INDICIUM

Débora Nunes Ferreira

Desafio: **Ciência de Dados**

Análise Cinematográfica para PProductions

Brasília, DF
2024

Introdução

Este projeto de análise de dados foi desenvolvido como parte de um processo seletivo da Indicium, uma renomada empresa especializada em tecnologia de dados que oferece soluções inovadoras para transformar dados em insights estratégicos. O objetivo principal é avaliar minha capacidade na resolução de problemas de negócio, análise de dados avançada e aplicação de modelos preditivos.

Neste desafio, fui designada para integrar a equipe da Indicium, contratada pelo estúdio de cinema PProduction, sediado em Hollywood. Meu papel consiste em realizar uma análise detalhada de um extenso banco de dados cinematográfico. A meta é orientar o estúdio na seleção do próximo filme a ser desenvolvido, considerando aspectos como a distribuição dos filmes no mercado, a análise das notas no IMDb, a categorização por gêneros e a resposta a outras questões estratégicas que serão exploradas adiante. Posteriormente, discutiremos a criação de um modelo preditivo desenvolvido para prever as notas no IMDb, contribuindo assim para decisões mais fundamentadas e eficazes no campo da produção cinematográfica.

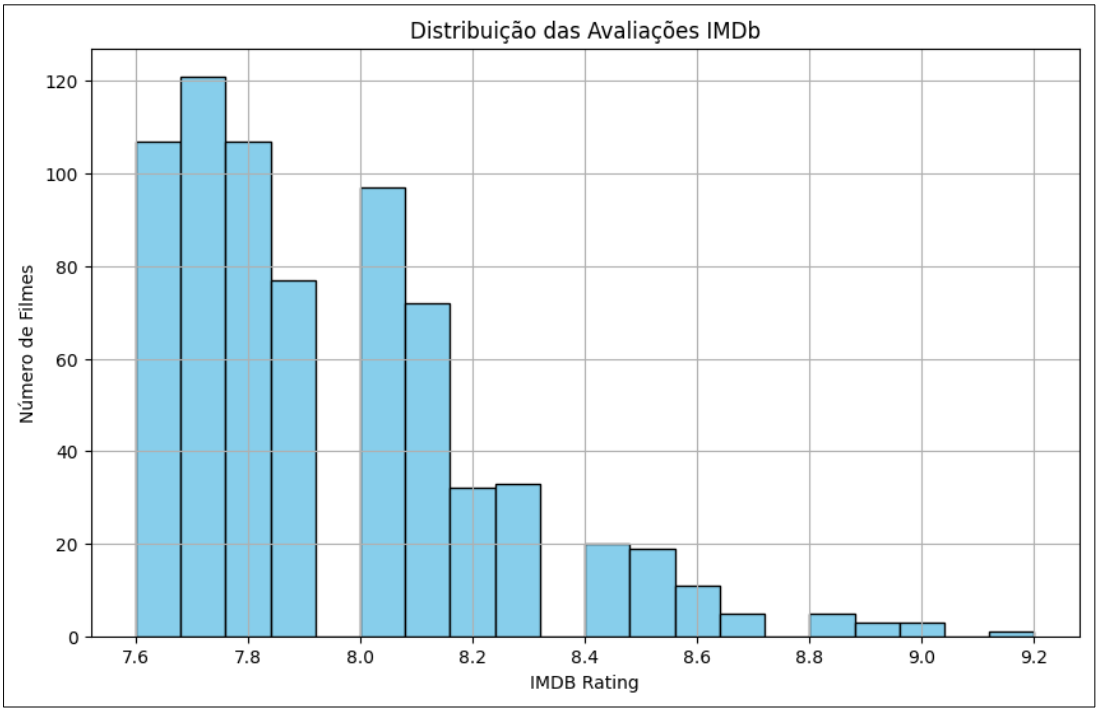
ESTÚDIO PPRODUCTIONS

No início do projeto, configuramos as bibliotecas essenciais para manipulação de dados, cálculos matemáticos, estatísticas e visualizações gráficas. Entre essas bibliotecas estão Pandas, NumPy, Seaborn, entre outras disponíveis no repositório GitHub para mais detalhes sobre o código. Após a importação das bibliotecas, procedemos com a coleta dos arquivos CSV e uma análise inicial dos dados para entender a estrutura do conjunto de dados, que possui 15 colunas e 999 linhas. Utilizando a função info(), identificamos colunas com dados faltantes e iniciamos o processo de limpeza, removendo duplicatas, nulos e tratando a coluna 'Gross' para converter seus valores de string para inteiro, removendo aspas e vírgulas. Essa preparação permite uma análise mais precisa e a criação de gráficos relevantes. Além disso, aplicamos o método describe() para obter estatísticas básicas das colunas numéricas.

