Desafio Cientista de Dados

Vinícius Franklin Pedroso Mansur de Azevedo

Juiz de Fora, setembro de 2025

Sumário

Introdução

Este documento apresenta meu desenvolvimento e abordagem para o **Desafio Cientista de Dados** da empresa **Indicium**. O desafio forneceu um dataset de filmes retirado do IMDb, contendo as seguintes colunas principais:

- **Series_Title** – Nome do filme
- **Released_Year** – Ano de lançamento
- **Certificate** – Classificação etária
- **Runtime** – Tempo de duração (em minutos)
- **Genre** – Gênero
- **IMDB_Rating** – Nota no IMDb
- **Overview** – Sinopse do filme
- **Meta_score** – Média ponderada das críticas
- **Director** – Diretor
- **Star1, Star2, Star3, Star4** – Principais atores/atrizes
- **No_of_Votes** – Número de votos recebidos
- **Gross** – Faturamento em bilheteria

As entregas propostas foram:

1. Realizar uma **análise exploratória dos dados (EDA)**, mostrando as principais características das variáveis e apresentando hipóteses.
2. Responder às perguntas:
    a. Qual filme recomendar para uma pessoa desconhecida?
    b. Quais fatores estão relacionados com a expectativa de faturamento de um filme?
    c. Quais insights podem ser extraídos da coluna *Overview*? É possível inferir o gênero do filme a partir dela?
3. Explicar como realizar a previsão da nota do IMDb: variáveis utilizadas, tipo de problema (regressão ou classificação), modelo escolhido, prós e contras e métricas de avaliação.
4. Realizar a previsão para um filme específico, a partir de suas características.

Neste relatório serão apresentados os resultados da análise e as conclusões possíveis. Já a parte mais técnica, incluindo pré-processamento e código, está disponível no arquivo `.ipynb` e no `README.md` do repositório.

1 Análise Exploratória dos Dados (EDA)

Nesta seção serão abordadas as análises realizadas sobre os dados fornecidos. Para complementar as informações, foram coletados dados adicionais pela API do **TMDb**, incluindo colunas ausentes no dataset original, como o **budget** (orçamento). Essa informação é essencial para análises financeiras.

Nosso foco principal será o aspecto financeiro, com o objetivo de avaliar formas de otimizar o lucro de produções cinematográficas. Podemos calcular tanto o **lucro bruto** quanto o **lucro percentual** em relação ao orçamento.

As fórmulas utilizadas foram:

$$\text{Profit} = \text{Gross} - \text{Budget}$$

$$\text{Lucro Percentual} = \frac{\text{Profit}}{\text{Budget}} \times 100$$

1.1 Análise em Relação ao Gênero

Uma questão relevante é identificar qual gênero deve ser escolhido para maximizar o lucro. Para isso, a coluna *Genre*, que originalmente contém uma lista com todos os gêneros de cada filme, foi decomposta em três colunas: gênero principal, gênero secundário e gênero terciário. A partir dessa divisão, foram gerados gráficos que apresentam métricas como *Count* (quantidade), *Profit (%)* (lucro percentual), *Budget Mean* (média de orçamento), *Gross Mean* (média de bilheteria) e *Profit Mean* (lucro médio).

Para o gênero principal temos:

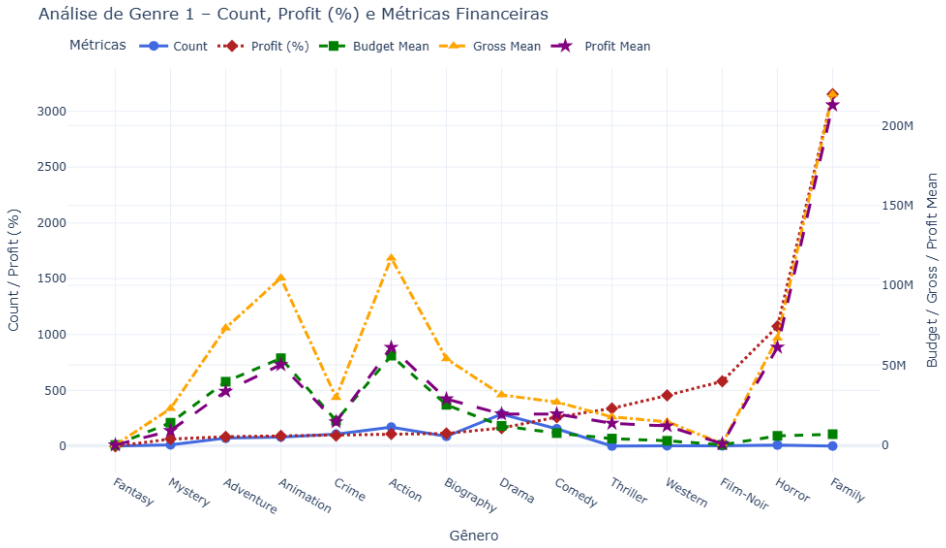

Figura 1: Indicadores em Relação ao Gênero Principal

Para o gênero secundário temos:

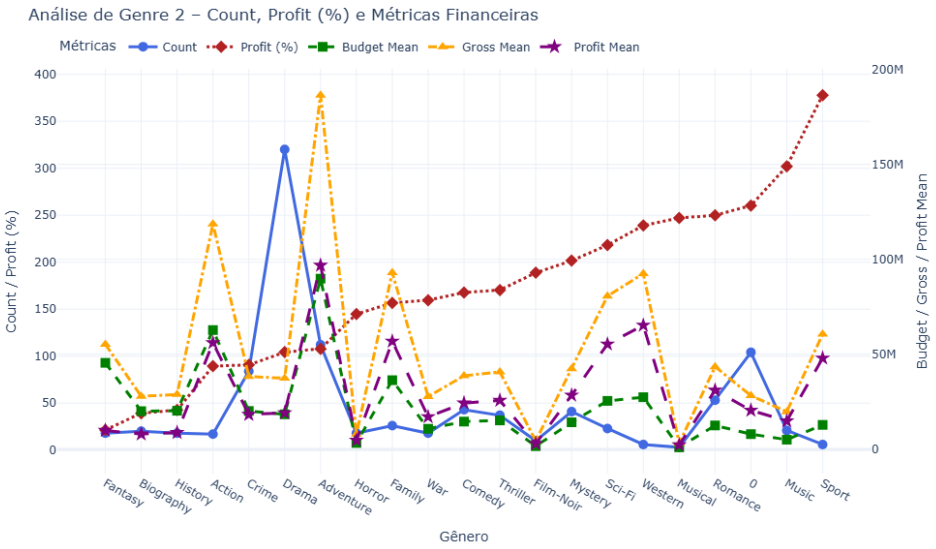

Figura 2: Indicadores em Relação ao Gênero Secundário

Para o gênero terciário temos:

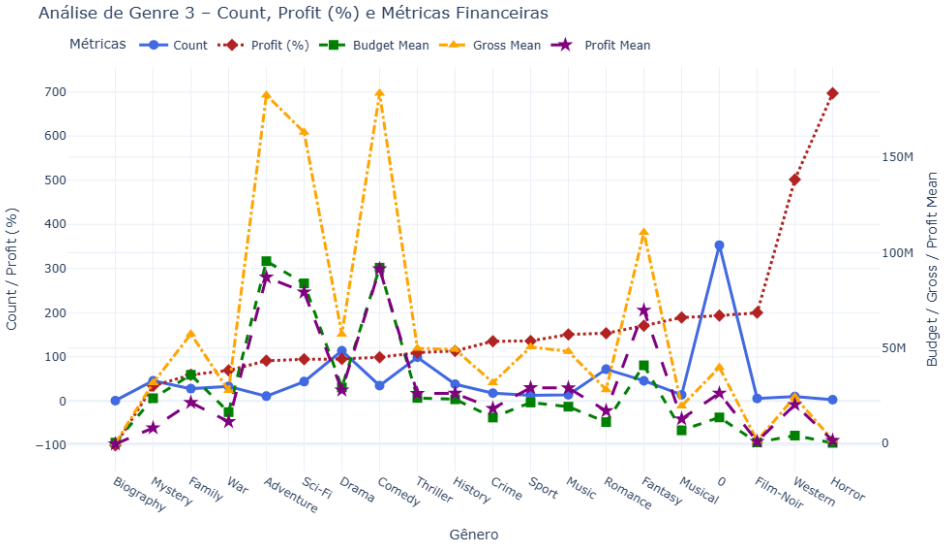

Figura 3: Indicadores em Relação ao Gênero Terciário

A análise desses gráficos evidencia a forte relação entre gênero e desempenho financeiro. Nota-se que a escolha do gênero depende do objetivo da produtora: se a intenção for obter um alto lucro percentual com baixo investimento, os gêneros *família* e, principalmente, *terror* se destacam, sendo este último mais representativo pela quantidade maior de amostras. Já gêneros como *ação*, *animação* e *aventura* apresentam altos orçamentos, mas também retornos proporcionais, reforçados por um número elevado de amostras que aumenta a confiabilidade da análise.

Os gráficos de gêneros secundários e terciários servem como complemento, ajudando a entender a multiplicidade de temas, mas os resultados confirmam que o gênero principal exerce maior peso sobre os indicadores financeiros.

1.2 Análise em Relação aos Atores

Outro ponto importante na idealização de um filme é a escolha dos atores que irão protagonizar a obra e o impacto dessa decisão na receita e no lucro. Naturalmente, quanto maior a celebridade, maior o custo associado, mas a questão central é: esse investimento compensa no resultado final? Além disso, pode ser interessante considerar combinações entre gênero e ator. Há casos de artistas que migram para gêneros diferentes dos habituais e obtêm sucesso, mas essa não costuma ser a regra.

Nos gráficos a seguir foram analisados os 100 atores com maior número de participações, independentemente de serem protagonistas ou coadjuvantes. As métricas consideradas incluem quantidade de filmes, lucro médio, orçamento médio e percentual de lucro. Para melhorar a visualização, a análise foi dividida em duas figuras. Vale destacar que, nesses gráficos, há uma escala de cores exibida ao lado: quanto mais próximo de 1 for o valor, maior a probabilidade de o ator estar associado a uma franquia de filmes.

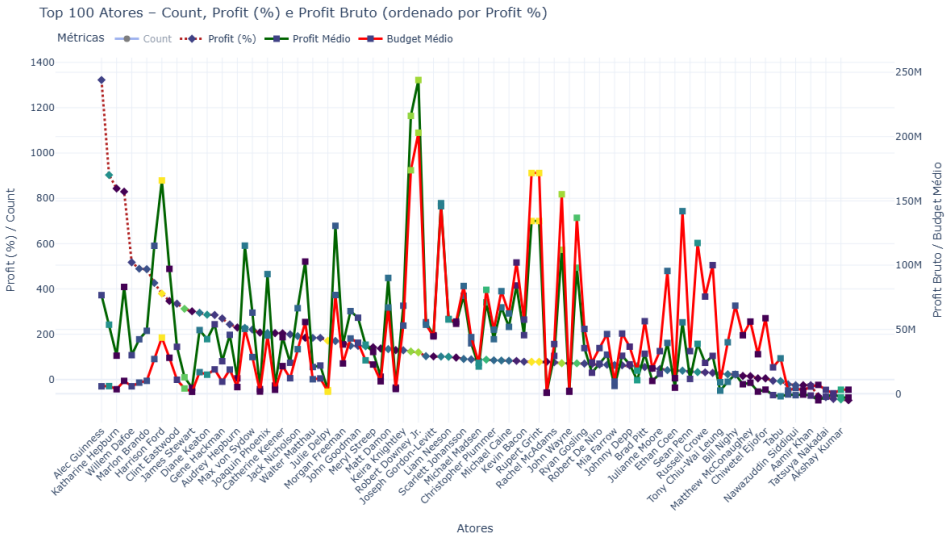

Figura 4: Indicadores Financeiros Relacionados a Atores (Parte 1)

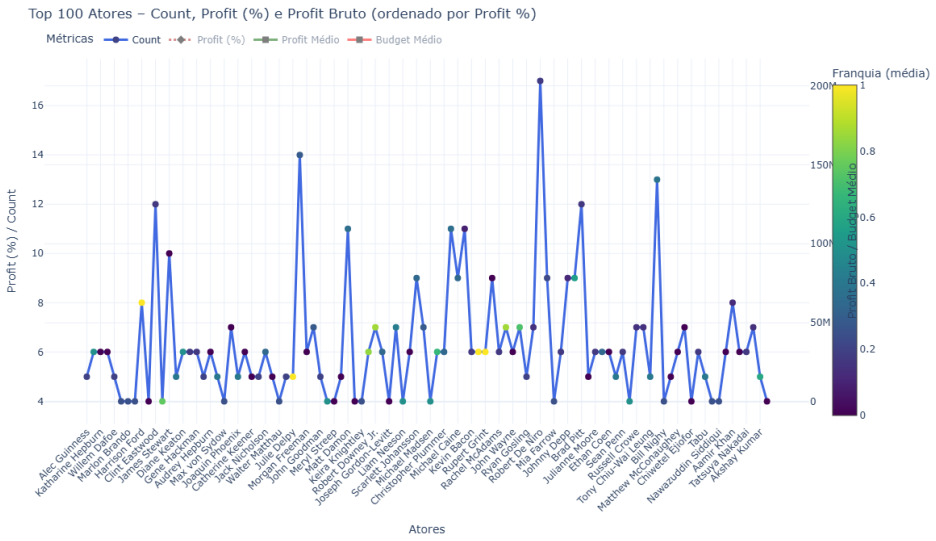

Figura 5: Indicadores Financeiros Relacionados a Atores (Parte 2)

Os resultados mostram que a escolha do elenco depende fortemente do objetivo do investidor: pode-se optar por altos orçamentos em busca de grandes bilheteiras, ou por estratégias que priorizam retorno percentual, com investimentos menores e multiplicadores mais altos. Como esperado, Robert Downey Jr. se destaca com a maior média bruta, refletindo seu papel em franquias de enorme sucesso. No entanto, ao observar tanto retorno absoluto quanto percentual, atores como Harrison Ford apresentam uma boa relação custo-benefício, com investimentos relativamente baixos e retornos elevados.

É importante ressaltar que essa análise utiliza dados históricos. No caso de Harrison Ford, por exemplo, seu grande destaque se deve à franquia *Star Wars*, mas quando iniciou sua participação ele ainda era um ator pouco conhecido e, portanto, barato para os estúdios. Situações semelhantes ocorrem com outros atores em diferentes fases da carreira: antes de se tornarem estrelas consolidadas, seus custos eram menores. Mesmo assim, os gráficos fornecem uma visão relevante sobre a influência dos atores no desempenho financeiro dos filmes.

1.3 Análise em Relação ao Diretor

Agora iremos realizar uma análise semelhante à feita para os atores, mas com foco nos diretores. A ideia é identificar quem seria o nome mais indicado a comandar nosso navio e nos levar ao sucesso. O diretor é, sem dúvida, uma das figuras mais importantes em um projeto cinematográfico. Abaixo temos o gráfico com os diretores:

Figura 6: Gráfico em Relação a Diretores, Sentido Financeiro

O grande problema dessa análise é a pequena amostragem por diretor. Em muitos casos há apenas um filme de cada, o que tende a destacar seus melhores trabalhos, já que este dataset contempla somente produções com maiores notas no IMDb. Por isso, entendo que essa comparação tenha menor relevância. Caso utilizássemos a API do TMDB para expandir consideravelmente o conjunto de dados, essas informações ganhariam mais robustez. Ainda assim, os números chamam atenção: o lucro absoluto gerado por Rogers Allers é impressionante, assim como o lucro percentual de John G. Avildsen, que chega a valores extraordinários.

