Análise Preditiva da Receita de Jogos na Plataforma Steam

Gabriel de França Marques (RA: 10395270)[1], **Henrique Magno dos Santos(RA: 10335286)**[1], **Pedro Machado Gomes Caixeta (RA: 10314309)**[1]

[1]Ciência da Computação (CC) – Faculdade de Ciência e Informação (FCI) – Universidade Presbiteriana Mackenzie (UPM)

10395270@, 10335286@, 10314309@mackenzista.com.br

Resumo. *Este trabalho apresenta uma análise preditiva da receita de jogos na plataforma Steam. Utilizando um conjunto de dados de 1500 jogos e o algoritmo Random Forest Regressor, o objetivo é prever a receita com base em características como preço, cópias vendidas, tempo médio de jogo, avaliação dos usuários, classe da publicadora e ano de lançamento. Os resultados são avaliados pelas métricas RMSE e R^2. O modelo apresentou um alto R^2 (0,997), indicando um bom ajuste aos dados de treinamento, porém um RMSE relativamente alto (1.177.379,51), sugerindo a possibilidade de overfitting e a necessidade de ajustes para melhorar a generalização.*

1. Introdução

1.1. Contextualização

O mercado de jogos digitais tem se mostrado um campo fértil para a aplicação de técnicas de análise de dados e inteligência artificial. A capacidade de prever métricas-chave, como receita, avaliações e preço de jogos, pode trazer insights valiosos para desenvolvedores e publishers, auxiliando-os na tomada de decisões estratégicas.

Neste projeto, o objetivo é realizar uma análise preditiva utilizando um dataset de 1500 jogos da plataforma Steam. A escolha desse dataset se justifica pela relevância e disponibilidade de informações detalhadas sobre o mercado de jogos digitais. O estudo e previsão dessas métricas relevantes para o sucesso de um jogo podem contribuir significativamente para o setor.

1.2. Justificativa

O estudo e a previsão de métricas relevantes para o sucesso de um jogo, como receita, avaliações e preço, podem trazer insights valiosos para o setor.

1.3. Objetivo

O objetivo deste projeto é realizar uma análise preditiva utilizando um dataset de jogos da plataforma Steam, buscando encontrar padrões e prever informações importantes para o sucesso de um jogo.

1.4. Opção do projeto

A escolha deste dataset de jogos da Steam se justifica pela relevância e disponibilidade de informações detalhadas sobre o mercado de jogos digitais.

2. Descrição do Problema

O principal problema a ser abordado neste projeto é a identificação de fatores-chave que influenciam a receita, as avaliações e o preço dos jogos na plataforma Steam. Além disso, pretende-se desenvolver modelos preditivos capazes de estimar essas métricas com base nas características dos jogos.

3. Dataset

O conjunto de dados utilizado neste projeto contém informações sobre 1500 jogos da Steam, coletadas do Kaggle [Topcu 2024]. As variáveis incluídas na análise são:

- name (Nome): Nome do jogo.
- revenue (Receita): Receita total gerada pelo jogo (variável alvo).
- price (Preço): Preço do jogo em dólares.
- copiesSold (Cópias Vendidas): Número de cópias vendidas.
- avgPlaytime (Tempo Médio de Jogo): Tempo médio de jogo em minutos.
- reviewScore (Avaliação): Pontuação média das avaliações dos usuários.
- publisherClass (Classe da Editora): Classificação da editora (AAA, AA, Indie, etc.).
- Release Year (Ano de Lançamento): Ano de lançamento do jogo.

3.1. Pré-processamento dos Dados

Antes do treinamento do modelo, os dados foram pré-processados para lidar com valores faltantes e garantir a compatibilidade com o algoritmo. As seguintes etapas foram realizadas:

- Tratamento de valores faltantes: Valores faltantes em variáveis numéricas foram imputados usando a mediana, enquanto valores faltantes em ‘publisherClass‘ foram imputados usando a moda.
- Conversão de data: A data de lançamento (‘releaseDate‘) foi convertida para o ano de lançamento (‘Release Year‘).
- One-Hot Encoding: A variável categórica ‘publisherClass‘ foi convertida para uma representação numérica usando one-hot encoding.
- Padronização: As variáveis numéricas foram padronizadas usando o ‘StandardScaler‘ para que tivessem média zero e desvio padrão igual a um.

4. Metodologia

Para a previsão da receita dos jogos, foi utilizado o algoritmo Random Forest Regressor, um método de aprendizado de máquina ensemble que combina múltiplas árvores de decisão. A escolha deste algoritmo se justifica por sua capacidade de lidar com dados não lineares e pela sua robustez a outliers.

O conjunto de dados foi dividido em conjuntos de treinamento (80%) e teste (20%) usando o train_test_split do scikit-learn, com random_state=42 para garantir a reprodutibilidade dos experimentos.

Um pipeline foi utilizado para encadear as etapas de pré-processamento e o treinamento do modelo, facilitando a aplicação das transformações aos dados de forma consistente.