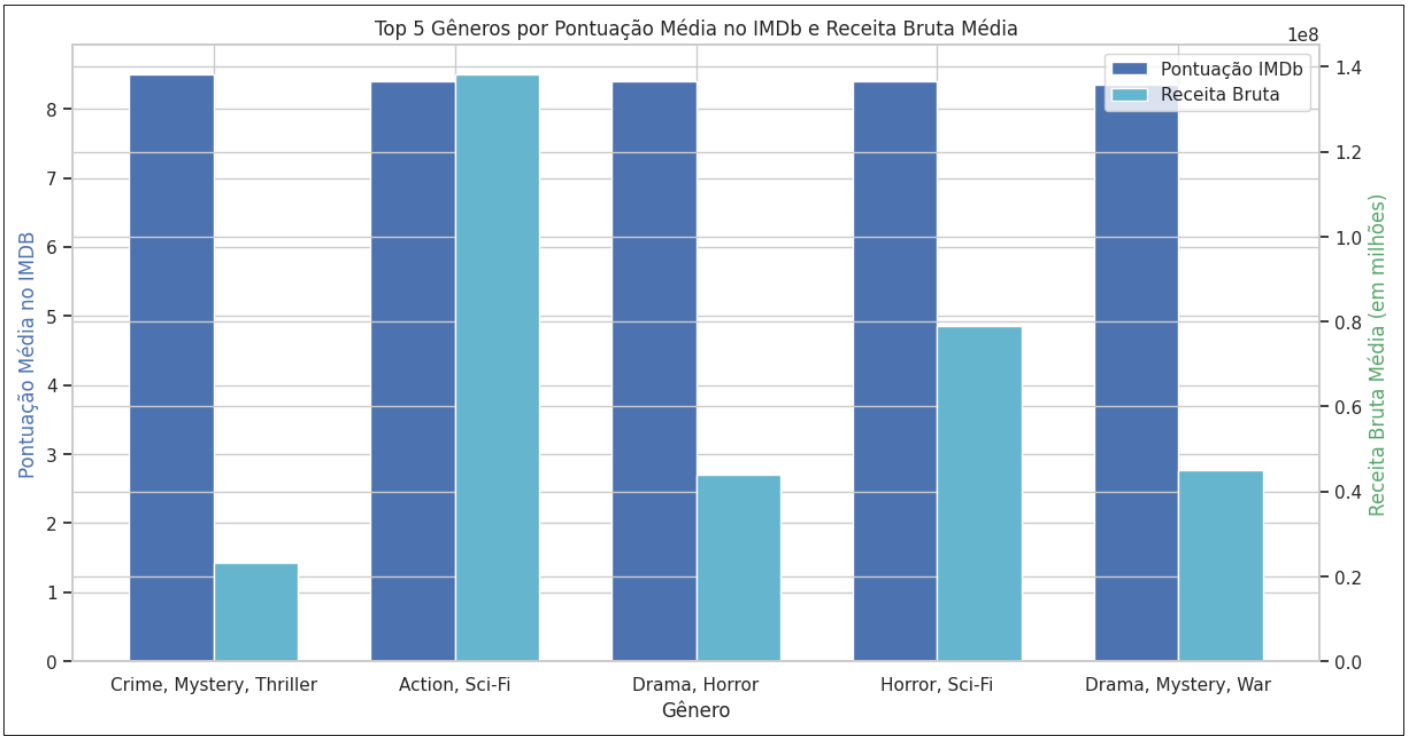
```
# Estatísticas Básicas
print(imdb.describe().T)
```

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	713.0	5.193001e+02	2.954163e+02	1.0	263.0	527.0	778.0	997.0
IMDB_Rating	713.0	7.935203e+00	2.889986e-01	7.6	7.7	7.9	8.1	9.2
Meta_score	713.0	7.715428e+01	1.240939e+01	28.0	70.0	78.0	86.0	100.0
No_of_Votes	713.0	3.533480e+05	3.462212e+05	25229.0	95826.0	236311.0	505918.0	2303232.0
Gross	713.0	7.858395e+07	1.150433e+08	1305.0	6153939.0	35000000.0	102515793.0	936662225.0

