apenas um pesquisador para a leitura e discussão dos artigos selecionados tornando a possibilidade de viés na seleção dos estudos muito alta. Sabendo disto foi tomado um cuidado extra na seleção e leitura para seguir as regras de inclusão e exclusão para minimizar ao máximo esse viés.

Capítulo

4

TRABALHOS RELACIONADOS

Nesse Capítulo iramos introduzir trabalhos relacionas quem desenvolveram modelos preditivos sobre evasão de jogadores, apresentaremos o contexto, objetivos dos trabalhos, resultados e compararemos as semelhanças e diferenças como o nosso trabalho.

O trabalho desenvolvido por Hadiji et al. (2014) é um dos mais importantes da área. Eles foram os primeiros a formalizar a definição de predição de evasão para jogos grátis-para-jogar, possuindo duas formalizações do problema com dois métodos diferentes de extração de dados. Além disso, eles introduziram um conjunto de *features* que permite predição de evasão para um grande conjunto de jogos grátis-para-jogar:

- Número de sessões: quantas vezes o jogador entrou no jogo desde sua instalação.
- Número de dias: o número de dias jogados desde o primeiro dia de instalação.
- Tempo de ausência atual: o tempo entre o início da sessão atual até o final da última sessão jogada.
- Tempo médio por sessão: o valor médio do tempo de todas as sessões jogadas.
- Tempo médio entre sessões: é o valor médio do tempo de ausência entre todas as sessões jogadas.
- Tempo de jogado: quanto tempo foi jogado atualmente.
- Valor de evasão: construindo um modelo de evasão baseado na média de evasão de jogadores (com uma função de ajuste de curva), é possível obter um valor de evasão no dia respectivo.
- Bandeira de usuário *premium*: é uma variável que se torna verdadeira caso o jogador tenha realizado alguma compra.
- Categoria de gasto predefinida: são definidas três categorias de gastos para classificar os jogadores, sendo os que mais gastam na categoria um, os que estão mais próximos da média na categoria dois e o que gastam menos na categoria três.

- Número de compras: é o número total de compras realizadas por um determinado jogador.
- Valor médio gasto por sessão: o valor médio gasto por sessão por um determinado jogador.

Por fim, Hadiji et al. (2014) treinaram e avaliaram o modelo com cinco jogos diferentes, contudo os jogos não foram divulgados. Divulgando Apenas uma métrica de comparação obteve valores médios de F-Score de 0.916 com árvore de decisão, 0,889 com regressão logística, 0,874 com rede neural e 0.804 com naive bayers.

Nosso trabalho propõem um conjunto diferente de *features* e para identificá-las catalogamos as *features* de diferentes trabalhos e selecionamos as mais comuns chegando a um conjunto amplo de *features* que podem ser mais facilmente replicadas. Além disso utilizamos uma definição de evasão diferente, focada na predição de evasão dos usuários em até 28 dias após o primeiro contato com o jogo. Outro fator determinante como o conjunto de algoritmos de aprendizagem de máquina, Aplicação da técnica SMOTE para balanceamento da classe minoritária e diferentes variações temporais para evasão de usuários são características diferentes entre os dois trabalhos.

Runge et al. (2014) apresentaram a predição de evasão para jogadores valiosos ou popularmente conhecidos como baleias. Tais jogadores detêm esse nome por serem o menor grupo de usuários que gera a maior receita. A evasão foi definida com um problema binário onde eram classificados como evadidos usuários com um período de inatividade de 14 dias após a primeira atividade no jogo. Os jogos utilizados foram *Diamond Dash* e *Monster World* dois jogos grátis-para-jogar com foco no público casual de dispositivos móveis.

Os modelos desenvolvidos utilizaram redes neurais, regressão logística, árvore de decisão e SVM, comparando a melhor performance dos respectivos modelos foi identificado que as redes neurais tiveram o melhor desempenho com uma AUC de 0.815 para o jogo *Diamond Dash* e 0.930 para o *Monster World*. Além disso foi realizado testes A/B para identificar o real impacto dos modelos no jogo *Monster World*, enviando quantias significantes das moedas do jogo para os jogadores categorizados como evadidos. Contudo não foi detectado uma mudança significativa na evasão dos jogadores valiosos, indicando que estes jogadores próximos da evasão não são engajados por esses tipos de incentivos (RUNGE et al., 2014).

Lee et al. (2016) desenvolveu um modelo preditivo para o jogo *Crazy Dragon*, um RPG de ação *mobile* grátis-para-jogar desenvolvido e publicado pela Mgame. Para a classificação foi definida como um problema binária entre evadidos e recorrentes, definindo o período de treinamento de duas semanas e quarta semana onde após estes períodos jogadores sem atividade eram classificados como evadidos. Como um jogo grátis-para-jogar a proporção de evasão e muito maior acaba criando um problema de classe minoritária e para resolver esse problema foi utilizado a técnica SMOTE. O SMOTE gera dados sintéticos para balancear o conjunto de dados (LEE et al., 2016). Utilizaram árvore de decisão, floresta aleatória e SVM (support vector machine), todos tiveram sua performance avaliada utilizando *10-fold cross validation* com 10.000 amostra de jogadores. Todos os modelos conseguiram pontuações expressiva com valores acima de 0.75 entre

todas as métricas, porém o melhor modelo foi a floresta aleatória com a aplicação da técnica SMOTE para o período de 2 semanas de treinamento. Obtendo um acurácia de 0.871, precisão de 0,833, *recall* de 0.925 e F-Score de 0.877.

Drachen et al. (2016) avaliaram a viabilidade da predição da evasão em jogos grátis-para-jogar baseados no comportamento de usuários a curto prazo, utilizando dados da primeira sessão, dia e semana. O modelo foi desenvolvido com dados do jogo *Jelly Splash* é um jogo de quebra-cabeça gratuito disponível em dispositivos móveis, onde o objetivo é combinar *jellies* da mesma cor para eliminá-las e alcançar a pontuação necessária para completar cada nível. Quando um jogo é lançado, muitas vezes há muitas coisas que precisam ser criadas. Por isso, é útil para grandes e pequenas empresas terem uma regra simples e fácil de implementar para tomar decisões. Para enfrentar esse desafio, introduzimos a ideia de heurística modelagem e previsão (DRACHEN et al., 2016).

Definindo evasão como um problema de classe binária, utilizando um período de até 7 dias para o treinamento e caso os jogadores não voltem a ter atividade nos 7 dias seguintes os categorizando como evadidos (DRACHEN et al., 2016). Foi comparado os resultados da regressão logística, *suport vector machine* (SVM) e floresta aleatória. Os modelos obtiveram uma acurácia razoável de 0.613 e 0.686 utilizando os dados de uma sessão e um dia respectivamente, melhorando consideravelmente para 0.786 quando se utiliza os dados de uma semana (DRACHEN et al., 2016).

Milošević, Živić e Andjelković (2017) na dinâmica de jogos grátis-para-jogar identificaram que a maioria dos usuários recém-registrados abandonam o jogo nos primeiros dias, portanto para evitar evasão uma metodologia de prevenção é essencial para o sucesso de jogos grátis-para-jogar. Para resolver esse problema desenvolveram um sistema em dois estágios, o primeiro um modelo preditivo de evasão utilizando uma classificação binária e segundo um sistema de *feedback* baseado em notificações aos jogadores categorizados como evadidos. Utilizaram uma amostra aleatória de dados do jogo *Top Eleven – Be a Football Manager online* que é um jogo para dispositivos móveis, online com foco no gerenciamento de um time de futebol (MILOŠEVIĆ; ŽIVIĆ; ANDJELKOVIĆ, 2017).

Os treinamentos dos modelos preditivos foram utilizados 5 técnicas distintas, regressão logística, naive bayes, árvore de decisão, gradient boosting e floresta aleatória além disso o período de 14 dias de inatividade foi utilizado na definição de evadidos e recorrentes. Analisando os resultados obtidos performando a técnica *10-fold cross validation* é observável que gradient boosting foi o que melhor performou obtendo AUC de 0.83, precisão de 0.75 e *recall* de 0,84. Em conjunto com o sistema de notificações foi possível reduzir a evasão em até 28% (MILOŠEVIĆ; ŽIVIĆ; ANDJELKOVIĆ, 2017)

Os trabalhos apresentados todos possuem características semelhantes com o nosso, o fundamento de utilizar uma classificação binária para abordar evasão, o contexto de jogos grátis-para-jogar, Várias técnicas de aprendizagem de máquina e balanceamento dos dados como o SMOTE. Contudo nenhum deles possuem a mesma definição de evasão que utilizamos, nenhum dos modelos considera a avaliação temporal de quando um determinado jogador evadirá além disso os modelos são referentes a jogos todos diferentes dos quais estamos utilizando.