Tabela 1. Exemplos de Previsões - Treino

Nome do Jogo	Receita Real	Receita Prevista
Starstruck Vagabond	162429.0	160572.52
Shipwrecked 64	202048.0	185827.81
hololive Treasure Mountain	291300.0	275343.24
Magical Delicacy	66573.0	61196.47
The Casting of Frank Stone™	1967699.0	2054725.86
Yaoling: Mythical Journey	1047362.0	1033506.56
Beneath the Mountain	102217.0	104950.20
Taora : Survival	78871.0	78892.21
GUNDAM BREAKER 4	8440898.0	7754458.83
Balatro	20479210.0	20689501.51

5. Resultados

Após o treinamento e avaliação do modelo no conjunto de teste, os seguintes resultados foram obtidos:

- RMSE: 1.177.379,51
- R-squared: 0.997

O R-squared de 0.997 indica que o modelo explica uma grande parte da variância na receita dos jogos no conjunto de *treinamento*. No entanto, o RMSE relativamente alto sugere que o modelo pode estar sofrendo de *overfitting*, ou seja, se ajustando muito bem aos dados de treinamento, mas não generalizando bem para novos dados. É importante notar a necessidade de uma análise mais aprofundada do desempenho em um conjunto de validação para confirmar a presença de overfitting.

5.1. Exemplos de Previsões

As tabelas a seguir mostram exemplos de previsões do modelo para os conjuntos de treinamento e teste. A coluna ”Real”representa a receita real do jogo, enquanto a coluna ”Previsto”representa a receita prevista pelo modelo.

5.2. Gráficos de Dispersão

Os gráficos de dispersão abaixo comparam os valores reais e previstos da receita para os conjuntos de treinamento e teste. A linha tracejada representa a situação ideal onde as previsões são perfeitas.

6. Conclusão

Este projeto demonstrou a aplicação de técnicas de aprendizado de máquina para prever a receita de jogos na Steam. O modelo Random Forest obteve um alto R-squared no conjunto de treinamento, mas o RMSE e a análise dos gráficos de dispersão sugerem a possibilidade de overfitting. Futuros trabalhos podem explorar técnicas para mitigar o overfitting, como regularização, aumento do conjunto de dados ou a utilização de modelos mais robustos à generalização. A exploração de outras variáveis e a engenharia de novas features também podem contribuir para a melhoria do desempenho preditivo.

Tabela 2. Exemplos de Previsões - Teste

Nome do Jogo	Receita Real	Receita Prevista
Time to Morp	47790.0	4.745999e+04
Real Dive World	36294.0	3.823038e+04
Immortal Family	327846.0	3.817441e+05
The Settlers: New Allies	498522.0	9.245065e+05
KinitoPET	1166563.0	1.392249e+06
Hospital 666	596521.0	6.436494e+05
Moonbreaker	1367965.0	1.533254e+06
Fate/stay night REMASTERED	764069.0	7.475598e+05
Stigma-ARIA	22306.0	2.340739e+04
Terra Randoma	46626.0	4.956908e+04

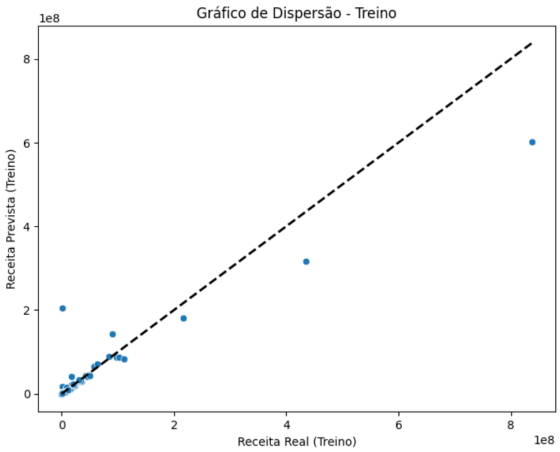

Figura 1. Gráfico de Dispersão - Treino

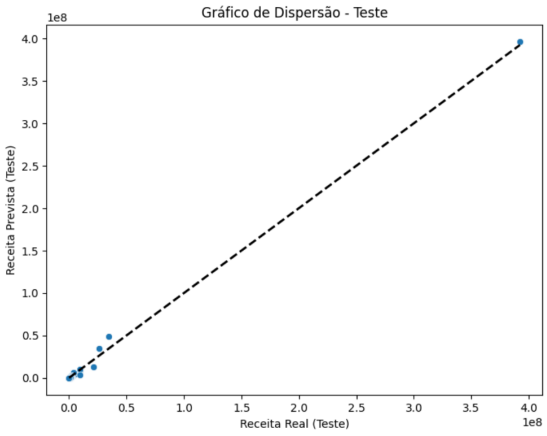

Figura 2. Gráfico de Dispersão - Teste

7. Bibliografia

O dataset foi obtido em [Topcu 2024]. Os slides de ”Preparação e Pré-processamento dos dados”utilizados como base de estudo foram ministrados pelo Prof. Dr. Ivan Carlos Alcântara de Oliveira, sendo disponibilizados nas seguintes partes: [de Oliveira 2024a], [de Oliveira 2024b] e [de Oliveira 2024c].

Referências

de Oliveira, P. D. I. C. A. (2024a). Preparação e pré-processamento dos dados parte i. Slide de aula. Acesso em: 2024-09-21.

de Oliveira, P. D. I. C. A. (2024b). Preparação e pré-processamento dos dados parte ii. Slide de aula. Acesso em: 2024-09-21.

de Oliveira, P. D. I. C. A. (2024c). Preparação e pré-processamento dos dados parte iii. Slide de aula. Acesso em: 2024-09-21.

Topcu, A. C. (2024). Top 1500 games on steam by revenue 09-09-2024. https://www.kaggle.com/datasets/alicemtopcu/top-1500-games-on-steam-by-revenue-09-09-2024. Acesso em: 2024-09-21.