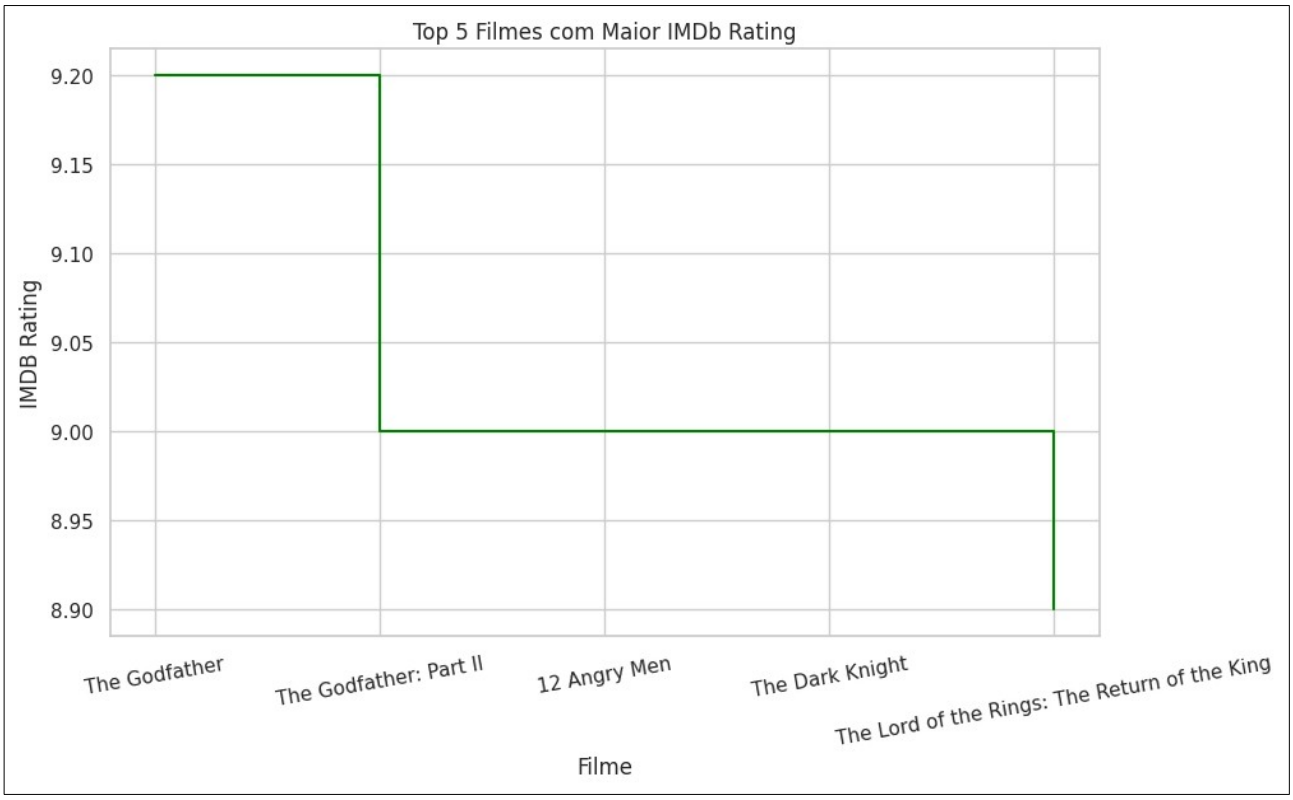
Realizei uma análise exploratória criando um histograma para investigar a distribuição das avaliações no IMDb. No eixo x, representamos o número de filmes, enquanto no eixo y, estão as notas no IMDb. Essa visualização nos proporciona uma compreensão mais aprofundada do conjunto de dados, auxiliando na consecução dos nossos objetivos propostos. Veja a imagem a seguir para mais detalhes:

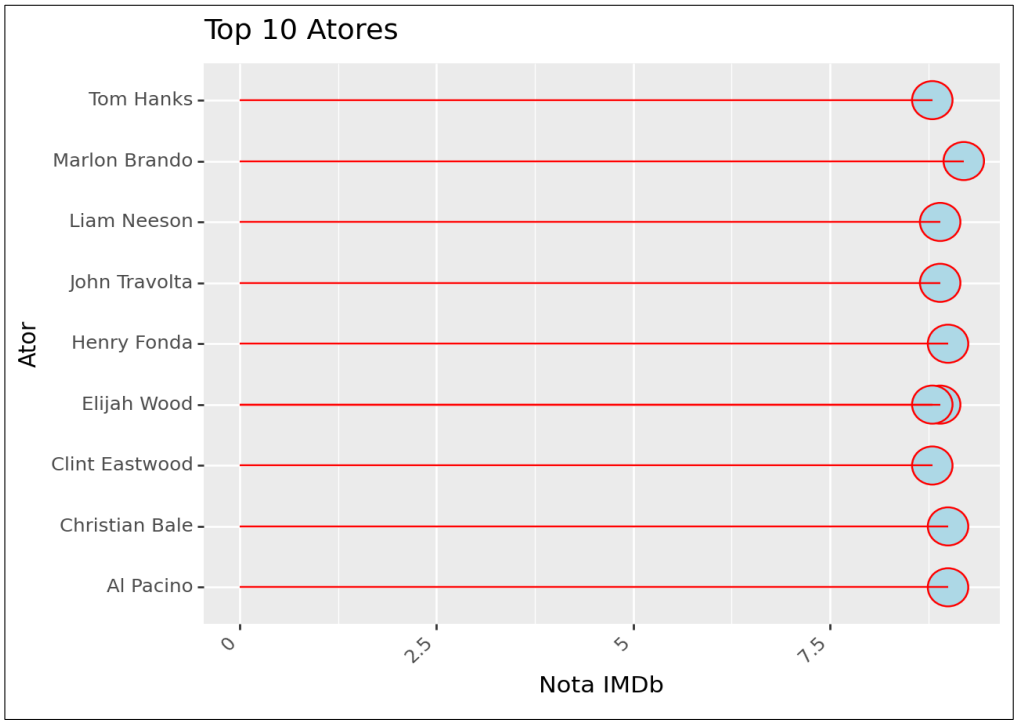


Então, avancei para responder perguntas que irão esclarecer minhas dúvidas e ajudar na tomada de decisão sobre qual será o próximo filme a ser desenvolvido. No gráfico abaixo, exploramos a seguinte questão: "Quais são os 5 gêneros de filmes com a maior média de pontuação no IMDb e qual é a média de faturamento?". Podemos observar que o gênero com maior faturamento e pontuação é 'Action, Sci-fi', seguido por 'Horror, Sci-Fi', 'Drama, Mystery, War', 'Drama, Horror' e 'Crime, Mystery, Thriller'.