Como complemento, podemos observar também o desempenho sob a ótica da qualidade das obras dirigidas. O gráfico a seguir está ordenado pela nota média no IMDb e inclui métricas adicionais como *Metascore*, duração, número de filmes, quantidade de votos e popularidade:

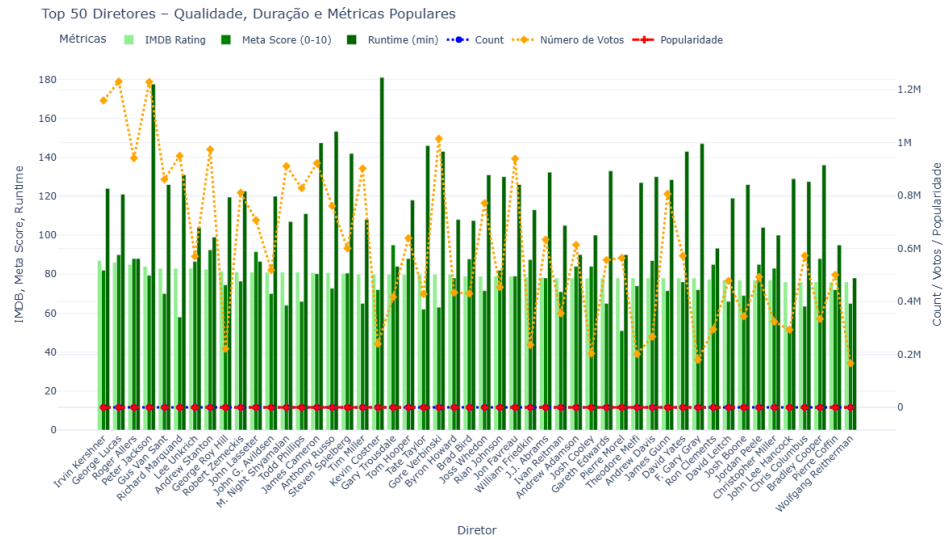

Figura 7: Popularidade por Diretor

1.4 Análise em Relação ao Tempo

Agora vamos realizar uma análise temporal, considerando tanto a data de lançamento quanto a duração dos filmes. Ambos os aspectos são importantes para a idealização do projeto, permitindo entender como algumas características se modificaram ao longo do tempo. Abaixo temos três gráficos:

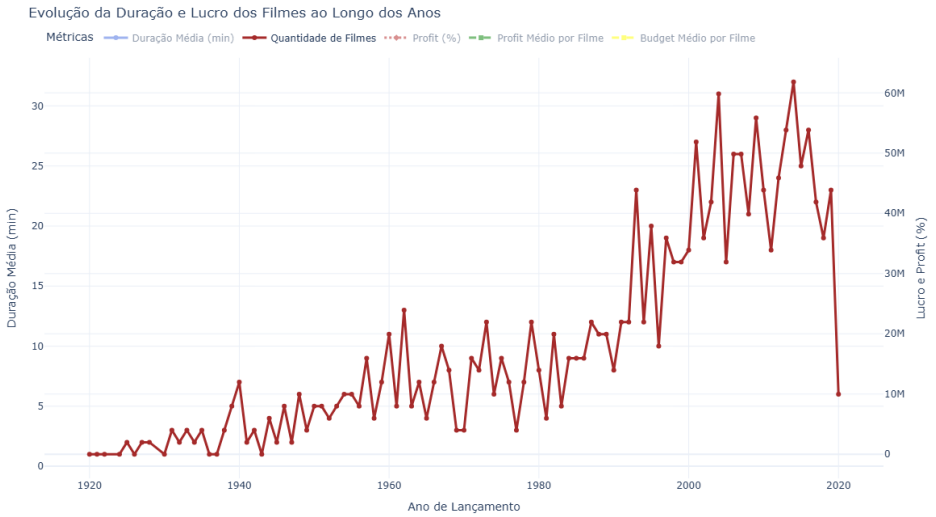

Figura 8: Quantidade de Filmes por Ano

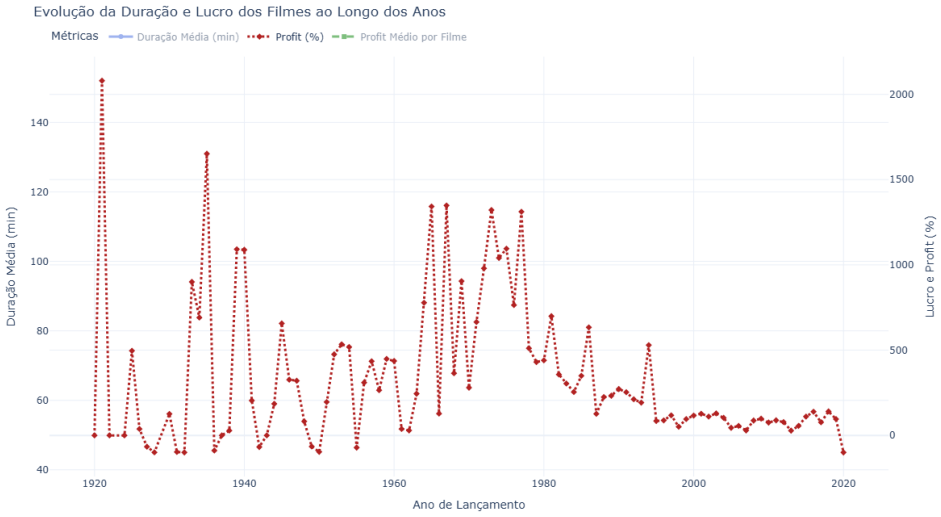

Figura 9: Lucro Percentual por Ano

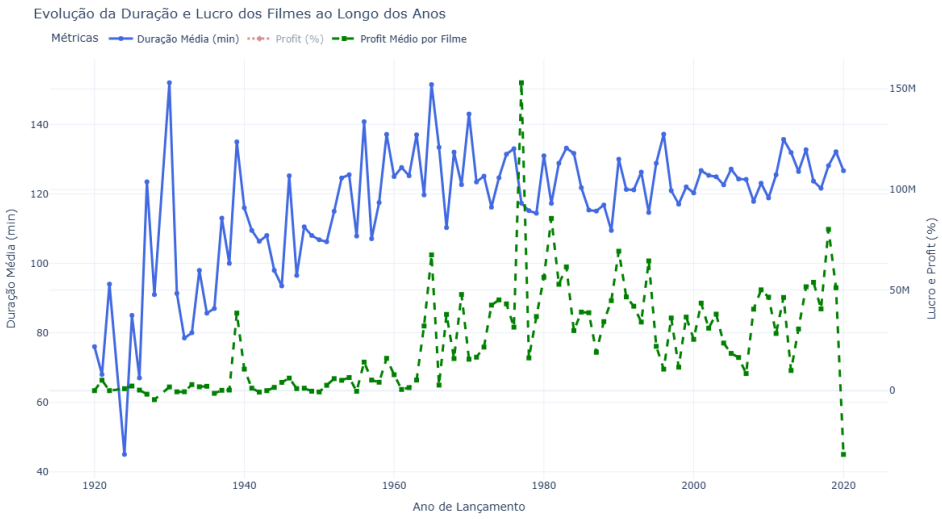

Figura 10: Duração Média dos Filmes por Ano

A primeira observação é relativamente óbvia, mas interessante: filmes lançados exclusivamente em streaming não apresentam lucro direto, e filmes de 2020 tiveram retorno negativo, possivelmente devido à pandemia, com muitos lançamentos direto no streaming.