Periáñez et al. (2016) desenvolveram um modelo preditivo para evasão de usuários em um jogo social chamado *Age of Ishtaria*, é um jogo para dispositivos móveis incorporando

elementos de RPG, focado no público casual com várias mecânicas sociais como guildas, grupos de chat e eventos. com o jogo obtendo relativo sucesso no Japão. A abordagem clássica para predição de evasão classificando o problema de forma binária entre evadidos(positivos) e recorrentes(negativo), apesar de ser uma forma muito intuitiva não prova quando um determinado jogador parará de jogar, além disso, as *features* são limitadas a prover dados estáticos(não temporais) (PERIÁÑEZ et al., 2016). Para capturar o tempo até a evasão o modelo construído utilizando análise de sobrevivência, uma abordagem que é tradicionalmente utilizada em pesquisas médicas e biológicas e consiste em um conjunto de técnicas utilizadas para predizer a expectativa de vida. Contudo, nesse contexto estas técnicas foram utilizadas para capturar o momento em que o jogador deixa o jogo (PERIÁÑEZ et al., 2016).

Definido como evasão os jogadores que não conectaram ao jogo por 10 dias consecutivos, para as *features* foram utilizadas quantidade da última compra, dias desde a última compra, level, index de lealdade, distancia de atividade, dias até a primeira compra, tempo médio de jogo dos primeiros 15 dias, quantidade da primeira compra e tempo de vida (PERIÁÑEZ et al., 2016). Foram desenvolvidos dois modelos um utilizando técnicas de análise de sobrevivência e outro utilizando regressão de cox. Por fim para comparação com demais modelos foi utilizado a métrica AUC onde o modelo de análise de sobrevivência pontuou 0.960 (PERIÁÑEZ et al., 2016).

Observando o trabalho de Periáñez et al. (2016) destaca-se várias similaridades com o nosso trabalho, sobre o objeto preditivo ser a evasão de jogadores, algumas de suas *features* são utilizadas em nossos modelos e o fato que tomamos como base os dados do jogador assim que o mesmo entre em contato com o jogo. Contudo nosso trabalho caracteriza o problema de evasão como uma abordagem binária entre evadidos(positivos) e recorrentes(negativo) não aplicando a análise de sobrevivência, além disso para capturar o aspecto temporal desenvolvemos 4 modelos diferentes com suas respectivas dimensões temporais e temos o foco de predizer a evasão dos jogadores até o primeiro mês de atividade. Utilizamos um conjunto de dados, *features* e técnicas de aprendizagem de máquinas diferentes.

Capítulo 5

MODELO PREDITIVO PARA EVASÃO DE JOGADORES

Neste capítulo, examinaremos de maneira abrangente os elementos essenciais da análise de dados e aprendizado de máquina. Iniciaremos com a identificação das fontes de dados, seguida detalharemos a da transformação de dados além da seleção das *features*. Posteriormente, abordaremos os modelos preditivos em detalhes, explorando suas características e aplicações em datas de corte específicas. Por fim, encerraremos este capítulo com uma análise detalhada das *features* nos modelos escolhidos como os mais promissores.

5.1 CRANK, O JOGO INCREMENTAL E OS DADOS

O jogo Crank[1] foi desenvolvido e publicado por Faedine[2] é um jogo incremental ou popularmente conhecido como *idle game*. Jogos incrementais se caracterizam por sua mecânica principal ser uma ação básica e repetitiva, que geralmente é o clique do mouse ou um toque em aparelhos móveis. No Crank, girar a manivela gera energia, e essa energia é utilizada para criar outros recursos, que por sua vez são utilizados para melhorar as produção dos mesmos, assim criando um ciclo.

Faedine publicou três bancos de dados sobre seu jogo[3], o crankStatus, crankCombat e o crankTimes. O crankStatus é uma série de *data points* de cada jogador, cada *data points* e criado em media a cada 3 minutos e descreve um retrato da progressão do jogo e seus diversos recursos, possuindo 75 distintas *features* em seu total sendo como exemplo tempo de jogo total, total de recursos obtidos, ID de cada jogador, a imagem 9 mostra todas as *features* disponíveis. Esse banco de dados possui 2 milhões de instâncias divididas entre mais de 14 mil jogos no período de 09/2015 a 03/2016.

O crankCombat é um banco de dados exclusivos de registros apenas dos combates, os dados datam do mesmo período do crankStatus possuindo 3 milhões de *data points*

[1] http://faedine.com/games/crank/b39/
[2] http://faedine.com
[3] https://hoffa.medium.com/gaming-analytics-for-crank-an-incremental-game-62323879d43c

EntryNumber	INTEGER	BatteryAmount	INTEGER	SolarPanelMax	INTEGER	SectorName	STRING	ResearchAntiMatter	INTEGER
GameID	STRING	BatteryMax	INTEGER	SolarPanelMFMax	INTEGER	SectorTechLevel	INTEGER	ResearchScanner	INTEGER
RealTime	TIMESTAMP	BatteryMFMax	INTEGER	SolarPanelBoostMax	INTEGER	SectorSolar	INTEGER	ResearchPhasers	INTEGER
GameTime	INTEGER	BatteryBoostMax	INTEGER	SolarPanelAmountFabricated	INTEGER	SectorEnemy	INTEGER	ResearchPlasma	INTEGER
Debug	INTEGER	BatteryAmountFabricated	INTEGER	ItemAntiMatterAmount	INTEGER	SectorTrader	INTEGER	ResearchPhotonTorp	INTEGER
TimeCranked	FLOAT	CrankBotAmount	INTEGER	ItemPhotonTorpAmount	INTEGER	SectorBeacon	INTEGER	ResearchScrapCannon	INTEGER
PowerGenerated	INTEGER	CrankBotMax	INTEGER	ItemQuantumComputerAmount	INTEGER	SectorWreckage	INTEGER	ResearchHull	INTEGER
PowerGeneratedCrank	INTEGER	CrankBotMFMax	INTEGER	ShipHP	INTEGER	SectorQuest	INTEGER	ResearchShields	INTEGER
PowerGeneratedSolar	INTEGER	CrankBotBoostMax	INTEGER	ShipMaxHP	INTEGER	HelmTotalJumps	INTEGER	ResearchHelm	INTEGER
PowerAmount	INTEGER	CrankBotAmountFabricated	INTEGER	Shields	INTEGER	ResearchSystem	INTEGER	ResearchDecrypter	INTEGER
PowerMax	INTEGER	DuraniumAmount	INTEGER	ShieldsMax	INTEGER	ResearchCrank	INTEGER	DecryptedMessages	INTEGER
ScrapMetalAmount	INTEGER	DuraniumMax	INTEGER	ShieldPower	INTEGER	ResearchScrapMetal	INTEGER		
ScrapMetalMax	INTEGER	DuraniumMFMax	INTEGER	ShieldPowerMax	INTEGER	ResearchBattery	INTEGER		
ScrapMetalMFMax	INTEGER	DuraniumBoostMax	INTEGER	ScrapCannonAmount	INTEGER	ResearchCrankBot	INTEGER		
ScrapMetalBoostMax	INTEGER	DuraniumAmountFabricated	INTEGER	ScannerPower	INTEGER	ResearchDuranium	INTEGER		
ScrapMetalAmountFabricated	INTEGER	SolarPanelAmount	INTEGER	ScannerPowerMax	INTEGER	ResearchSolarPanel	INTEGER		

Figura 9: crankStatus variáveis. Fonte: Autor.

contudo possuindo apenas 13 *features*. CrankTimes consiste em dados relacionados a catalogar quando o jogador atinge certas conquistas, relacionadas principalmente a reparo de itens ou a desbloquear certos itens de progressão. Para o nosso estudo e desenvolvimento do modelo utilizamos apenas o crankStatus.

5.2 TRANSFORMAÇÃO E LIMPEZA DE DADOS

A qualidade e utilidade dos dados são aspectos fundamentais em qualquer análise ou projeto que envolva informações. No contexto do cranckStatus, a organização dos dados foi um ponto crítico para extrair *insights* valiosos. Inicialmente, observamos que os *data points* dos usuários eram gerados em um intervalo médio de 3 minutos. No entanto, para os propósitos de análise e compreensão da interação dos jogadores, era essencial definir o conceito de sessão. Nesse sentido, estabelecemos que uma sessão é encerrada quando o próximo *data point* é criado com um intervalo de tempo real maior do que 10 minutos. Essa abordagem nos permitiu condensar os dados de várias interações em uma única sessão.