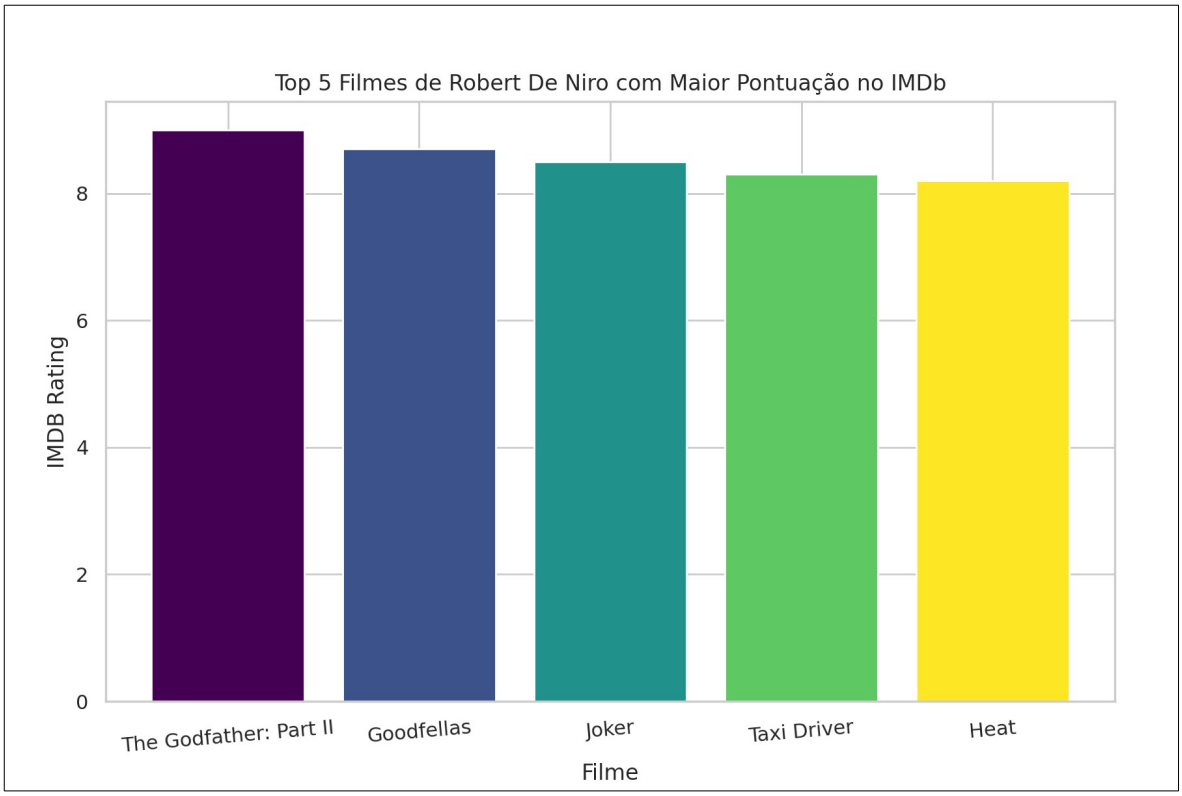


Em seguida, abordei as seguintes questões: "Quais são os filmes com maior nota do IMDb?" e "Quais são os atores em filmes com maior nota no IMDb?"