O lucro médio apresenta certa constância, com altos e baixos ano a ano, mas sua média geral indica um comportamento estável. Já o lucro percentual apresenta tendência de queda ao longo dos anos. Em relação ao orçamento, observa-se um aumento quase linear com o tempo.

Também é perceptível o aumento no número de filmes produzidos por ano, possivelmente refletindo a preferência do público por lançamentos mais recentes, já que a nota do IMDb reflete a percepção da audiência.

1.5 Análise em Relação ao País de Origem e Linguagem

Mesmo que o filme seja produzido em Hollywood, é interessante considerar o idioma original, pois ele pode impactar tanto na percepção de qualidade quanto no desempenho financeiro. Para isso, apresentamos os seguintes gráficos:

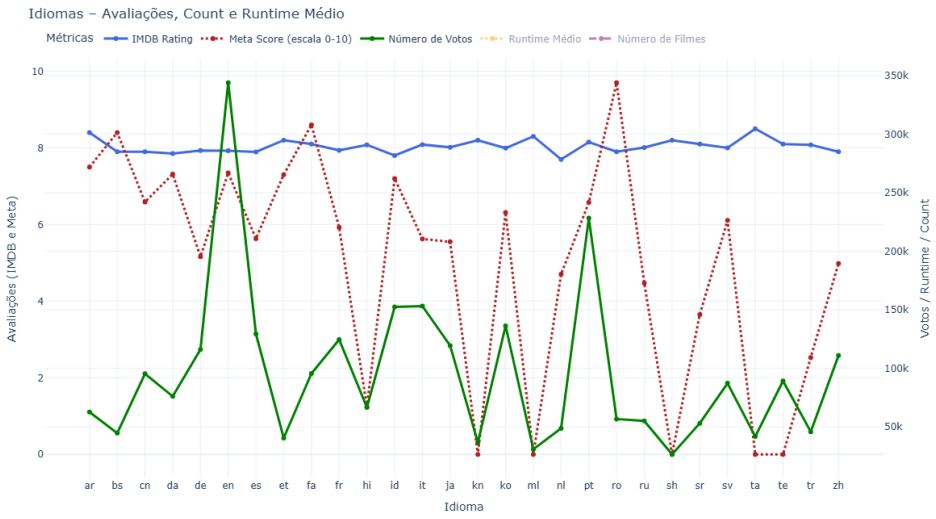

Figura 11: Notas por Idiomas

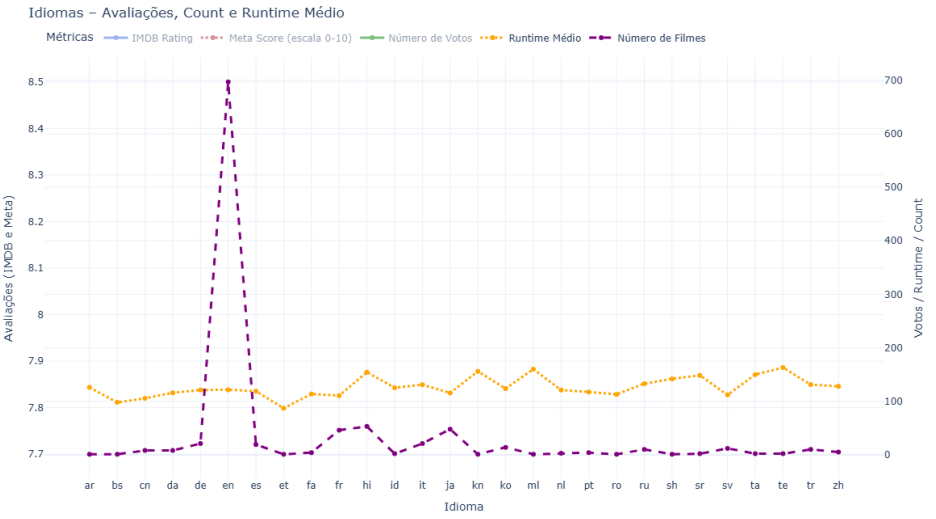

Figura 12: Duração Média e Número de Filmes por Idioma

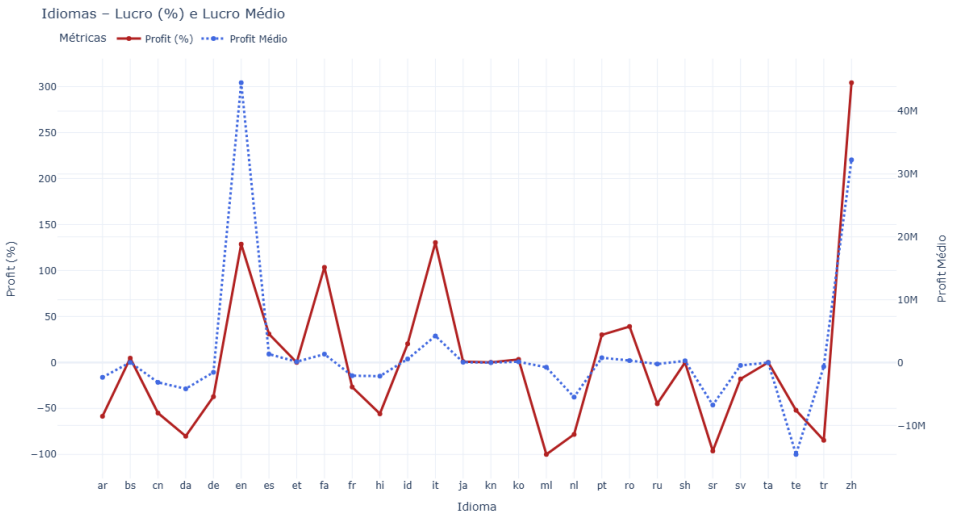

Figura 13: Lucro por Idioma

Constata-se que o inglês é, de forma evidente, o idioma mais popular, concentrando um número muito superior de votos em relação aos demais. Destaca-se também o português em segundo lugar, possivelmente relacionado ao forte engajamento do público brasileiro, especialmente em mídias sociais voltadas para produções nacionais. Em contrapartida, o chinês ("zh", referente ao mandarim) apresenta baixa representatividade, possivelmente devido às restrições da intranet na China, que limitam o acesso da população a sites internacionais como IMDb e TMDb.

Quanto à qualidade, observa-se que ela não depende exclusivamente do inglês, já que outros idiomas também apresentam boas avaliações. No entanto, é necessário cautela nessa análise, pois a quantidade de amostras em inglês (697) é muito superior às demais, tornando a comparação estatisticamente desigual.

No aspecto financeiro, os maiores valores de lucro aparecem em produções em inglês e mandarim, como esperado. É interessante notar, porém, que o desempenho do mandarim se aproxima do inglês, evidenciando a relevância do mercado asiático. Nos demais idiomas, os resultados são mais dispersos, muitas vezes negativos, possivelmente devido a menores investimentos e a uma audiência reduzida.