A agregação de todos os *data points* de uma sessão em um único ponto de dados resumiu eficazmente toda a atividade do jogador durante aquela sessão. Isso resultou na redução de cerca de 2 milhões de *data points* para apenas 53 mil, sem qualquer perda significativa de informação. Esse processo simplificou enormemente a manipulação dos dados, tornando-os mais gerenciáveis e úteis para análise. Além disso, outra etapa crucial no tratamento dos dados foi a remoção de usuários e dados nulos ou mal formatados do banco de dados. Essa ação foi fundamental para garantir a integridade e a qualidade dos dados utilizados em nossa análise.

5.3 SELEÇÃO DE *FEATURES*

Um dos passos mais importantes para predizer evasão é a seleção de *features* que capturem o comportamento do jogador de forma efetiva. Para este propósito identificamos

Tabela 3: Descrição das *Features* Selecionadas

Tipo das *Features*	*Features*	Número de trabalhos
Atividade	Número de dias	9
Atividade	Total de sessões	6
Atividade	Duração média da sessão	2
Atividade	Tempo de ausência atual	3
Atividade	Tempo de ausência total	2
Atividade	Tempo médio entre sessões	2
Atividade	Tempo de jogo	6
Atividade	Media do tempo jogado	2
Atividade	Desvio padrão do tempo jogado	2

24 trabalhos relacionados a predição de evasão e dentre estes trabalhos encontramos duas formas de categorizar as *features* selecionadas. Uma proposta por foi apresentada por Sifa et al. (2018) subdividido as *features* em 4 categorias distintas, telemetria, temporal, composta e meta. A categoria de telemetria iria possui todas as *features* que não são derivadas, geralmente coletadas sem nenhum tipo de transformação. As categorias temporal e composta são dedicadas as *features* derivadas da telemetria e por fim a categoria meta e relacionada a *features* de meta dados dos jogadores. Proposto por Lee et al. (2016) subdividiu em 4 categorias, Atividade para *features* relacionadas ao tempo e atividades do jogador, Compra para *features* relacionadas a transações envolvendo dinheiro real, Transação para relações de comerciais envolvendo moedas virtuais do jogo e Sociabilidade para interações entre jogadores. Para o presente trabalho utilizamos a categorização proposta por Lee et al. (2016) adicionando uma categoria de Outros para as *features* que não se encaixavam entre as 4 demais categorias.

No contexto de nossa análise, começamos com um conjunto inicial de 73 *features* distintas, que foram descritas nos 24 trabalhos estudados. Essas *features* se distribuem em diferentes categorias, sendo 43 relacionadas à Atividade, 9 à Compra, 7 a Transações, 7 à Sociabilidade e 7 na categoria denominada "Outras". A lista completa de todas as *features* pode ser encontrada no Apêndice, na Tabela A.1.

Para tornar nosso conjunto de *features* mais adequado e relevante para a criação de modelos preditivos, adotamos um critério de seleção rigoroso. Inicialmente, consideramos apenas as *features* que apareceram em mais de um dos 24 trabalhos estudados. Essa abordagem permitiu uma redução significativa no número inicial de *features*, de 73 para 22. No entanto, o processo de seleção não se limitou apenas a critérios de recorrência. Também consideramos a viabilidade de derivar essas *features* a partir dos dados disponíveis em nossa análise. Como resultado, chegamos a uma lista final, apresentada na Tabela 3, que consiste em 9 *features* selecionadas.

5.4 DEFINIÇÃO DE EVASÃO

Para predizer evasão a definimos como um problema de classificação binário, então cada jogador será considerado como evadido(positivo) ou recorrente(negativo). No escopo do nosso trabalho temos como objetivo predizer a evasão de novos jogadores uma vez que a aquisição de novos usuário é mais caro do que a manutenção dos mesmo, logo maximizar a aquisição de novos jogadores é essencial. Sendo assim adotamos uma abordagem similar a proposta por Lee et al. (2016) onde primeiramente utilizamos todos os jogadores disponíveis no *dataset*. Adotamos a data de corte com base na data do primeiro dia de atividade de cada usuário, contudo identificamos que a definição fixa de apenas uma data de corte deixa muito a desejar sobre a janela de tempo de quando um determinado jogador vai deixar de jogar. Sendo assim definimos 4 datas de corte diferentes, uma para 2 dias, 7 dias, 14 dias e 28 dias após a primeira atividade. Para cada caso se um determinado jogador tiver qualquer atividade após a data de corte ele é considerado com jogador recorrente e caso contrário ele é classificado como evadido.

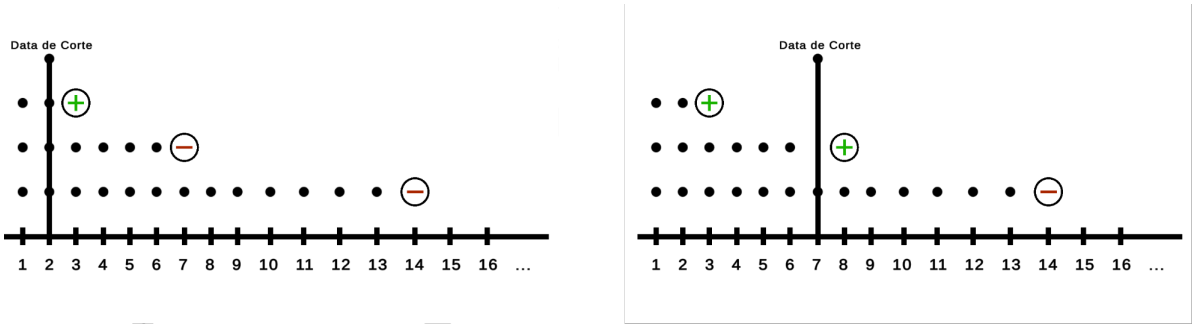

Figura 10: A esquerda temos uma demonstração da data de corte com 2 dias e a direita um exemplo a 7 dias. Fonte: Autor.

Com base em nossa definição de evasão uma vez que determinado jogador é definido como evadido e ou recorrente todas as instâncias de dados deste jogador recebem a mesma categoria sendo estas antes ou após a data de corte. A imagem ilustra como este processo de rotulação acontece em três possíveis jogadores com duas datas de corte distintas, no primeiro exemplo apenas um jogador é categorizado como evadido pois não possui atividade após a data de corte e no segundo exemplo dois jogadores são classificados como evadidos com a data de corte em 7 dias. Sendo assim geramos 4 data sets distintos respectivamente utilizando as datas de corte de 2, 7, 14 e 28 dias.

Observando as estatísticas de evadidos e recorrentes encontrados nos dados evidenciamos na Tabela 4 que as categorias de 7, 14 e 28 dias possuem um desbalanceamento com a quantidade de evadidos crescendo ao decorrer que a data de corte é ampliada, para corrigir esse problema utilizamos uma técnica chamada SMOTE. O SMOTE é uma técnica que utiliza *oversampling* para criar dados sintéticos da classe minoritária, estes dados sintéticos são gerados utilizando amostragens aleatoriamente selecionadas levando em conta dos dados vizinhos com base no algorítimo do K-NN (CHAWLA et al., 2002),

Tabela 4: Estatísticas dos data sets

Data de Corte	Positivos	Negativos
2 dias	0.4477848	0,5522152
7 dias	0.7174314	0,2825686
7 dias com SMOTE	0.5714286	0,4285714
14 dias	0.8348026	0,1651974
14 dias com SMOTE	0.5714286	0,4285714
28 dias	0.9121835	0,0878165
28 dias com SMOTE	0.4705882	0,5294118

em nossa pesquisa utilizamos a biblioteca do R DMwR (versão 0.4.1)[4]. Utilizamos os 5 vizinhos mais próximos para a gerar os dados sintéticos e a Tabela 4 apresenta também as estatísticas das categorias de 7, 14 e 28 dias balanceadas através do SMOTE. Como a categoria de 2 dias se encontra balanceada, possui 44% de positivos e 55% de negativos não identificamos a necessidade de utilizar o SMOTE.

5.5 EXPLORANDO OS MODELOS

Para o desenvolvimento dos modelos adotamos 6 algorítimos de aprendizagem de máquina sendo esta árvore de decisão, floresta aleatória, regressão logística, regressão linear, rede neural e SVM. Para a implementação dos modelos utilizamos da linguagem de programação R com a biblioteca R stats (versão 3.6.2)[5] para os modelos de regressão logística e regressão linear. Usamos a biblioteca rpart (versão 4.1.19)[6] para a árvore de decisão, para a floresta aleatória utilizamos a biblioteca randomForest (versão 4.7-1.1)[7], utilizamos neuralnet (versão 1.44.2)[8] para as redes neurais e por fim a biblioteca e1071 (versão 1.7-12)[9] para a SVM.

