Relatório Estatístico do Desafio de Data Science

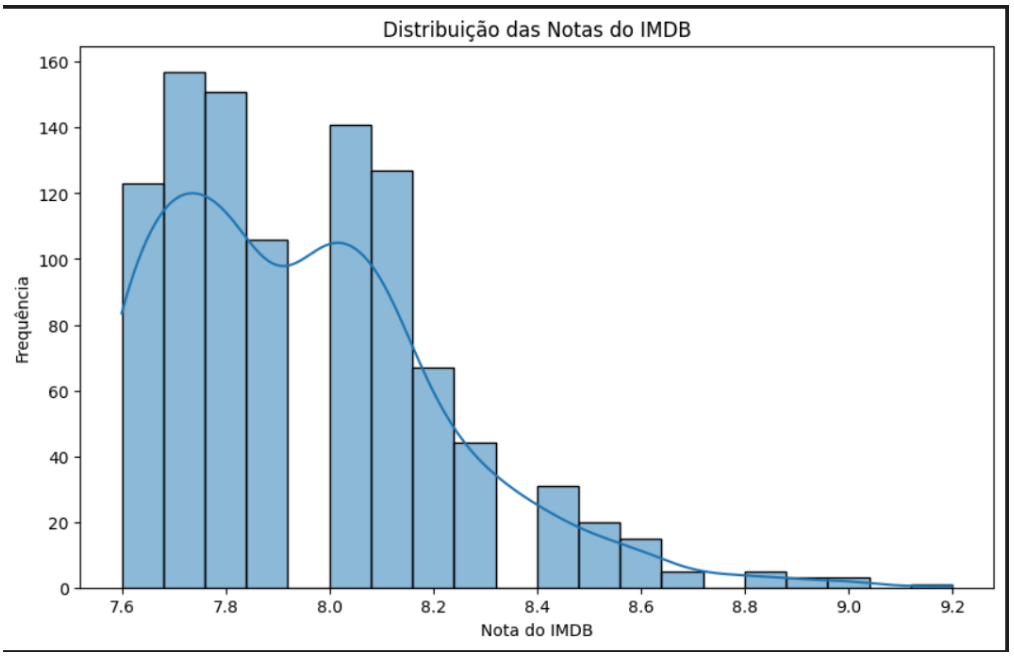
Introdução

Este relatório apresenta uma análise estatística detalhada do dataset de filmes, conforme solicitado no desafio de Data Scientist da Indicium para a PProductions. Foram utilizados diversos modelos de Machine Learning, como regressão linear e árvores de decisão, e métricas como precisão e revocação foram calculadas a partir de uma matriz de confusão.

1. Análise Exploratória dos Dados

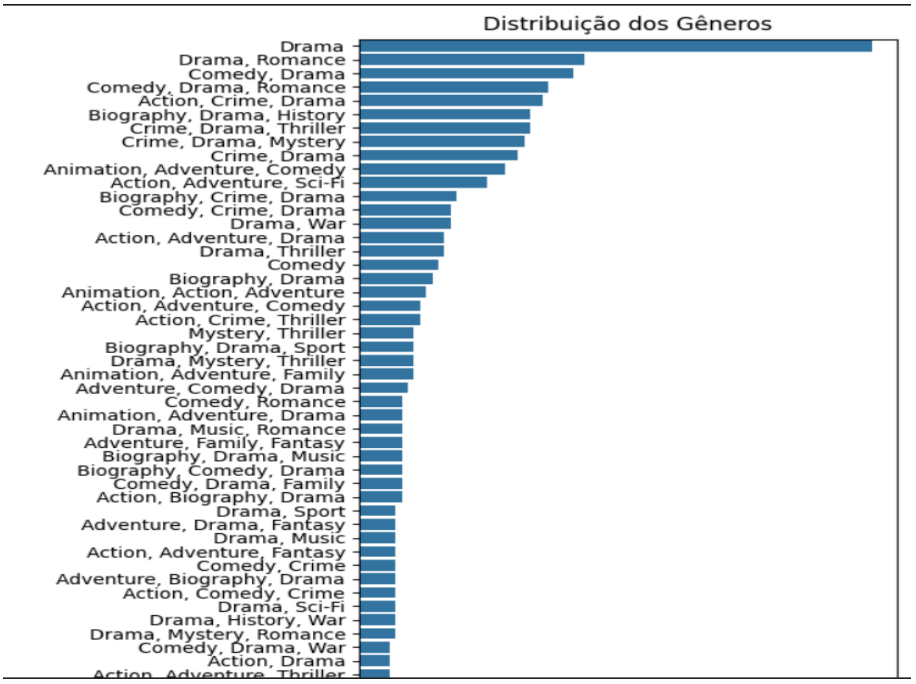
1.1 Distribuição das Notas do IMDB

Conclusão: A maioria dos filmes possui notas entre 7.6 e 8.2, com um pico próximo de 7.8. A distribuição é levemente assimétrica à direita, indicando que há mais filmes com notas menores que 8.0 do que maiores.