1.6 Análise em Relação a Métricas da Nota do IMDb

Agora podemos analisar a relação entre a nota IMDb e o desempenho financeiro. É importante ter em mente que a nota IMDb reflete a avaliação dos usuários do site. Abaixo temos o gráfico correspondente:

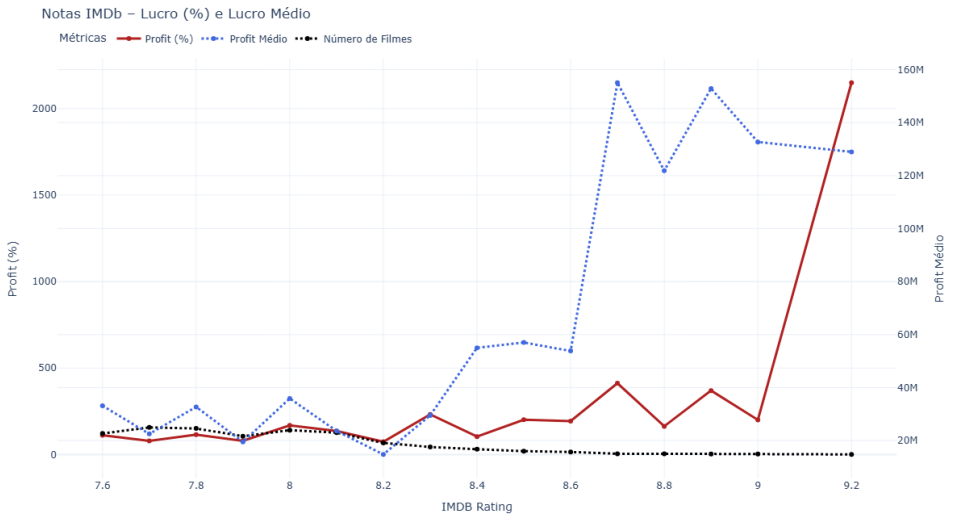

Figura 14: Lucro por Nota IMDb

Podemos observar uma tendência de crescimento proporcional entre as duas métricas, principalmente no que se refere ao lucro médio. Nos filmes com notas mais altas, verifica-se um valor extremamente elevado de lucro percentual. Como esperado, por se tratar de avaliações do público, há uma correlação com o sucesso de bilheteria. Isso reforça a ideia de que a nota IMDb pode servir como um indicativo da qualidade percebida pelos espectadores, influenciando a decisão de ir ao cinema ou assistir a um filme em casa.

1.7 Análise em Relação a Métricas da Nota do Metascore

Agora analisamos uma métrica que não reflete a opinião popular: o Metascore do IMDb, que utiliza avaliações de críticos. Isso muda completamente a análise. O gráfico abaixo ilustra a relação entre o Metascore e o desempenho financeiro:

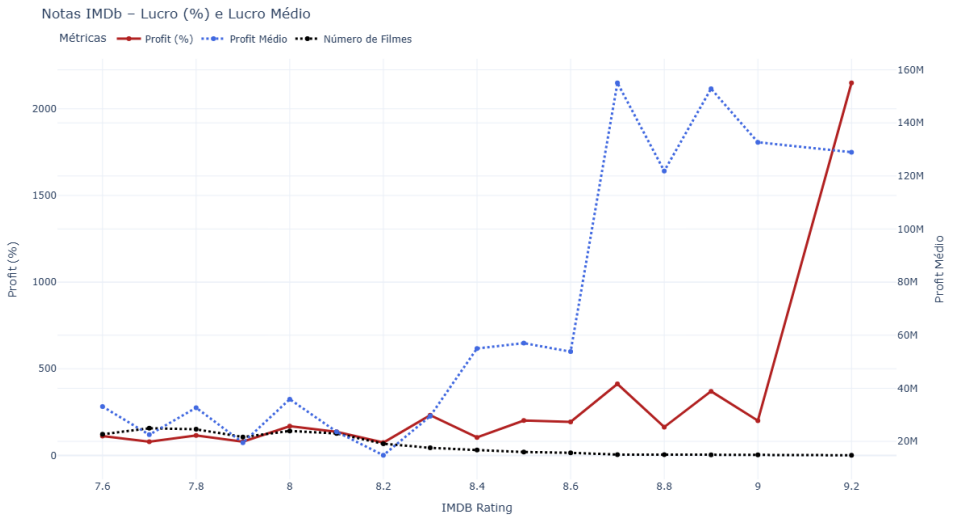

Figura 15: Lucro por Nota Metascore

Observa-se que não há uma relação clara entre o ganho financeiro e a nota atribuída pelos críticos. Isso ilustra exatamente o que ocorre: a avaliação crítica muitas vezes difere da percepção do público. Dessa forma, os dois indicadores refletem maneiras diferentes de se apreciar a obra cinematográfica.

1.8 Análise em Relação a Classificação Indicativa

A classificação indicativa é um tema delicado na indústria cinematográfica. Apesar da crença de que filmes blockbuster devem alcançar todo o público, casos como *Deadpool* mostram que produções voltadas para faixas etárias específicas também podem gerar bons resultados. No dataset, as classificações seguem o padrão dos EUA, incluindo G, PG, PG-13, R, A, Approved, Passed, U, U/A, UA, 16, TV-14, TV-MA, TV-PG e Unrated. No Brasil, as equivalentes seriam L (Livre), 10, 12, 14, 16 e 18 anos, de acordo com a faixa etária recomendada pelo Ministério da Justiça.

Os gráficos abaixo mostram a relação entre a classificação indicativa, a nota IMDb e o desempenho financeiro:

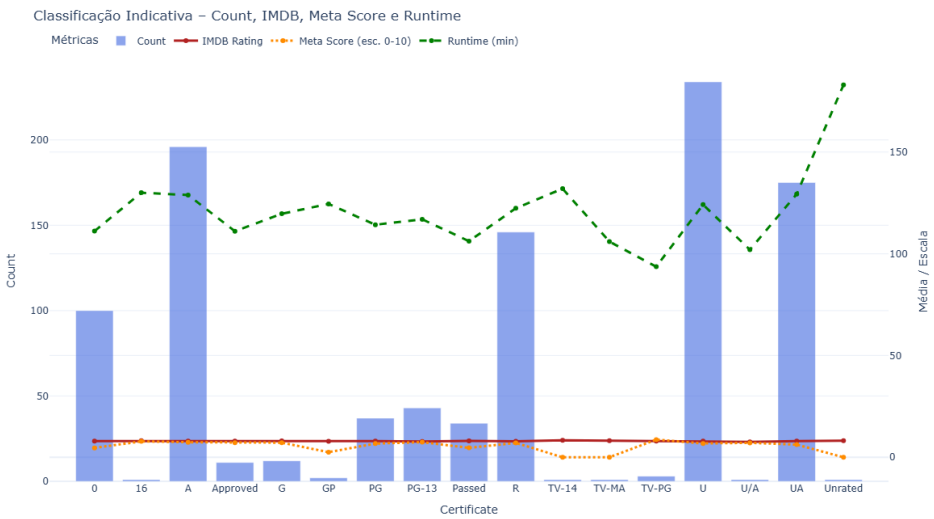

Figura 16: Notas por Classificação Indicativa

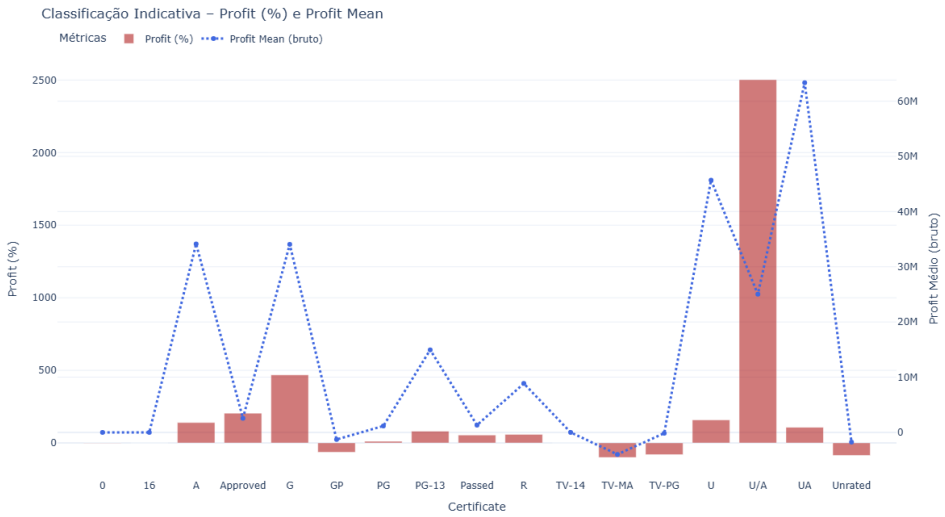

Figura 17: Lucro por Classificação Indicativa

Ao observar os dados, percebe-se que filmes voltados para público amplo ou familiar, como as categorias A, G, U e UA, apresentam altos lucros médios e percentuais, indicando que esse tipo de produção pode ser financeiramente vantajoso. Já filmes com classificação R e TV-MA mostram lucros médios elevados, mas com maior variabilidade, refletindo investimentos maiores e público mais restrito. Algumas categorias com poucas amostras, como 16, TV-14 e Unrated, apresentam resultados instáveis, reforçando a necessidade de cautela ao interpretar esses valores.