A performance dos modelos de aprendizagem de máquina é avaliada através de uma matriz de confusão para problemas de classificação binários. A Tabela 5 descreve bem o funcionamento da matriz onde as colunas representam os dados preditos e a linha a classificação real. Os TP (*true positive*) são o número de casos positivos classificados corretamente, os FP (*false positive*) são o número de casos positivos classificados incorretamente, os TN (*true negative*) são o número de casos negativos classificados corretamente e por fim os FN (*false negative*) são o conjunto de casos negativos classificados de forma incorreta (CHAWLA et al., 2002).

Através da matriz de confusão somos capazes de definir as métricas de avaliação dos modelos como a acurácia que é definida por (TP + TN)/(TP + FP + TN + FN), A precisão é definida pela fórmula TP/(TP + FP) e a *recall* é TP/(TP + FN). A acurácia

[4]https://www.rdocumentation.org/packages/DMwR/versions/0.4.1
[5]https://www.rdocumentation.org/packages/stats/versions/3.6.2
[6]https://www.rdocumentation.org/packages/rpart/versions/4.1.19
[7]https://www.rdocumentation.org/packages/randomForest/versions/4.7-1.1
[8]https://www.rdocumentation.org/packages/neuralnet/versions/1.44.2/topics/neuralnet
[9]https://www.rdocumentation.org/packages/e1071/versions/1.7-12

Tabela 5: Matriz de Confusão

	Preditos Positivos	Preditos Negativos
Real Verdadeiro	TP	FN
Real Falso	FP	TN

Tabela 6: Modelos de 2 dias

Modelo	Acurácia	Precisão	*Recall*	AUC
Arvore de Decisão	0.792	0.722	0.874	0.801
Regressão Logística	0.782	0.684	0.953	0.845
Rede Neural	0.764	0.665	0.952	0.832
Floresta Aleatória	**0.797**	**0.702**	**0.952**	**0.881**
SVM	0.782	0.697	0.908	0.881
Regressão Linear	0.744	0.679	0.812	0.809

indica com que frequência o modelo está correto em suas previsões. A precisão observa dentre todas as classificações da classe positiva quantas o modelo classificou corretamente e o *recall* dentre as classificações positivas como valor esperado quantas estão corretas.

A curva ROC pode ser considerada como representando a família de melhores limiares de divisão por um relativo custo entre os TP e FP. Na curva ROC o eixo X representa %FP = FP/(TN+FP) e o eixo Y representa %TP = TP/(TP +FN). Área abaixo da curva(AUC) é uma métrica útil para avaliar a performance dos modelos uma vez que é independente do critério de decisão e probabilidades prévias. A comparação da AUC pode estabelecer uma dominância entre classificadores(CHAWLA et al., 2002).

Para o treinamento e validação dos modelos dividimos os dados de forma aleatória para treino e teste após a geração dos rótulos, sendo 80% dedicados ao treinamento dos modelos e 20% para o teste, contudo nessa primeira fase não utilizamos *10-fold cross validation*. A Tabela 6 demonstra os resultados dos modelos gerados para 2 dias, é possível observar que a acurácia não chegou a ultrapassar 80% em nenhum dos modelos com também uma precisão baixa estando em torno de 70%. Contudo o *recall* manteve-se alto, acima dos 90% em vários modelos o que indica que dentre a classificação dos que realmente evadiram conseguimos identificar com 95% todos, contudo a precisão baixa indica que ainda tem uma alta taxa de FP(falsos positivos) sendo categorizados. Dentro todos os modelos de 2 dias a Floresta Aleatória se destaca com os melhores métricas.

Nos modelos gerados para 7 dias temos dois casos, um com o conjunto de dados desbalanceados e o outro utilizando o SMOTE pare balancear os dados como referenciado na Tabela 4. Nos modelos da Tabela 7 referente a 7 dias sem o SMOTE é notável que a maioria dos modelos tiveram métricas semelhantes sendo a floresta aleatória que teve melhor desempenho com a melhor AUC dentre todos os modelos. Na Tabela 8 referente a 7 dias com SMOTE temos uma queda geral nas métricas da maioria dos modelos, contudo a floresta aleatória teve um ganho significativo de desempenho melhorando todas as suas métricas se destacando sendo o primeiro modelo a ter uma AUC acima de 95%.

Tabela 7: Modelos de 7 dias

Modelo	Acurácia	Precisão	*Recall*	AUC
Arvore de Decisão	0.851	0.836	0.986	0.746
Regressão Logística	0.825	0.860	0.902	0.823
Rede Neural	0.797	0.847	0.877	0.794
Floresta Aleatória	**0.856**	**0.875**	**0.932**	**0.868**
SVM	0.845	0.839	0.969	0.748
Regressão Linear	0.805	0.856	0.876	0.802

Tabela 8: Modelos de 7 dias com SMOTE

Modelo	Acurácia	Precisão	*Recall*	AUC
Arvore de Decisão	0.785	0.735	0.972	0.755
Regressão Logística	0.792	0.760	0.927	0.828
Rede Neural	0.767	0.728	0.945	0.788
Floresta Aleatória	**0.918**	**0.914**	**0.946**	**0.975**
SVM	0.792	0.754	0.943	0.768
Regressão Linear	0.771	0.755	0.885	0.808

Analisando os modelos de 14 dias sem SMOTE da Tabela 9 é notável um ganho de desempenho em todas as métricas por todos os modelos com a árvore de decisão e SVM tendo os melhores resultados referentes a acurácia, precisão e com o *recall* dos dois modelos em 99%, contudo é observado que a AUC dos dois modelos é baixa e sabendo que se trata de um conjunto de dados não balanceado é provável que os modelos em questão estejam sobre-ajustados por consequência do conjunto de dados. Por fim temos a floresta aleatória com índices semelhantes de acurácia, precisão e *recall* porém com uma AUC muito melhor do que os demais modelos.

Aplicando o SMOTE para balancear os dados de 14 dias, observado na Tabela 10 novamente é notável uma queda em todas as métricas em quase todos os modelos, semelhante ao que aconteceu com o modelo de 7 dias com SMOTE. Contudo a floresta aleatória novamente se destaca tendo uma melhora muito significativa entre todas as suas métricas sendo o segundo modelo com uma AUC acima de 95%.

Os modelos de 28 dias seguiram o padrão observador pelos modelos de 7 e 14 dias, onde quando treinado sobre o conjunto de dados não balanceado todos os modelos tiveram um aumento significado nas métricas de acurácia, precisão e *recall* contudo ao custo de uma baixa AUC, o que indica um alto nível de sobre-ajustados dos modelos observados na Tabela 11. Em contra partida a floresta aleatória mesmo com o conjunto de dados desbalanceado manteve métricas muito boas e uma AUC superior a 85%. Observando a Tabela 12 temos os dados agora balanceados utilizando o SMOTE, novamente observamos um queda generalizada nas métricas de todos os modelos, Contudo a floresta aleatória volta a ter um desempenho muito bom com todas suas métricas acima de 90% sendo o terceiro modelo com uma AUC acima de 95%.

Tabela 9: Modelos de 14 dias

Modelo	Acurácia	Precisão	*Recall*	AUC
Arvore de Decisão	0.901	0.898	0.994	0.711
Regressão Logística	0.836	0.926	0.873	0.818
Rede Neural	0.848	0.902	0.917	0.777
Floresta Aleatória	**0.902**	**0.933**	**0.950**	**0.873**
SVM	0.905	0.901	0.995	0.720
Regressão Linear	0.856	0.920	0.906	0.804

Tabela 10: Modelos de 14 dias com SMOTE

Modelo	Acurácia	Precisão	*Recall*	AUC
Arvore de Decisão	0.779	0.738	0.951	0.751
Regressão Logística	0.777	0.765	0.881	0.818
Rede Neural	0.717	0.704	0.872	0.757
Floresta Aleatória	**0.904**	**0.905**	**0.930**	**0.965**
SVM	0.778	0.741	0.940	0.752
Regressão Linear	0.770	0.741	0.917	0.805

Tabela 11: Modelos de 28 dias

Modelo	Acurácia	Precisão	*Recall*	AUC
Arvore de Decisão	0.944	0.945	0.997	0.689
Regressão Logística	0.890	0.957	0.920	0.796
Rede Neural	0.743	0.951	0.758	0.713
Floresta Aleatória	**0.895**	**0.968**	**0.915**	**0.863**
SVM	0.942	0.941	0.999	0.667
Regressão Linear	0.864	0.958	0.891	0.787

Tabela 12: Modelos de 28 dias Com SMOTE

Modelo	Acurácia	Precisão	*Recall*	AUC
Arvore de Decisão	0.739	0.684	0.823	0.745
Regressão Logística	0.735	0.678	0.827	0.801
Rede Neural	0.595	0.538	0.951	0.764
Floresta Aleatória	**0.925**	**0.911**	**0.932**	**0.980**
SVM	0.716	0.636	0.923	0.729
Regressão Linear	0.713	0.651	0.836	0.779