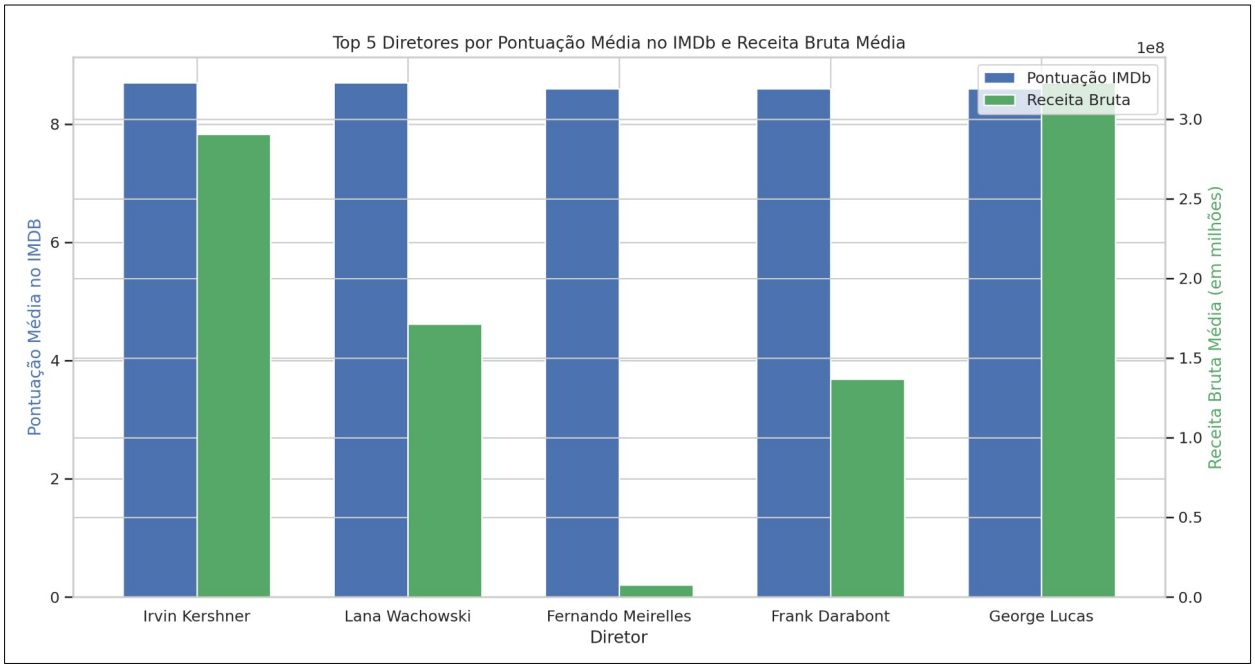




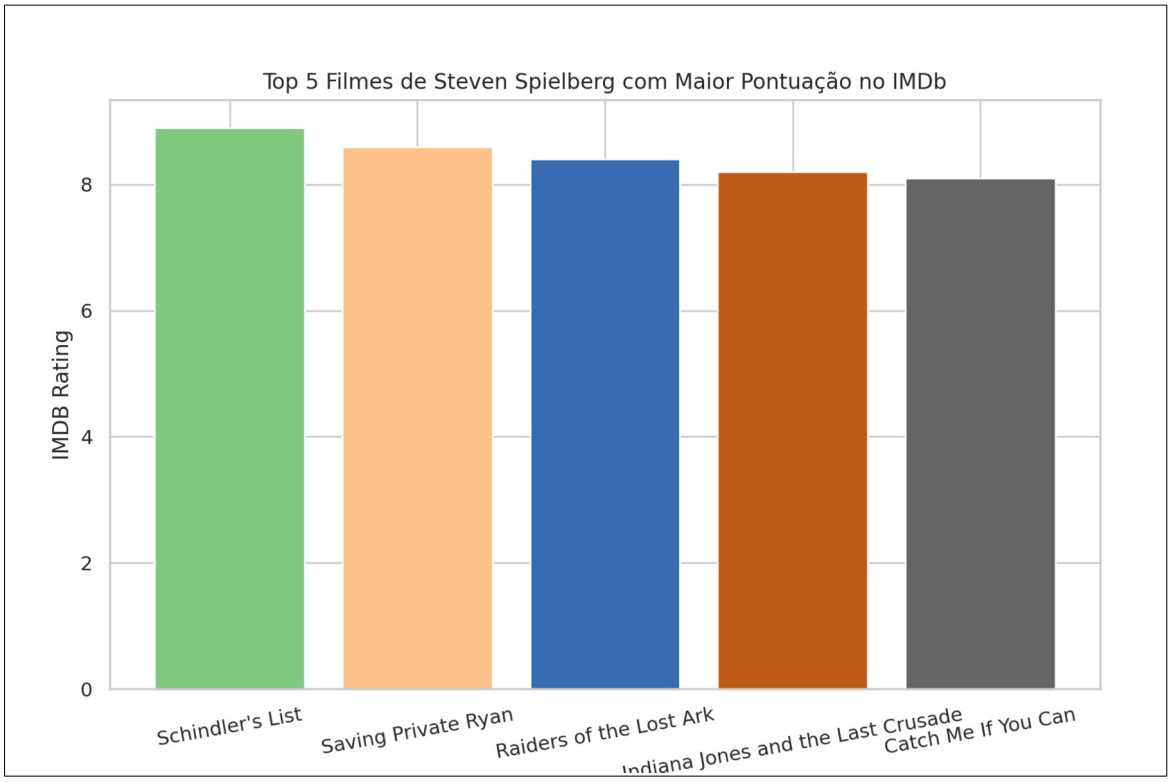
Logo após analisar os atores, busquei a informação sobre qual ator tinha protagonizado mais filmes. Assim, conseguimos escolher um ator não só com boa nota, mas também com experiência para protagonizar o próximo filme. Robert De Niro foi o ator que mais protagonizou filmes, com um total de 16. Em seguida, investiguei os cinco filmes com a maior pontuação no IMDb para entender um pouco mais sobre sua carreira.



Analisando possíveis atores, dediquei-me à análise dos diretores. Elaborei um gráfico que apresenta a pontuação média no IMDb e a receita bruta média (em milhões). George Lucas foi o diretor com a maior pontuação e faturamento, seguido por Irvin Kershner, Lana Wachowski, Frank Darabont e Fernando Meirelles.



Busquei o diretor com o maior número de filmes produzidos, que foi Steven Spielberg com um total de 13 filmes. A média de pontuação em seus filmes dirigidos é 8.03. Em seguida, desenvolvi um gráfico para verificar quais filmes têm maior pontuação.