Em resumo, a escolha da classificação indicativa deve considerar o objetivo do projeto: maximizar o lucro percentual com menor investimento ou buscar maior retorno bruto mesmo com um público mais limitado.

1.9 Correlação

Uma forma eficiente de entender como as variáveis do nosso dataset se relacionam é através da análise de correlação, considerando apenas colunas numéricas, como inteiros e floats. A tabela de correlação nos permite observar como mudanças em uma variável podem influenciar outras, sendo essencial para identificar padrões ou dependências entre métricas financeiras, notas, votos e características do filme.

O gráfico abaixo apresenta a matriz de correlação de forma visual:

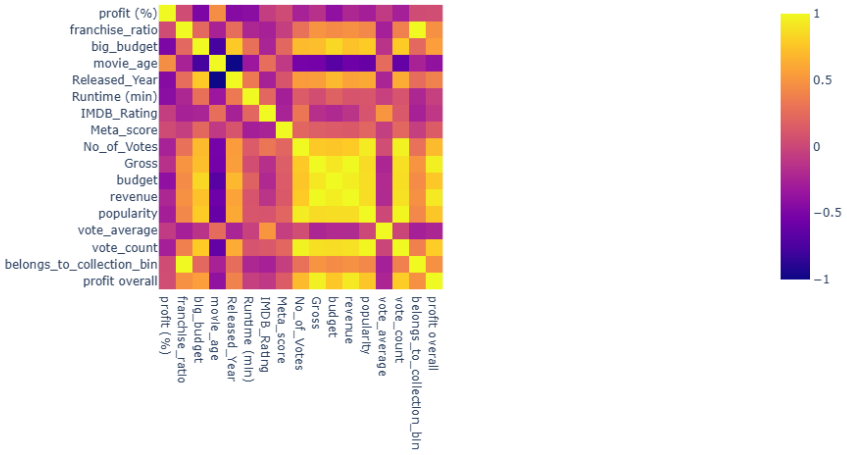

Figura 18: Gráfico de Correlação entre Variáveis Numéricas

Ao analisar os dados, algumas relações se destacam. O lucro total apresenta correlação positiva moderada com o número de votos (0,54) e receita bruta (0,92), indicando que filmes mais populares e de maior bilheteria tendem a gerar mais lucro. Por outro lado, o orçamento possui alta correlação com receita e lucro bruto, mas menor relação com o percentual de lucro, mostrando que investimentos maiores podem aumentar o ganho absoluto, mas não garantem eficiência financeira.

Outras observações importantes incluem a correlação positiva entre franquias e lucro (0,37) e entre popularidade e número de votos (0,77), evidenciando que filmes de franquias ou mais conhecidos pelo público tendem a gerar resultados melhores em várias métricas. Já a idade do filme e o ano de lançamento apresentam correlação negativa, refletindo que filmes mais antigos possuem métricas financeiras e de audiência diferentes em relação aos lançamentos mais recentes.

O lucro absoluto depende fortemente de faturamento e presença em franquias, enquanto o lucro percentual mostra que filmes de baixo orçamento podem ser mais eficientes. Orçamentos altos não garantem lucro proporcional. Popularidade e votos estão mais ligados a faturamento e orçamento do que à avaliação crítica.

Para a nota IMDb, observa-se que o tempo de filme tem alguma relação, curiosamente negativa com o Metascore. O número de votos impacta diretamente a nota, indicando que filmes com boas avaliações tendem a receber mais interações dos usuários, mas não há relação forte com outras variáveis, confirmando que a avaliação crítica e a popularidade do público nem sempre caminham juntas.

2 Recomendação

Segundo as orientações do trabalho, devemos recomendar um filme para uma pessoa aleatória, sem informações sobre suas preferências.

Para isso, defini alguns critérios que tornam o filme adequado para qualquer público:

1. Filme popular, considerando o número de votos e a métrica de popularidade. 2.

Filme que não faça parte de uma franquia, tornando-o fácil de assistir isoladamente. 3. Duração menor que 120 minutos, evitando longas que podem cansar o espectador. 4. Número de votos maior que a média, garantindo engajamento do público. 5. Boa nota combinada de IMDb e Metascore, usando a média das duas. 6. Classificação indicativa geral (G / U), adequada para qualquer faixa etária. 7. Filmes mais recentes (lançados após 2000), considerando que muitas pessoas preferem produções atuais.

Aplicando esses filtros na base de dados, obtivemos o top 5, sendo que os dois primeiros se destacam:

- **Up**: uma excelente indicação, agradando todas as idades, bem avaliado pela crítica e pelo público.

- **Wall-E**: segue o mesmo padrão, reforçando que a estratégia funcionou.

Como a pessoa é aleatória, a atenção à classificação indicativa é importante. A popularidade, combinada com boas notas, garante a chancela da crítica e do público quanto à qualidade do filme. A duração curta ajuda a manter o interesse, enquanto a escolha de filmes mais recentes atende a preferências de públicos mais jovens.

A estratégia aqui aplicada é simples, mas eficiente, utilizando esses filtros. Caso conhecêssemos melhor a pessoa, seria possível criar recomendações mais sofisticadas, ou ainda utilizar o endpoint de recomendação da TMDb, ajustando o filtro com base em gostos ou filmes já apreciados pelo usuário.

3 Inferir Informações do Overview

A partir da coluna `overview`, é possível tentar inferir o gênero principal de um filme através da ocorrência de palavras. Essa tarefa poderia ser feita manualmente, verificando a presença de palavras-chave e delimitando o final delas por espaços ou pela terminação das sentenças. Nesse caso, os valores seriam binários e poderiam ser utilizados em modelos de classificação simples, como KNN. Porém, fazer isso manualmente seria muito trabalhoso, então optamos por utilizar bibliotecas que automatizam o processo.

Para este exercício, usamos a biblioteca `TfidfVectorizer` com os seguintes parâmetros:

```
TfidfVectorizer(
    max_features=8000,
    stop_words='english',
    min_df=2,
    lowercase=True,
    max_df=0.60,
    token_pattern=r'\b[a-zA-Z]{3,}\b',
    ngram_range=(1,2)
)
```

Essa abordagem nos permite extrair informações relevantes do `overview`, do `tagline` do TMDb e de descrições adicionais, limitando palavras muito frequentes (como artigos) e tratando a multiplicidade de gêneros em um filme.

A partir dessa base de dados, testamos diversos modelos de classificação, sendo que a `Logistic Regression` apresentou o melhor desempenho. O classification report obtido foi:

Tabela 1: Performance do modelo Logistic Regression para classificação de gêneros a partir do `overview`

Gênero	**Precision**	**Recall**	**F1-Score**	**Support**
Action	0.43	0.56	0.49	16
Adventure	0.17	0.25	0.20	8
Animation	0.43	0.38	0.40	8
Biography	0.40	0.67	0.50	9
Comedy	0.33	0.21	0.26	14
Crime	0.25	0.30	0.27	10
Drama	0.62	0.38	0.47	34
Fantasy	0.00	0.00	0.00	0
Horror	1.00	1.00	1.00	1
Mystery	0.00	0.00	0.00	0
Accuracy	-	-	0.40	-
Macro Avg	0.36	0.38	0.36	100
Weighted Avg	0.44	0.40	0.41	100

Para o propósito do exercício, consideramos esse desempenho aceitável, com acurácia e F1-score médios de cerca de 41%.