Tabela 13: Melhores modelos com *10-fold cross validation*

Floresta Aleatória	2 dias	7 dias com SMOTE	14 dias com SMOTE	28 dias com SMOTE
Acurácia	0.797	0.922	0.906	0.934
Precisão	0.700	0.944	0.916	0.924
Recall	0.953	0.917	0.919	0.936
AUC	0.881	0.976	0.967	0.984

Após a avaliação dos modelos selecionamos os melhores modelos de cada categoria de tempo e retreinamos os modelos aplicando a técnica *10-fold cross validation*, esta técnica é comumente aplicada para melhor estimar o erro de teste do modelo (GARETH et al., 2013). O erro de teste é definido pelas variais que estamos utilizando para avaliar o modelo sendo esta acurácia, precisão, *recall* e AUC. *10-fold cross validation* consiste em aleatoriamente dividir o conjunto de dados em 10 grupos, ou *folds*, de tamanhos aproximados. O primeiro *fold* é tratado como o conjunto de validação e o método é treinar os *folds* restantes usando o *fold* de validação como teste, esse procedimento é repetido 10 vezes; cada vez com um *fold* diferente sendo usado como validação (GARETH et al., 2013). Esse processo resulta em 10 estimativas de erros de testes, onde a média está disponível na Tabela 13.

Observando as métricas dos modelos utilizando *10-fold cross validation* é possível notar que não ouve queda de desempenho se comparado com suas versões prévias, em alguns casos até possuindo um pequeno ganho de desempenho como os modelos de 7 e 28 dias com SMOTE. Assim validamos os resultados obtidos e mitigamos possíveis viés inerentes do conjunto de dados utilizado.

5.6 ANALISE DAS *FEATURES* E MODELOS

Observando os quatro melhores modelos gerados, decidimos investigar a importância de cada *features*, para este fim criamos quatro gráficos que demonstram o nível de importância de cada *features* aos quais os modelos determinam esse valor para categorizar o nível de importância da informação na tomada de decisão ao categorizar um jogador como evadido ou recorrente. A imagem 11 é o gráfico de importância das *features* geradas pelo modelo de 2 dias com a floresta aleatória, nesse gráfico é notável que duas *features* se sobressaem sobre as demais. O número de dias e o tempo de ausência são indicados como as *features* mais importantes pelo modelo. Além deles outras duas *features* sugerem de forma significativas que são o tempo entre sessões e o tempo médio das sessões, temos cinco *features* que contribuem muito pouco para o modelo se comparado com as quatro principais que são em ordem de importância, desvio padrão, Tempo de jogo, sessões, média das sessões e a média de tempo de jogo.

O modelo de 7 dias com SMOTE gerou o gráfico da imagem 12, aqui já é possível notar um padrão que permanece sobre os 4 modelos uma vez que temos novamente como as *features* mais importantes o número de dias e o tempo de ausência, contudo nesse

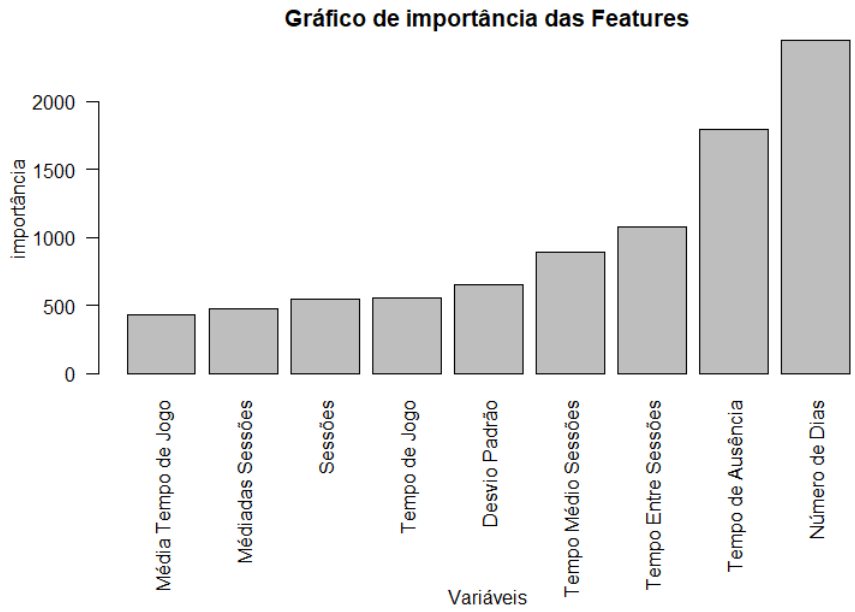

Figura 11: Importância das *Features* para o modelo de 2 dias Fonte: Autor.

modelo é notável que estas duas *features* possuem ainda mais importância derivando a maior parte da tomada de decisão. As duas *features* subsequente o tempo entre sessões e o tempo médio das sessões teve um notável declínio na importância do modelo se aproximando das demais cinco menos importantes. Contudo apesar de existir uma certa variação na importância do desvio padrão, Tempo de jogo, sessões, média das sessões e a média de tempo de jogo, aqui é notável que o desvio padrão perdeu importância e junto a sessões se comparado ao modelo de 2 dias.

O modelo de 14 dias com SMOTE continua seguindo o padrão de importância das *features* estabelecido. A imagem 13 demonstra como o número de dias e o tempo de ausência continuam sendo as mais importantes, contudo perderam importância se comparado ao de 7 dias com SMOTE, voltando aos valores similares do modelo de 2 dias. As demais *features* seguem o mesmo padrão sem variação se comparado ao modelo de 7 dias com SMOTE.

O modelo de 28 dias com SMOTE permanece com o padrão das *features* estabelecido, contudo este demonstra ser o modelo mais balanceado com relação a importância das *features* demonstrado na imagem 14. O número de dias e o tempo de ausência continuam as mais relevantes, contudo o tempo entre sessões e o tempo médio das sessões ganharam muita importância estando muito próximos das duas primeiras e esse padrão se estabelece para todas as demais. Apesar da ordem de importância não ter se alterado todas as *features* se tornaram muito mais importantes para a tomada de decisão do modelo se comparado aos três prévios.

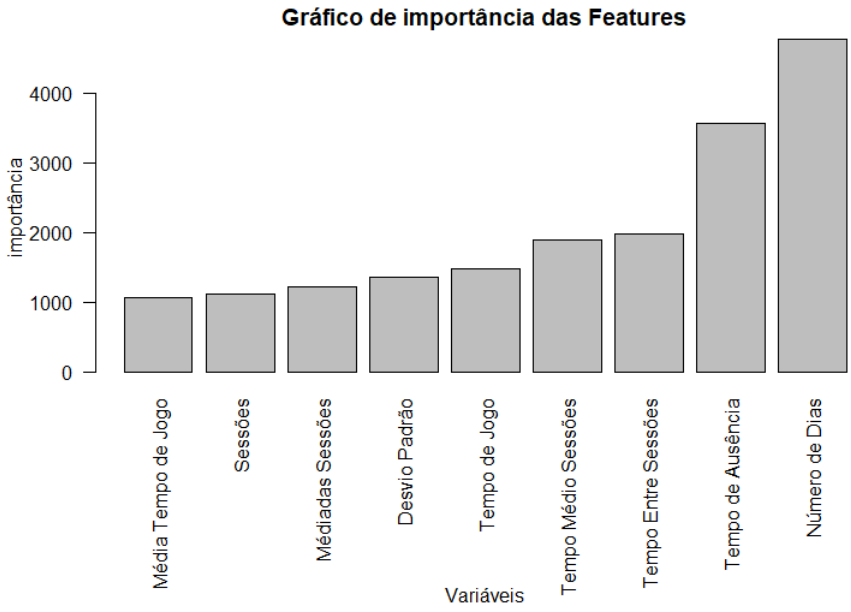

Figura 12: Importância das *Features* para o modelo de 7 dias Fonte: Autor.

