FATEC RUBENS LARA

CIÊNCIA DE DADOS – 4° CICLO

Mapeamento do Futuro - Analisando as Tendências Globais dos Jogos

Pedro Jorge de Souza Colombrino

Matheus Ferreira Amaral Madeira

Guilherme Vieira Rodrigues

Santos - SP

2024

Pedro Jorge de Souza Colombrino

Matheus Ferreira Amaral Madeira

Guilherme Vieira Rodrigues

Mapeamento do Futuro - Analisando as Tendências Globais dos Jogos

Atividade apresentada ao curso de

Ciência de Dados como

requisito para a obtenção de nota

Santos - SP

2024

1. INTRODUÇÃO

Desde os primórdios dos videogames, quando simples pontos luminosos em telas monocromáticas simulavam jogos de tênis, a indústria de games passou por uma evolução vertiginosa. Dos *arcades* dos anos 80, com seus clássicos inesquecíveis, até os consoles de última geração e a era dos jogos online, a jornada tem sido marcada por inovações tecnológicas e mudanças nos hábitos dos jogadores.

Com o avanço da tecnologia, os jogos se tornaram cada vez mais complexos e imersivos, envolvendo milhões de jogadores em todo o mundo. Paralelamente a esse crescimento, as empresas do setor perceberam a necessidade de compreender melhor seu público e as tendências do mercado. A coleta e análise de dados se tornaram essenciais para tomar decisões estratégicas, desenvolver novos produtos e personalizar experiências.

Neste contexto, o presente estudo propõe utilizar a metodologia *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*) para mapear o futuro dos jogos, analisando as tendências globais do setor. Através de uma abordagem estruturada e sistemática, buscamos identificar padrões, prever comportamentos e gerar insights valiosos para empresas, desenvolvedores e jogadores.

2. CRISP-DM

2.1. Entendimento do negócio

A análise e o acompanhamento de tendências são práticas consolidadas no ambiente corporativo. No competitivo mercado de games, a compreensão das preferências do público é essencial para tomar decisões estratégicas. Com o objetivo de identificar oportunidades de investimento, este projeto busca analisar dados de mercado, buscando insights sobre as tendências emergentes.

Seguindo os princípios do CRISP-DM, que garantem a atualização contínua dos dados, visamos desenvolver um projeto escalável e aprimorável, capaz de gerar valor a longo prazo.

2.2. Entendimento dos Dados

A fim de obter insights sobre o mercado de games, realizaremos uma análise exploratória dos dados disponíveis. Utilizando a biblioteca Pandas do Python, investigaremos as variáveis presentes no dataset vgsales.csv e seus respectivos tipos. Para facilitar a compreensão, geraremos um dicionário de dados completo com o auxílio da biblioteca Python-Docx.

Essa etapa é fundamental para identificar as métricas de sucesso mais adequadas para avaliar o desempenho dos jogos e, consequentemente, direcionar nossas análises.

Variável	Tipo	Descrição
Rank	int64	Posição do jogo em um ranking geral.
Name	object	Nome do jogo.
Platform	object	Plataforma em que o jogo foi lançado (neste caso, 2600).
Year	int64	Ano de lançamento.
Genre	object	Gênero do jogo.
Publisher	object	Editora do jogo.
NA_Sales	float64	Vendas na América do Norte.
EU_Sales	float64	Vendas na Europa.
JP_Sales	float64	Vendas no Japão.
Other_Sales	float64	Vendas em outras regiões.
Global_Sales	float64	Vendas globais

A variável 'Rank' representa a classificação de um jogo em relação às suas vendas globais, que são obtidas somando as vendas em todas as regiões registradas no dataset.

As variáveis 'Plataform', 'Genre' e 'Publisher' oferecem um perfil detalhado de cada jogo, permitindo analisar as preferências dos distribuidores em relação a plataformas e gêneros específicos. Essa análise pode revelar padrões e tendências importantes no mercado de games.

2.3. Preparação de dados

Para garantir a confiabilidade de nossas análises, utilizamos a biblioteca Pandas para identificar as colunas que contêm valores ausentes ou nulos em nosso conjunto de dados.

Essa etapa é crucial para realizar o tratamento adequado dos dados, eliminando ou imputando os valores faltantes e evitando que comprometam os resultados da análise.