Percebe-se que, mesmo com um dataset pequeno e desbalanceado, algumas classes se destacam. Por exemplo, a classe `Biography` teve bom rendimento mesmo com apenas 9 amostras, superando até o `Drama` com 34 amostras, indicando um padrão de utilização de palavras mais consistente nesse gênero.

Para melhorar o modelo, seria interessante:

- Obter o máximo de dados possíveis via API do TMDb, aumentando a base para cada gênero.
- Ajustar os parâmetros do `TfidfVectorizer` e testar combinações diferentes.
- Considerar abordagens para lidar com múltiplos gêneros por filme.

Mesmo com as limitações do dataset, a análise foi válida para avaliar as palavras mais frequentes e demonstrar a viabilidade de inferir informações do `overview`.

4 Previsão de Nota do IMDb

O range de notas do dataset é pequeno, o que permite até considerar o problema como classificação. Entretanto, a regressão faz mais sentido, especialmente se expandirmos o modelo para prever filmes fora do dataset. Caso quiséssemos usar classificação, poderíamos dividir em faixas (boa, média e ruim), baseando-se na média.

Porém, como se trata de valores contínuos (ainda que discretizados no dataset), a regressão é a abordagem mais adequada. Ressalta-se que, apesar do dataset conter 1000 linhas, esse número é relativamente pequeno e ainda enviesado, pois contempla apenas os 1000 melhores filmes segundo o IMDb, com notas entre 7.6 e 9.2. Isso é uma faixa reduzida, considerando o intervalo total possível de 0 a 10. Dessa forma, prever notas menores que 7.6 com este modelo torna-se difícil ou até impossível.

4.1 Seleção de Colunas

As colunas de interesse selecionadas foram:

```
colunas_modelo = [ 'Released_Year', 'Certificate', 'Runtime (min)',
        'Genre 1', 'Genre 2', 'Genre 3',
        'Meta_score','Star1', 'Star2', 'Star3', 'Star4',
        'No_of_Votes', 'Gross', 'budget','origin_country',
        'original_language', 'popularity', 'production_countries',
        'revenue', 'vote_average', 'vote_count',
        'belongs_to_collection_name','production_companies_names',
        'belongs_to_collection_bin', 'profit overall', 'IMDB_Rating']
```

Quando havia múltiplos valores em uma coluna (listas ou dicionários), selecionou-se o mais relevante, como no caso de *production_companies_names*, em que se utilizou a principal.

As colunas categóricas foram tratadas com *LabelEncoder*. Em seguida, o dataset foi dividido em treino (80%) e teste (20%), de forma aleatória e embaralhada.

```
train, test = np.split(
    tabela_modelo.sample(frac=1, random_state=42),
    [int(0.8*len(tabela_modelo))]
)
```

Após a divisão, aplicou-se o *StandardScaler* para normalizar os valores numéricos, evitando discrepâncias entre variáveis.

4.2 Métricas de Avaliação

Foram utilizadas as seguintes métricas:

- **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **Root Mean Squared Logarithmic Error (RMSLE):**

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(1+\hat{y}_i) - \log(1+y_i)\right)^2}$$

- **Coeficiente de Determinação (R^2):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

É importante destacar o motivo da escolha dessas métricas. O **Mean Squared Error (MSE)** é especialmente útil porque, ao elevar as diferenças ao quadrado, evita que erros positivos e negativos se anulem, semelhante ao que ocorre na análise de sinais em engenharia: um cosseno centrado em zero teria média nula, mas ao elevarmos ao quadrado conseguimos medir sua potência real, independentemente do sinal ser positivo ou negativo. O **Root Mean Squared Error (RMSE)** segue a mesma ideia, mas retorna o erro na mesma escala da variável prevista, facilitando a interpretação. Já o **Root Mean Squared Log Error (RMSLE)** atenua discrepâncias em valores maiores, sendo útil quando diferenças relativas são mais importantes que absolutas. Por fim, o **coeficiente de determinação (R^2)** mede a proporção da variância dos dados explicada pelo modelo, permitindo avaliar sua capacidade de generalização. Assim, o uso conjunto dessas métricas fornece uma visão ampla e equilibrada do desempenho do modelo.

4.3 Resultados dos Modelos

4.3.1 Random Forest Regressor

Configuração: $n\_estimators = 10$, $random\_state = 42$, $oob\_score = True$.

Métrica	**Valor**
Out-of-Bag Score	-11.4814
Mean Squared Error	0.0446
Root Mean Squared Error	0.2111
Root Mean Squared Log Error	0.0230
R-squared	0.4019

Tabela 2: Resultados do Random Forest Regressor

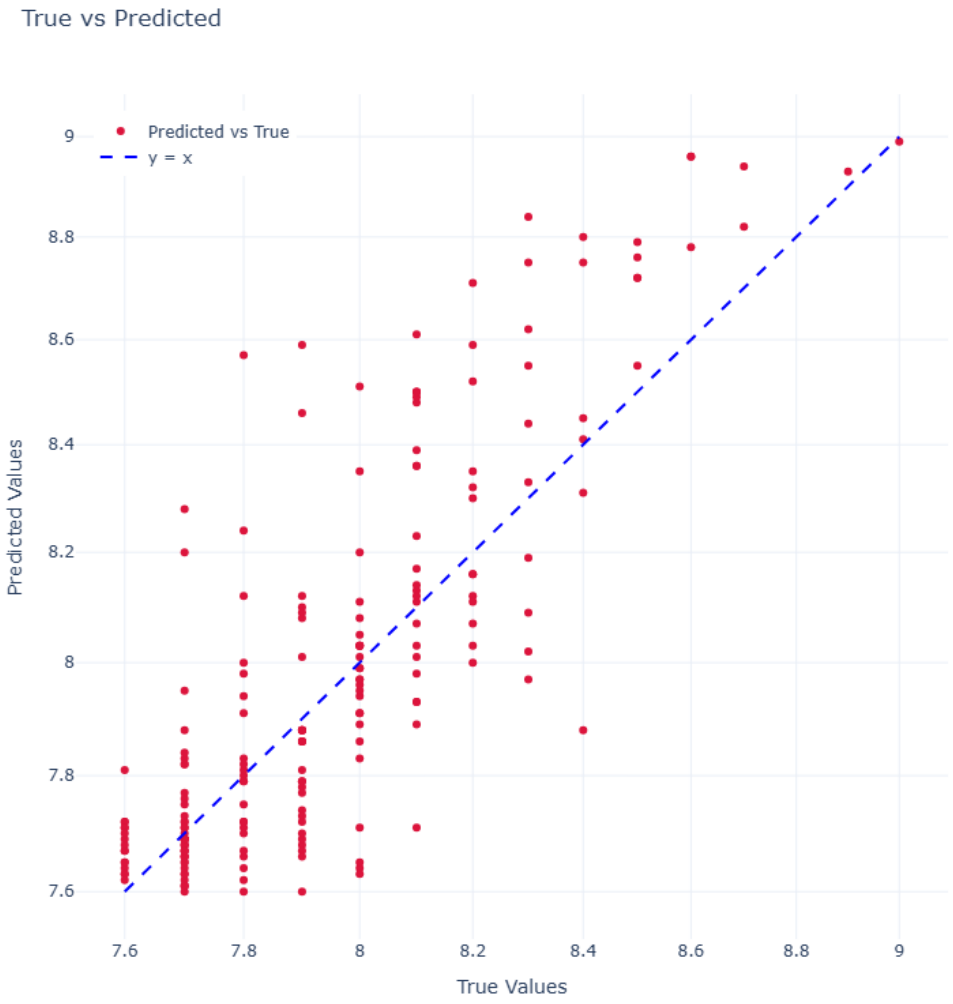

Figura 19: Gráfico do Random Forest Regressor

4.3.2 XGBoost

Configuração: $test\_size = 0.25$, $max\_depth = 3$, $learning\_rate = 0.1$, $n = 50$.