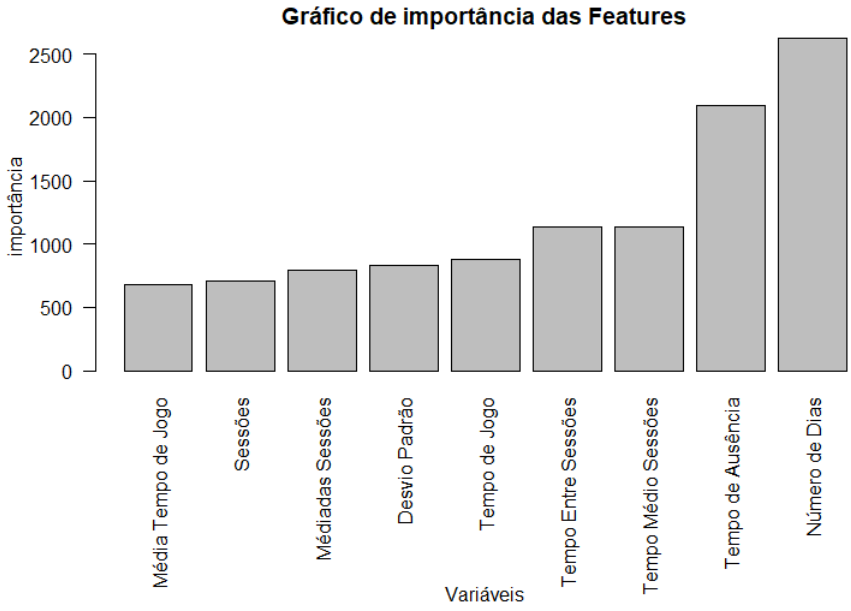

Figura 13: Importância das *Features* para o modelo de 14 dias Fonte: Autor.

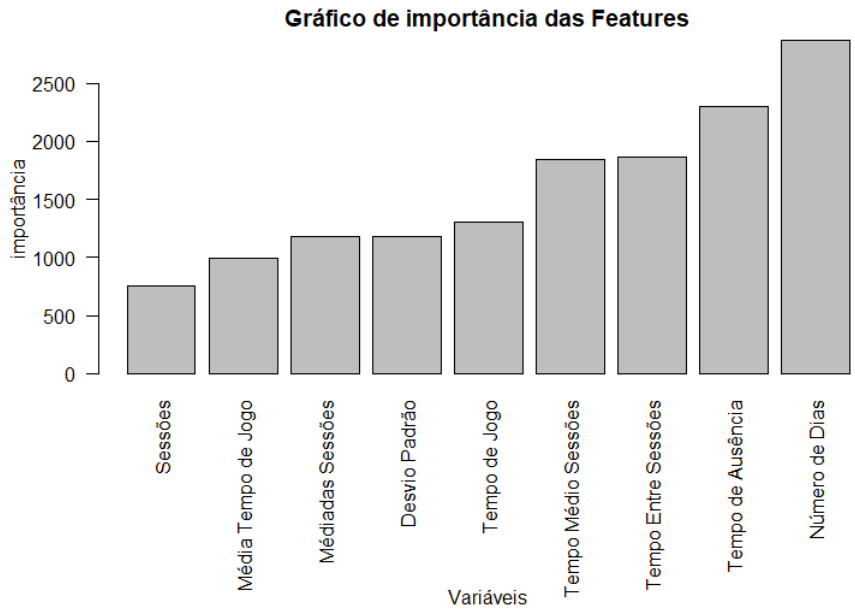

Figura 14: Importância das *Features* para o modelo de 28 dias Fonte: Autor.

5.7 AMEAÇAS A VALIDADE

O desenvolvimento de modelos preditivos para a evasão de jogadores em um ambiente de jogo online é uma tarefa complexa que envolve muitos aspectos críticos. Nesta seção, abordaremos as ameaças à validade do nosso trabalho, considerando os processos e métodos adotados.

Definição de Evasão e Data de Corte: Uma das principais ameaças à validade do nosso trabalho está relacionada à definição de evasão e às datas de corte. Optamos por considerar um problema de classificação binária, onde os jogadores são classificados como evadidos ou recorrentes. No entanto, essa definição pode ser subjetiva e variar dependendo do contexto do jogo. Além disso, o uso de datas de corte fixas, como 2, 7, 14 e 28 dias após a primeira atividade, pode não capturar com precisão o momento da evasão de um jogador. Isso pode levar a classificações errôneas, pois os jogadores podem interromper o jogo em momentos não coincidentes com as datas de corte.

Desequilíbrio de Classes e Uso de SMOTE: Observamos que o desequilíbrio nas classes de evadidos e recorrentes era evidente em várias das datas de corte, com um aumento na quantidade de evadidos à medida que a data de corte era estendida. Para mitigar esse desequilíbrio, aplicamos a técnica SMOTE para gerar dados sintéticos da classe minoritária. Embora isso tenha contribuído para a melhoria das métricas de desempenho, o uso do SMOTE também apresenta ameaças à validade. O SMOTE gera dados sintéticos com base nos vizinhos mais próximos, o que pode introduzir artefatos nos dados e potencialmente levar a resultados distorcidos. A aplicação de técnicas de balanceamento de classes deve ser realizada com cautela, pois pode afetar a qualidade e a interpretabilidade do modelo.

Importância das features: A análise da importância das *features* em nossos modelos revelou as características mais relevantes para a previsão da evasão. No entanto, a importância das *features* pode variar entre diferentes conjuntos de dados e contextos de jogo. Portanto, a interpretação das *features* mais importantes deve ser feita com consideração das peculiaridades do jogo estudado.

Em conclusão, nossa análise sobre a previsão da evasão de jogadores em um ambiente de jogo grátis-para-jogar apresenta resultados promissores, mas também destaca diversas considerações críticas. Ao adotar uma abordagem de classificação binária e a definição de datas de corte para identificar a evasão, enfrentamos desafios em termos de subjetividade e variação na interpretação da evasão. A aplicação do SMOTE para equilibrar as classes foi essencial, embora tenha introduzido complexidades potenciais na qualidade dos dados. A seleção de algoritmos de aprendizado de máquina, métricas de avaliação e a técnica de *10-fold cross validation* foram decisões significativas que moldaram nossa análise. Além disso, a importância das *features* destacou aspectos críticos para a previsão da evasão, mas sua interpretação requer uma compreensão profunda do contexto do jogo. Em última análise, nosso trabalho demonstra o potencial de prever a evasão de jogadores, mas ressalta a importância de abordagens cuidadosas, consideração do contexto e avaliação contínua da validade dos resultados para obter *insights* confiáveis e relevantes para a maximização da aquisição de novos jogadores em ambientes de jogos grátis-para-jogar.

Capítulo

6

CONCLUSÃO

Análise Preditiva é um componente central e indispensável para a tomada de decisões baseadas em dados em diversos tipos de negócios. Nos jogos, não é diferente. Como observado em nosso mapeamento, a análise preditiva pode ser aplicada em diferentes aspectos do desenvolvimento de jogos, como probabilidade de vitória e previsão da experiência dos jogadores, entre outros descritos na Seção 3.2. Contudo, escolhemos estudar a predição de evasão, uma vez que este é um tópico amplamente estudado e de alto impacto, especialmente quando associado a jogos do modelo grátis-para-jogar. Isso ocorre devido às altas taxas de evasão, uma vez que uma grande quantidade de jogadores desinstalam os jogos nos primeiros dias após a interação (DRACHEN et al., 2016).

Nosso trabalho tem como referência o estudo desenvolvido por Hadiji et al. (2014), onde foi proposto um modelo preditivo para evasão de jogadores, utilizando uma seleção de *features* genéricas e testado com cinco jogos distintos. No entanto, algumas *features* selecionadas em seu modelo não podem ser aplicadas a qualquer jogo, como é o caso de *features* de monetização, gastos por sessão ou identificação de usuários pagantes. Conscientes das limitações desse modelo, em nosso estudo, identificamos, categorizamos e selecionamos as *features* mais comuns com base no mapeamento da literatura apresentado na Seção 3.

As *features* selecionadas para o nosso modelo estão descritas na Tabela 3. No total, foram nove *features* selecionadas. Para evitar o problema de *features* que não podem ser derivadas, todas as *features* do nosso modelo podem ser geradas a partir de dois dados fundamentais: o tempo de jogo e a data de quando foi jogado. Essas informações são simples de serem coletadas. Essa abordagem torna a seleção de *features* facilmente replicável para outros jogos. Desenvolvemos os modelos preditivos com base em quatro variações temporais: 2 dias, 7 dias, 14 dias e 28 dias. Esses modelos foram treinados usando vários algoritmos de aprendizado de máquina, dos quais selecionamos com base nas técnicas encontradas em nosso mapeamento, como descrito na Seção 3.2. Isso resultou em 42 modelos finais para todos os períodos de corte, aplicando a técnica SMOTE, conforme evidenciado na Seção 5.5.

Selecionamos os melhores modelos gerados, um para cada período de corte, resultando em quatro modelos finais. Os resultados e métricas desses modelos podem ser observados na Tabela 13. Obtemos excelentes valores de precisão, acurácia, *recall* e AUC para todos os períodos de corte, o que demonstra que é possível prever a evasão com alta precisão usando *features* genéricas em todos os períodos temporais. Comparando nossos resultados com os da literatura, como o estudo de Sifa et al. (2018), observamos que, em muitos casos, nossos modelos apresentaram valores de *recall* semelhantes ou superiores.