1.2 Distribuição dos Gêneros

Conclusão: O gênero "Drama" é o mais comum entre os filmes analisados, seguido de combinações como "Drama, Romance" e "Comedy, Drama". Gêneros como "Adventure, Family, Fantasy" são menos frequentes.



1.3 Correlação entre variáveis

O mapa de calor mostra as correlações entre diferentes variáveis. A variável 'No_of_Votes' tem a maior correlação positiva com 'Gross' (0.62), indicando que um maior número de votos está associado a um maior faturamento.
Outras variáveis, como 'Runtime' (0.14) e 'IMDB_Rating' (0.085), têm correlações fracas com 'Gross'. 'Meta_score' tem uma correlação negativa fraca com 'Gross' (-0.052).

2. Resposta às Perguntas

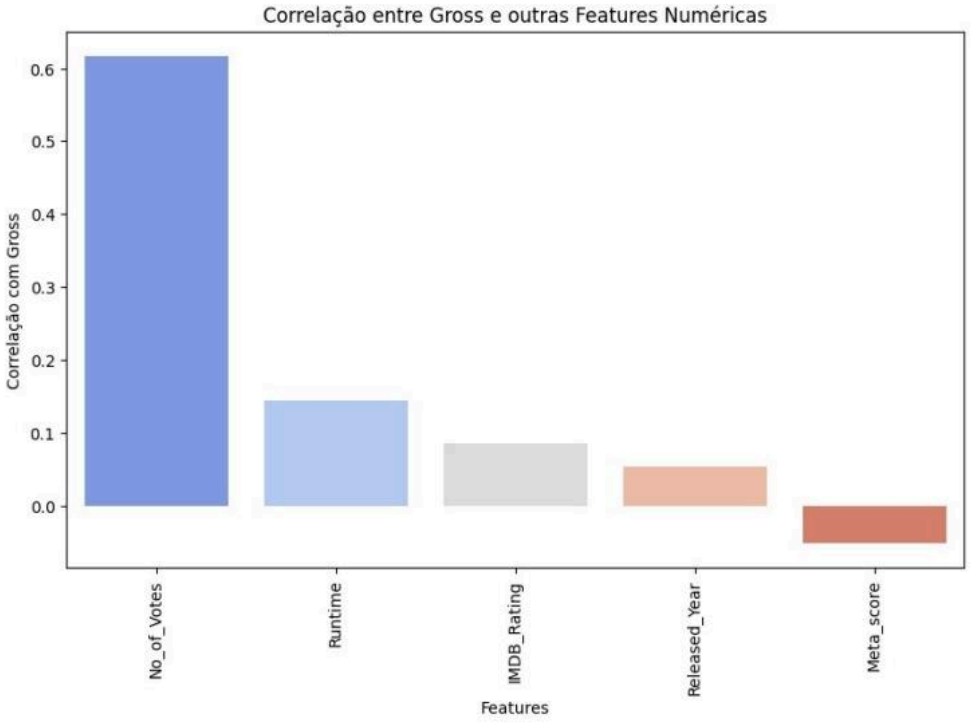
2.1 Qual filme você recomendaria para uma pessoa que você não conhece?

Recomendação baseada na maior nota do IMDB: **Recomendação: The Godfather**

2.2 Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?

Os principais fatores relacionados ao faturamento (Gross) são:

- Número de Votos (No_of_Votes)
- Tempo de Duração (Runtime)
- Nota do IMDB (IMDB_Rating)



- Este gráfico confirma que a variável 'No_of_Votes' tem a correlação mais forte com 'Gross'.
- 'Runtime' tem uma correlação moderada com 'Gross'.
- 'IMDB_Rating', 'Released_Year' e 'Meta_score' têm correlações fracas com 'Gross'.

Esses insights são importantes para entender quais fatores podem influenciar o faturamento de um filme.