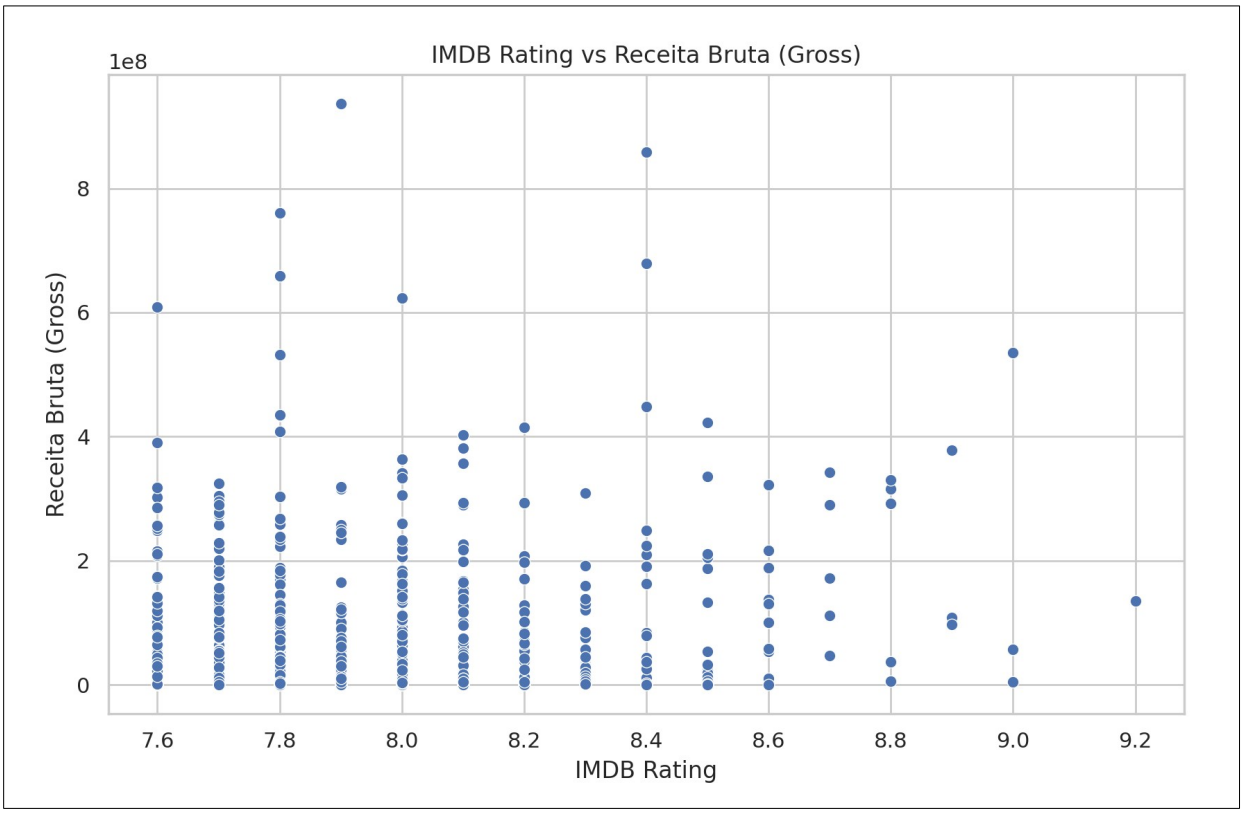
Perguntas de Negócio:

a) Qual filme recomendaria para uma pessoa que você não conhece?

Primeiramente, eu consideraria a classificação etária para recomendar o filme adequado. Para adultos, indicaria "The Godfather", um clássico do cinema aclamado pela crítica e pelo público. Para adolescentes, sugiro "Jodaeiye Nader az Simin", um filme cativante que aborda temas universais de forma emocionante. Para crianças, "Bacheha-Ye aseman" é uma escolha excelente, conhecido por sua narrativa envolvente e visualmente cativante. E para todos os públicos, "Modern Times", um filme icônico de Charlie Chaplin que continua a encantar gerações com sua comédia atemporal e crítica social.

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

Para analisar isso, plotei um gráfico de dispersão com a receita bruta (Gross) no eixo x e a nota do IMDB no eixo y. No gráfico, identifiquei alguns outliers, que são pontos de dados que se destacam significativamente do padrão geral dos dados, podendo indicar casos excepcionais. No entanto, os fatores que podem estar relacionados com altas expectativas de faturamento incluem gêneros cinematográficos populares, diretores renomados e atores de destaque, que têm influência significativa na recepção e sucesso comercial de um filme.



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

A coluna Overview permite a aplicação de Processamento de Linguagem Natural (PLN), que é uma técnica para análise e interpretação de texto. Através do resumo do filme, podemos identificar o gênero ao qual ele pertence. Utilizei uma nuvem de palavras para visualizar as palavras mais frequentes. E criei outra para o gênero 'Action, Adventure, Drama', explorei as palavras mais relevantes para compreender melhor o contexto e os temas predominantes do filme.