6.1 TRABALHOS FUTUROS

Observando os resultados no contexto deste estudo, várias oportunidades de pesquisa emergem para trabalhos futuros, como a melhoria dos modelos preditivos, uma vez que é possível explorar técnicas de aprendizado de máquina mais avançadas e sofisticadas para melhorar a precisão das previsões de evasão em jogos grátis-para-jogar. Além disso a aplicação do modelo a outras bases de dados com a possibilidade de testes em múltiplos jogos. Generalização para Outros Gêneros de Jogos uma vez que concentramos apenas em jogos grátis-para-jogar, mas a pesquisa pode ser estendida para outros gêneros de jogos, como jogos de console, jogos para dispositivos móveis e jogos de realidade virtual, para avaliar a eficácia dos modelos preditivos em diferentes contextos.

REFERÊNCIAS BIBLIOGRÁFICAS

ALEEM, S.; CAPRETZ, L. F.; AHMED, F. Game development software engineering process life cycle: a systematic review. *Journal of Software Engineering Research and Development*, SpringerOpen, v. 4, n. 1, p. 6, 2016.

ALHA, K. et al. Free-to-play games: Professionals' perspectives. *Proceedings of nordic DiGRA*, v. 2014, 2014.

ANDRADE, L. A. Jogos pervasivos: Educação, cultura e cidade digital. *Revista Opara*, v. 3, n. 1, 2013.

BORBORA, Z. H.; SRIVASTAVA, J. User behavior modelling approach for churn prediction in online games. In: IEEE. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. [S.l.], 2012. p. 51–60.

CAILLOIS, R. *Man, play, and games*. [S.l.]: University of Illinois Press, 1961.

CASTRO, E. G.; TSUZUKI, M. S. Churn prediction in online games using players' login records: A frequency analysis approach. *IEEE Transactions on Computational Intelligence and AI in Games*, IEEE, v. 7, n. 3, p. 255–265, 2015.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

DEMEDIUK, S. et al. Player retention in league of legends: a study using survival analysis. In: *Proceedings of the Australasian computer science week multiconference*. [S.l.: s.n.], 2018. p. 1–9.

DRACHEN, A.; EL-NASR, M. S.; CANOSSA, A. Game analytics – the basics. In: _____. *Game Analytics: Maximizing the Value of Player Data*. London: Springer London, 2013. p. 13–40. ISBN 978-1-4471-4769-5. Disponível em: <https://doi.org/10.1007/978-1-4471-4769-5_2>.

DRACHEN, A. et al. Rapid prediction of player retention in free-to-play mobile games. *arXiv preprint arXiv:1607.03202*, 2016.

EL-NASR, M. S.; DRACHEN, A.; CANOSSA, A. Introduction. In: _____. *Game Analytics: Maximizing the Value of Player Data*. London: Springer London, 2013. p. 3–12. ISBN 978-1-4471-4769-5. Disponível em: <https://doi.org/10.1007/978-1-4471-4769-5_1>.

EL-NASR, M. S.; DRACHEN, A.; CANOSSA, A. *Game analytics*. [S.l.]: Springer, 2016.

ENGSTRÖM, H. et al. Game development from a software and creative product perspective: A quantitative literature review approach. *Entertainment Computing*, Elsevier, v. 27, p. 10–22, 2018.

ERDEM, A. N.; HALICI, U. Applying computational aesthetics to a video game application using machine learning. *IEEE Computer Graphics and Applications*, IEEE, v. 36, n. 4, p. 23–33, 2016.

FARRIER, M. et al. Game development. 2012.

GAGNÉ, A. R.; EL-NASR, M. S.; SHAW, C. D. A deeper look at the use of telemetry for analysis of player behavior in rts games. In: SPRINGER. *International Conference on Entertainment Computing*. [S.l.], 2011. p. 247–257.

GARETH, J. et al. *An introduction to statistical learning: with applications in R*. [S.l.]: Spinger, 2013.

HADIJI, F. et al. Predicting player churn in the wild. In: IEEE. *Computational intelligence and games (CIG), 2014 IEEE conference on*. [S.l.], 2014. p. 1–8.

HAMARI, J. et al. Why do players buy in-game content? an empirical study on concrete purchase motivations. *Computers in Human Behavior*, Elsevier, v. 68, p. 538–546, 2017.

HAMARI, J.; HANNER, N.; KOIVISTO, J. Service quality explains why people use freemium services but not if they go premium: An empirical study in free-to-play games. *International Journal of Information Management*, Elsevier, v. 37, n. 1, p. 1449–1459, 2017.

HUIZINGA, J. *Homo ludens: o jogo como elemento da cultura*. [S.l.]: Editora da Universidade de S. Paulo, Editora Perspectiva, 1971.

HULLETT, K. et al. Empirical analysis of user data in game software development. In: IEEE. *Empirical Software Engineering and Measurement (ESEM), 2012 ACM-IEEE International Symposium on*. [S.l.], 2012. p. 89–98.

KITCHENHAM STUART CHARTERS, D. B. P. B. M. T. S. L. M. J. E. M. G. V. B. Guideline for performing systematic literature reviews in software engineering. *School of Computer Science and Mathematics Keele University*, v. 2007, 2007.

KOULIERIS, G. A. et al. Gaze prediction using machine learning for dynamic stereo manipulation in games. In: IEEE. *2016 IEEE virtual reality (VR)*. [S.l.], 2016. p. 113–120.

LEE, S.-K. et al. Predicting churn in mobile free-to-play games. In: IEEE. *Information and Communication Technology Convergence (ICTC), 2016 International Conference on*. [S.l.], 2016. p. 1046–1048.

LI, S.; CHEN, C. Interest scheme: A new method for path prediction. In: *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*. [S.l.: s.n.], 2006. p. 41–es.

LIM, C.-U.; HARRELL, D. F. Modeling player preferences in avatar customization using social network data: A case-study using virtual items in team fortress 2. In: IEEE. *2013 IEEE Conference on Computational Inteligence in Games (CIG)*. [S.l.], 2013. p. 1–8.

MAHLMANN, T. et al. Predicting player behavior in tomb raider: Underworld. In: IEEE. *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*. [S.l.], 2010. p. 178–185.

MILOŠEVIĆ, M.; ŽIVIĆ, N.; ANDJELKOVIĆ, I. Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, Elsevier, v. 83, p. 326–332, 2017.

OLIVEIRA, J. K. Coerção, manipulação e tecnologia: estratégias de monetização em jogos free-to-play. Universidade Federal de São Carlos, 2022.

PAAVILAINEN, J. et al. Social network games: Players' perspectives. *Simulation & Gaming*, SAGE Publications Sage CA: Los Angeles, CA, v. 44, n. 6, p. 794–820, 2013.

PERIÁÑEZ, Á. et al. Churn prediction in mobile social games: towards a complete assessment using survival ensembles. In: IEEE. *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. [S.l.], 2016. p. 564–573.

PETERSEN, K. et al. Systematic mapping studies in software engineering. In: *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*. [S.l.: s.n.], 2008. p. 1–10.

RAMADAN, R.; WIDYANI, Y. Game development life cycle guidelines. In: IEEE. *Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on*. [S.l.], 2013. p. 95–100.

RUNGE, J. et al. Churn prediction for high-value players in casual social games. In: IEEE. *Computational Intelligence and Games (CIG), 2014 IEEE Conference on*. [S.l.], 2014. p. 1–8.

SALEN, K.; ZIMMERMAN, E. *Rules of play: Game design fundamentals*. [S.l.]: MIT press, 2004.

SHEN, C.; WILLIAMS, D. Unpacking time online: Connecting internet and massively multiplayer online game use with psychosocial well-being. *Communication Research*, Sage Publications Sage CA: Los Angeles, CA, v. 38, n. 1, p. 123–149, 2011.

SIFA, R. et al. Customer lifetime value prediction in non-contractual freemium settings: Chasing high-value users using deep neural networks and smote. In: *Proceedings of the 51st Hawaii International Conference on System Sciences*. [S.l.: s.n.], 2018.

SIFA, R. et al. Predicting retention in sandbox games with tensor factorization-based representation learning. In: IEEE. *2016 IEEE Conference on Computational Intelligence and Games (CIG)*. [S.l.], 2016. p. 1–8.

TAMASSIA, M. et al. Predicting player churn in destiny: A hidden markov models approach to predicting player departure in a major online game. In: IEEE. *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*. [S.l.], 2016. p. 1–8.

TYCHSEN, A. Crafting user experience via game metrics analysis. In: *Workshop Research Goals and Strategies for Studying User Experience and Emotion, part of NordiCHI 2008*. [S.l.: s.n.], 2008.