- No_of_Votes (Correlação: 0.616440): Tem a maior correlação positiva com o faturamento. Isso sugere que filmes que recebem mais votos tendem a ter um faturamento maior. Isso pode indicar que a popularidade de um filme (medida pelo número de votos) é um fator importante para determinar seu sucesso financeiro. Runtime (Correlação: 0.144242): O tempo de duração (Runtime) também apresenta uma correlação positiva com o faturamento, embora menor que No_of_Votes. Filmes mais longos podem ter uma maior profundidade de enredo ou produção, o que pode atrair mais espectadores e, consequentemente, gerar mais receita.
- IMDB_Rating (Correlação: 0.084732): A nota do IMDB tem uma correlação positiva, mas relativamente fraca com o faturamento. Isso sugere que a qualidade percebida do filme (medida pela nota do IMDB) pode ter algum impacto no faturamento, mas não é um fator tão determinante quanto o número de votos.
- Released_Year (Correlação: 0.053068): O ano de lançamento tem uma correlação muito fraca com o faturamento, indicando que filmes mais recentes ou mais antigos não têm uma diferença significativa em termos de receita.
- Meta_score (Correlação: -0.052202): A média ponderada das críticas (Meta_score) apresenta uma correlação negativa, ainda que muito fraca, com o faturamento. Isso sugere que críticas melhores não necessariamente se traduzem em maior faturamento e que outros fatores podem ser mais importantes para determinar o sucesso financeiro de um filme.

2.3 Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?

A coluna "Overview" fornece uma breve descrição do enredo do filme. Estas descrições contêm informações sobre o tema, a trama e os personagens do filme, o que pode ser útil para entender o conteúdo do filme de uma forma resumida.

Possível Inferência do Gênero:

A tentativa de inferir o gênero do filme a partir da coluna "Overview" utilizando técnicas de processamento de linguagem natural (PLN) e um modelo de classificação mostra uma baixa acurácia (aproximadamente 9,5%). Isso indica que, com o método e os dados utilizados, não é possível inferir com precisão o gênero do filme apenas a partir das descrições fornecidas na coluna "Overview". A partir desta análise, a conclusão é que a coluna "Overview" contém informações sobre a trama dos filmes, mas não é suficientemente discriminativa para inferir com precisão o gênero dos filmes utilizando o modelo e a abordagem apresentados. A baixa acurácia indica que ou a abordagem utilizada (regressão logística com vetorização TF-IDF) não é adequada para este problema, ou que as descrições dos filmes são muito gerais e não contêm informações específicas o suficiente para identificar o gênero.

2.4 Explique como você faria a previsão da nota do imdb a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê?

As variaveis:

- Released_Year,
- Runtime,
- Meta_score,
- No_of_Votes
- Gross

Foram escolhidas por serem potencialmente influentes na avaliação de um filme. O faturamento e o número de votos podem indicar popularidade, enquanto o tempo de execução e a média das críticas fornecem uma noção da qualidade percebida do filme.

Preparação dos Dados

1 Preenchimento de Valores Nulos: Qualquer valor nulo nas variáveis selecionadas foi preenchido com 0;

2 Conversão de Tipos: Garanti que as variáveis 'Runtime' e 'Gross' estivessem no formato correto (inteiro e float, respectivamente);

3 Separação dos Dados: Os dados foram divididos em conjuntos de treino e teste, com 80% dos dados sendo utilizados para treino e 20% para teste. Isso permite avaliar o desempenho do modelo em dados não vistos;

4 Treinamento do Modelo: Usei um modelo de Random Forest Regressor para prever a nota do IMDB. Este modelo é uma escolha sólida devido à sua capacidade de lidar com dados complexos e não lineares, e por ser robusto a overfitting.

5 Avaliação do Modelo: A performance do modelo foi avaliada usando o Mean Squared Error (MSE), que mede a média dos quadrados dos erros, ou seja, a diferença média quadrática entre os valores previstos e os reais.

2.5 Qual tipo de problema estamos resolvendo (regressão, classificação)?

Estamos resolvendo um problema de regressão.

2.6 Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

Modelo de Árvore de Decisão para Classificação

Performance

- Acurácia: 0.31
- Este modelo classifica as avaliações do IMDb em classes discretas, o que é útil se você estiver interessado em prever categorias específicas de classificação, como faixas de avaliação (por exemplo, ruim, mediano, bom, excelente).
- Prós: Pode ser interpretável dependendo da profundidade da árvore, captura relações não lineares entre as features e a variável alvo.
- Contras: Pode ser propenso a overfitting se não for regularizado adequadamente ou se a profundidade da árvore for muito grande.

O modelo de Árvore de Decisão parece ser mais apropriado, pois obteve uma acurácia de 0.31, indicando que ele é capaz de classificar as avaliações do IMDb com uma taxa razoável de precisão.

3 Supondo um filme com as seguintes características:

{'Series_Title': 'The Shawshank Redemption',
'Released_Year': '1994',
'Certificate': 'A',
'Runtime': '142 min',
'Genre': 'Drama',
'Overview': 'Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency.',
'Meta_score': 80.0,
'Director': 'Frank Darabont',
'Star1': 'Tim Robbins',
'Star2': 'Morgan Freeman',
'Star3': 'Bob Gunton',
'Star4': 'William Sadler',
'No_of_Votes': 2343110,
'Gross': '28,341,469'}

Qual seria a nota do IMDB?

Predicted IMDB Rating: 8.759

Letícia Oliveira Gobbi