2.4. Valores Nulos por Coluna

Coluna	Quantidade de Nulos
Rank	0
Name	0
Platform	0
Year	0
Genre	0
Publisher	36
NA_Sales	0
EU_Sales	0
JP_Sales	0
Other_Sales	0
Global_Sales	0

Durante a análise exploratória dos dados, constatamos a presença de 36 valores nulos em nosso dataframe. Para tratar esses valores ausentes, aplicamos os algoritmos de tratamento de dados implementados no arquivo main.ipynb. Na sequência, utilizamos a técnica do z-score para identificar os outliers presentes no conjunto de dados. Após uma análise detalhada dos outliers detectados, concluímos que eles representam fenômenos reais e não erros de coleta de dados.

Dessa forma, optamos por manter esses valores na análise, uma vez que a remoção poderia levar à perda de informações importantes e comprometer a precisão do modelo.

2.5. Outliers

Ao analisar os dados, detectamos outliers por meio de gráficos de boxplot e cálculo do Z-score. Esses valores atípicos foram isolados no arquivo 'outliers.csv' para investigação. Após análise cuidadosa, constatamos que esses dados correspondem a eventos específicos do negócio, como promoções ou lançamentos, que podem ser cruciais para entender o comportamento dos dados.

Decidimos, portanto, preservá-los na análise, pois sua exclusão poderia levar à perda de insights importantes e distorcer os resultados.

2.6. Clusterização

A fim de identificar grupos de jogos com características similares, realizamos uma análise de clusterização. Utilizando o método do cotovelo, determinamos que o número ideal de clusters para nossos dados era dois.

Em seguida, aplicamos o algoritmo k-means ao dataset 'vgsales_limpo.csv', gerando o arquivo 'vgsales_com_clusters.csv'. Para visualizar e analisar os resultados, utilizaremos gráficos de dispersão, os quais nos permitirão identificar quais clusters concentram os jogos de maior sucesso e quais características os diferenciam.

2.7. Modelagem:

Iniciamos com um processo de clusterização para entender melhor a estrutura dos nossos dados. Para determinar o número ideal de clusters, utilizamos o método do cotovelo. Este método foi implementado através de um algoritmo em Python, que nos indicou que o uso de 2 clusters seria apropriado para nosso dataset. Com base nessa informação, aplicamos o algoritmo k-means ao dataframe limpo (vgsales_limpo.csv), o que resultou em um novo arquivo CSV chamado 'vgsales_com_clusters.csv', onde podemos identificar as duas categorias criadas, rotuladas como 0 e 1.

Após a clusterização, prosseguimos com a modelagem preditiva. Dividimos os dados em conjuntos de treino e teste, e aplicamos a normalização usando StandardScaler. Implementamos dois modelos distintos: uma Rede Neural Artificial (RNA) e uma Árvore de Decisão. Para otimizar o desempenho, utilizamos GridSearchCV para ajustar os hiper parâmetros de ambos os modelos.

2.8. Avaliação:

A avaliação dos modelos foi realizada calculando métricas de regressão como o Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2). Comparamos o desempenho dos modelos e identificamos o melhor mercado com base nas previsões. Além disso, criamos uma visualização gráfica para comparar as vendas reais com as previsões de ambos os modelos.

3.0. Conclusão:

Os resultados obtidos mostraram que o modelo [RNA/Árvore de Decisão] apresentou melhor desempenho geral, com um R^2 de [valor], indicando que [X]% da variabilidade nas vendas é explicada por este modelo. O erro médio absoluto (MAE) do melhor modelo foi de [valor] milhões de unidades vendidas, o que nos dá uma ideia da precisão das previsões. Identificamos também o mercado mais promissor com base nas previsões de cada modelo.

Com base nesses resultados, recomendamos que a empresa concentre seus esforços de marketing e distribuição no mercado identificado como mais promissor. Sugerimos continuar refinando o modelo que mostrou melhor desempenho, possivelmente incluindo variáveis adicionais para melhorar a precisão das previsões com o objetivo do planejamento estratégico, ajustando a produção, estoque, preços e promoções de jogos.

Recomendamos também a implementação de um sistema de monitoramento contínuo para comparar as previsões do modelo com as vendas reais, ajustando o modelo periodicamente para manter sua precisão.

Por fim, sugerimos considerar a expansão desta metodologia de análise para outros gêneros de jogos e investigar possíveis correlações entre diferentes gêneros e mercados para uma estratégia de negócios mais abrangente.