Modelo Preditivo

3) 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? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

Para realizar a previsão da nota do IMDb a partir dos dados fornecidos, seguimos os seguintes passos:

Seleção das Variáveis Preditoras:

Primeiro, selecionamos as variáveis que serão usadas para prever a nota do IMDb. No código fornecido, as variáveis escolhidas foram:

- - Meta_score: A média ponderada das críticas.
- - No_of_Votes: O número de votos recebidos no IMDb.
- - Genre: O gênero do filme.
- - Certificate: A classificação etária do filme.

Transformação das Variáveis Categóricas:

As variáveis categóricas (Genre e Certificate) foram transformadas usando One-Hot Encoding. Esse processo cria variáveis dummy para cada categoria, permitindo que o modelo de aprendizado de máquina capture correlações entre diferentes categorias sem assumir uma ordem natural entre elas.

Divisão dos Dados em Treino e Teste:

Os dados foram divididos em conjuntos de treino e teste utilizando train_test_split do sklearn.model_selection. No código, utilizou-se 70% dos dados para treino e 30% para teste.

Criação e Treinamento do Modelo:

Foi escolhido um modelo de regressão utilizando RandomForestRegressor do sklearn.ensemble. O Random Forest é um modelo de aprendizado de máquina que combina múltiplas árvores de decisão para reduzir o overfitting e melhorar a precisão das previsões.

O modelo foi treinado com os dados de treino utilizando model_rf.fit(X_train, y_train).

Realização de Previsões:

Após treinar o modelo, foram feitas previsões utilizando os dados de teste com model_rf.predict(X_test).

Avaliação do Desempenho do Modelo:

O desempenho do modelo foi avaliado utilizando duas métricas principais:

- **Erro Quadrático Médio (MSE):** Mede a média dos quadrados dos erros entre os valores previstos e os valores reais. Quanto menor o MSE, melhor o modelo.
- **Coeficiente de Determinação (R^2):** Indica a proporção da variância na variável dependente que é previsível a partir das variáveis independentes. Um valor próximo de 1 indica que o modelo explica bem a variação nos dados.

Persistência do Modelo Treinado:

Por fim, o modelo treinado foi salvo em um arquivo .pkl usando joblib.dump do sklearn.externals.joblib, para ser utilizado posteriormente para fazer previsões em novos dados.

Esses passos formam um pipeline básico para construir e avaliar um modelo de regressão que prevê a nota do IMDb com base em características como meta-score, número de votos, gênero e classificação etária do filme.

4) 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?

Neste modelo, utilizei duas características principais, Meta_score (80.0) e No_of_Votes (2,343,110), para prever a nota do filme no IMDb. Os resultados da avaliação do modelo mostram um Erro Quadrático Médio (MSE) de 0.0443 e um Coeficiente de Determinação

(R^2) de 0.4485. Isso indica que o modelo explica aproximadamente 44.85% da variação na nota do IMDb com base nas características fornecidas.

Além disso, a nota prevista do IMDb para o filme com as características específicas é de aproximadamente 8.96. Isso sugere que o modelo prevê que o filme teria uma boa avaliação no IMDb com base nas métricas de Meta_score e No_of_Votes utilizadas.

Conclusão

Concluímos que o projeto de análise cinematográfica para PProductions, conduzido pela Indicium, demonstra uma abordagem robusta e estruturada para orientar decisões estratégicas na produção cinematográfica. O projeto incluiu desde a preparação inicial dos dados, como limpeza e análise exploratória, até a construção de um modelo preditivo para prever as notas no IMDb.

A análise exploratória revelou insights valiosos, como a relação entre gêneros cinematográficos e pontuações no IMDb, destacando categorias como 'Action, Sci-fi' e 'Horror, Sci-Fi' com altas pontuações e faturamento médio. Além disso, foram identificados diretores e atores com maior impacto nas avaliações dos filmes, como Steven Spielberg e Robert De Niro, respectivamente. Observamos que o gênero exerce uma influência significativa no faturamento, sugerindo que para a próxima estreia, filmes de ação, dirigidos por diretores especializados nesse gênero e estrelados por atores renomados, são escolhas estratégicas promissoras.

O modelo preditivo desenvolvido, utilizando RandomForestRegressor, foi treinado com variáveis como Meta_score, No_of_Votes, gênero e classificação etária do filme. Este modelo demonstrou uma capacidade significativa de prever as notas do IMDb, explicando aproximadamente 44.85% da variação com base nos dados fornecidos.

Este documento não apenas evidencia nossa capacidade analítica e técnica, mas também nossa dedicação em transformar complexidade em clareza, promovendo inovação e excelência no campo da análise cinematográfica.