VELLOSO, L. M. R. *O espaço nos videogames: dentro e fora do círculo mágico*. Tese (Doutorado) — Universidade de São Paulo, 2017.

VILJANEN, M. et al. Playtime measurement with survival analysis. *IEEE Transactions on Games*, IEEE, v. 10, n. 2, p. 128–138, 2017.

VILJANEN, M. et al. Modelling user retention in mobile games. In: IEEE. *2016 IEEE Conference on Computational Intelligence and Games (CIG)*. [S.l.], 2016. p. 1–8.

VINKEMEIER, D.; VALSTAR, M.; GRATCH, J. Predicting folds in poker using action unit detectors and decision trees. In: IEEE. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. [S.l.], 2018. p. 504–511.

WEBER, B. G. et al. Modeling player retention in madden nfl 11. In: *IAAI*. [S.l.: s.n.], 2011.

XIE, H.; DEVLIN, S.; KUDENKO, D. Predicting disengagement in free-to-play games with highly biased data. In: *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*. [S.l.: s.n.], 2016.

XIE, H. et al. Predicting player disengagement and first purchase with event-frequency based data representation. In: IEEE. *2015 IEEE Conference on Computational Intelligence and Games (CIG)*. [S.l.], 2015. p. 230–237.

Apêndice A

A.1 LISTA COMPLETA DE *FEATURES*

Tipo de Feature	Features	Número de trabalhos	Referências
Atividade	Número de dias	9	(LEE et al., 2016),(PERIÁÑEZ et al., 2016),(DRACHEN et al., 2016),(RUNGE et al., 2014),(HADIJI et al., 2014),(SIFA et al., 2018),(XIE; DEVLIN; KUDENKO, 2016),(XIE et al., 2015),(SIFA et al., 2016)
	Distribuição de logins diaria	2	(LEE et al., 2016),(SIFA et al., 2018)
	Distribuição da hora do dia	2	(SIFA et al., 2018),(SIFA et al., 2016)
	Distribuição das sessões do dia	2	(SIFA et al., 2018),(SIFA et al., 2016)
	Logs de eventos por dia	1	(LEE et al., 2016)
	Frequência de eventos	2	(XIE; DEVLIN; KUDENKO, 2016),(PERIÁÑEZ et al., 2016)
	Precisão	3	(RUNGE et al., 2014),(XIE; DEVLIN; KUDENKO, 2016),(XIE et al., 2015)
	Quantidade de espada no último dia de logout	1	(LEE et al., 2016)
	Distribuição de espada diaria	1	(LEE et al., 2016)
	Nível do personagem no último dia de logout	1	(LEE et al., 2016)
	Distribuição de Nível Diário	1	(LEE et al., 2016)
	Duração da vida	1	(PERIÁÑEZ et al., 2016)
	Level	3	(PERIÁÑEZ et al., 2016),(DRACHEN et al., 2016),(RUNGE et al., 2014)
	Index de lealdade	1	(PERIÁÑEZ et al., 2016)
	Ação atividade distância	1	(PERIÁÑEZ et al., 2016)
	Tempo medio jogado das primerias 15 semanas	1	(PERIÁÑEZ et al., 2016)

Total de sessões	6	(DRACHEN et al., 2016),(BORBORA; SRIVASTAVA, 2012),(MILOŠEVIĆ; ŽIVIĆ; ANDJELKOVIĆ, 2017),(HADIJI et al., 2014),(SIFA et al., 2018),(SIFA et al., 2016)
Duração média da sessão	2	(DRACHEN et al., 2016),(HADIJI et al., 2014)
Duração media recente de uma partida	1	(DEMEDIUK et al., 2018)
Duração da sessão	1	(VILJANEN et al., 2016)
Tempo de jogo total	1	(DRACHEN et al., 2016)
Tempo de ausência atual	3	(DRACHEN et al., 2016),(HADIJI et al., 2014),(TAMASSIA et al., 2016)
Tempo de ausência total	2	(SIFA et al., 2018),(SIFA et al., 2016)
Tempo médio entre sessões	2	(DRACHEN et al., 2016),(HADIJI et al., 2014)
Tempo médio entre partidas recente,	1	(DEMEDIUK et al., 2018)
Tempo até próxima partida	1	(DEMEDIUK et al., 2018)
Média do tempo de vida	1	(TAMASSIA et al., 2016)
Proporção de Atividades Concluídas	1	(TAMASSIA et al., 2016)
Tempo de jogo	6	(MILOŠEVIĆ; ŽIVIĆ; ANDJELKOVIĆ, 2017),(HADIJI et al., 2014),(SIFA et al., 2018),(VILJANEN et al., 2017),(SIFA et al., 2016),(MAHLMANN et al., 2010)
Transformada discreta de wavelet	1	(CASTRO; TSUZUKI, 2015)
Decomposição do sinal por Wavelet Packet	1	(CASTRO; TSUZUKI, 2015)
Espectro de Potência Wavelet	1	(CASTRO; TSUZUKI, 2015)
Plano de tempo-frequência wavelet	1	(CASTRO; TSUZUKI, 2015)

	Número de Rounds	4	(RUNGE et al., 2014),(SIFA et al., 2018),(XIE; DEVLIN; KUDENKO, 2016),(XIE et al., 2015)
	Tempo total entre sessões	1	(SIFA et al., 2018)
	Tempo total entre partidas	1	(SIFA et al., 2018)
	Número total de dias da semana	1	(SIFA et al., 2018)
	Tempo entre a primeira e a ultima sessão diária	2	(SIFA et al., 2018),(SIFA et al., 2016)
	Número de Ações	1	(SIFA et al., 2016)
	Coeficientes de correlação no tempo	2	(SIFA et al., 2018),(SIFA et al., 2016)
	Média e desvio padram do tempo	2	(SIFA et al., 2018),(SIFA et al., 2016)
	Progresso	1	(SIFA et al., 2016)
	Número de mortes	2	(SIFA et al., 2016),(MAHLMANN et al., 2010)
Compra	Número de Compras	5	(LEE et al., 2016),(PERIÁÑEZ et al., 2016),(MILOŠEVIĆ; ŽIVIĆ; ANDJELKOVIĆ, 2017),(HADIJI et al., 2014),(SIFA et al., 2018)
	Quantia da última compra	4	(PERIÁÑEZ et al., 2016),(RUNGE et al., 2014),(XIE; DEVLIN; KUDENKO, 2016),(XIE et al., 2015)
	Quantia da primeira compra	1	(PERIÁÑEZ et al., 2016)
	Dias desde a última compra	4	(PERIÁÑEZ et al., 2016),(RUNGE et al., 2014),(XIE; DEVLIN; KUDENKO, 2016),(XIE et al., 2015)
	Dias desde a primeira compra	1	(PERIÁÑEZ et al., 2016)
	Tag de Usuário premium	1	(HADIJI et al., 2014)
	Categoria de gastos predefinida	1	(HADIJI et al., 2014)
	Gasto médio por sessão	1	(HADIJI et al., 2014)
	Valor total da compra	1	(SIFA et al., 2018)

Transação	Quantidade de ouro no último dia de logout	1	(LEE et al., 2016)
	Quantidade de Ruby no último dia de logout	1	(LEE et al., 2016)
	Distribuição diária de ouro	1	(LEE et al., 2016)
	Distribuição diária de rubis	1	(LEE et al., 2016)
	Número de Vezes Consumindo Ruby	1	(LEE et al., 2016)
	Dinheiro no jogo gasto	2	(MILOŠEVIĆ; ŽIVIĆ; ANDJELKOVIĆ, 2017),(RUNGE et al., 2014)
	Quantidade de moeda do jogo	1	(LEE et al., 2016)
Sociabilidade	Quantidade de corações no último dia de logout	1	(LEE et al., 2016)
	Distribuição diária de corações	1	(LEE et al., 2016)
	Número de vezes que participa da guilda	2	(LEE et al., 2016),(RUNGE et al., 2014)
	Participação em uma guilda	1	(PERIÁÑEZ et al., 2016)
	amigos conectados	1	(DRACHEN et al., 2016)
	Convites enviados	2	(RUNGE et al., 2014),(XIE; DEVLIN; KUDENKO, 2016)
	Conexões de redes sociais	1	(SIFA et al., 2016)
Outra	País	1	(LEE et al., 2016)
	Nível de temporada mais alto alcançado	1	(DEMEDIUK et al., 2018)
	Tipo de dispositivo	1	(LEE et al., 2016)
	Sistema operacional	1	(LEE et al., 2016),(SIFA et al., 2018)
	Tipo de aquisição	1	(LEE et al., 2016)
	Plataforma jogada	1	(SIFA et al., 2016)
	Idioma	1	(SIFA et al., 2016)