Métrica	**Valor**
Mean Squared Error	0.0163
Root Mean Squared Error	0.1278
Root Mean Squared Log Error	0.0143
R-squared	0.7639

Tabela 3: Resultados do XGBoost

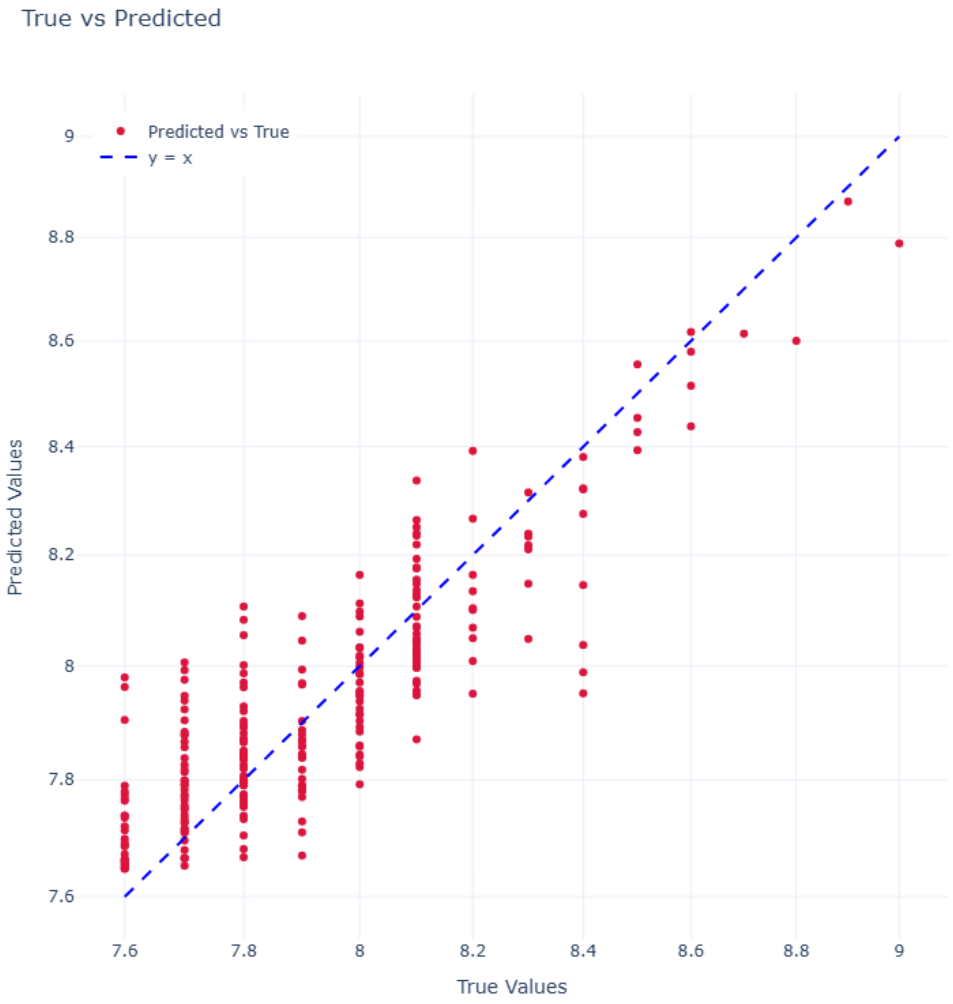

Figura 20: Gráfico do XGBoost

4.4 Comparação dos Modelos

Métrica	**Random Forest**	**XGBoost**
Mean Squared Error	0.0446	0.0163
Root Mean Squared Error	0.2111	0.1278
Root Mean Squared Log Error	0.0230	0.0143
R-squared	0.4019	0.7639

Tabela 4: Comparação entre Random Forest e XGBoost

Observa-se que o **XGBoost apresentou resultados superiores**, com menor erro e maior R^2, mostrando-se mais eficiente para este problema.

4.5 Previsão para o Filme The Shawshank Redemption

Para prever um filme isolado, categorias novas (atores, gêneros, produtoras) que não apareceram no treino foram mapeadas para o valor `-1`.

Rodando o modelo XGBoost obteve-se:

$$\text{Predicted IMDb Rating} = 8.82$$

O valor real é 9.3, resultando em um erro de aproximadamente 0.48. Esse erro, embora maior que a média do modelo, ainda é considerado baixo, principalmente porque notas acima de 8.8 representam apenas 0.7% do dataset, o que torna a previsão mais difícil.

5 Conclusão

Neste trabalho, foi possível atender a todas as entregas propostas no desafio, unindo análise exploratória, modelagem preditiva e aplicação prática em um caso real.

Na **Análise Exploratória dos Dados (EDA)**, investigamos diferentes perspectivas como gênero, atores, diretores, classificação indicativa, idioma, tempo de lançamento e duração, além de métricas de qualidade (IMDb e Metascore). Foram extraídos insights financeiros relevantes, como a relação entre orçamento, bilheteria e lucro, mostrando quais escolhas de gênero e elenco podem maximizar o retorno de investimento.

Na etapa de **Recomendação**, aplicamos critérios objetivos (popularidade, notas, classificação indicativa, duração, ausência de franquia) para selecionar filmes adequados a qualquer espectador. O resultado apontou produções como *Up* e *Wall-E*, ambas com forte apelo universal, notas altas e boa recepção crítica.

A análise da coluna **Overview** demonstrou a viabilidade de extrair informações textuais. Utilizando TF-IDF e regressão logística, foi possível inferir gêneros com acurácia aceitável, mesmo em um dataset pequeno e desbalanceado, evidenciando o potencial de enriquecimento com mais dados.

No desenvolvimento do modelo para **Previsão da Nota do IMDb**, foi justificada a opção pela regressão em detrimento da classificação. Foram aplicados modelos de *Random Forest Regressor* e *XGBoost*, avaliados com métricas robustas (MSE, RMSE, RMSLE e R^2). O *XGBoost* apresentou melhor desempenho, com erro médio menor e maior capacidade de generalização.

Um dos principais diferenciais deste trabalho foi a **integração com a base do TMDb**. Como os títulos vinham em formatos diferentes, foi necessário aplicar técnicas de normalização e *fuzzy matching* para casar corretamente os filmes. Esse esforço permitiu enriquecer a análise com variáveis críticas como orçamento, receita global e participação em franquias, possibilitando insights mais realistas sobre fatores que afetam a bilheteria. Esse passo extra, embora trabalhoso, aproximou o estudo de um cenário real da indústria cinematográfica, agregando grande valor à análise.

Por fim, a previsão do filme *The Shawshank Redemption* validou a abordagem: o modelo estimou uma nota de 8.82 contra a real de 9.3, erro considerado pequeno dentro do contexto do dataset. Esse resultado mostra que, mesmo com limitações na base (força

de amostra reduzida e enviesada para notas altas), o modelo conseguiu entregar previsões consistentes.

Em suma, o trabalho cumpriu todos os objetivos propostos, entregando não apenas respostas às perguntas do desafio, mas também um pipeline analítico que une exploração, enriquecimento de dados externos, modelagem e avaliação crítica. Com mais dados e ajustes finos, este framework pode ser expandido para aplicações reais no setor cinematográfico, auxiliando estúdios em decisões estratégicas de alto impacto.