Estudo de Caso: Dados PProductions

Cauã Braga de Lima
Processo Seletivo Lighthouse

RELATÓRIO DE ANÁLISES E RESULTADOS

indicium

SUMÁRIO

1. Introdução

O setor cinematográfico é uma indústria de grande impacto cultural e econômico em escala global. Com o aumento da disponibilidade de dados relacionados a filmes, tornou-se possível aplicar técnicas de Ciência de Dados para extrair insights valiosos que auxiliem estúdios, produtoras e profissionais do ramo em tomadas de decisão estratégicas. Neste contexto, a PProductions, um estúdio cinematográfico fictício, busca compreender melhor os fatores que influenciam o sucesso de seus filmes, com ênfase na previsão de notas do IMDB, na identificação de elementos associados a alto faturamento e na inferência de gêneros com base nas sinopses.

Este relatório descreve um estudo de caso completo de análise de dados e modelagem preditiva realizado como parte do processo seletivo Lighthouse da Indicium. O conjunto de dados utilizado contém informações detalhadas sobre 999 filmes, incluindo metadados como elenco, diretor, ano de lançamento, sinopse, classificação indicativa, duração, faturamento, número de votos e notas do IMDB e Metacritic.

O estudo foi conduzido com base na metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining), que orienta todas as etapas do projeto desde o entendimento do negócio e dos dados até a implantação de modelos de machine learning. Foram desenvolvidos dois cenários de previsão de notas do IMDB: um antes do lançamento de um filme (sem informações de bilheteria ou número de votos) e outro após o lançamento (com todos os dados disponíveis). Além disso, foram realizadas análises exploratórias, extração de features textuais com técnicas de NLP e modelagem com diversos algoritmos de regressão.

Os resultados obtidos fornecem modelos preditivos robustos e também insights acionáveis sobre quais elementos contribuem para o sucesso comercial e crítico de um filme, atendendo assim aos principais questionamentos do estúdio e demonstrando o valor da análise de dados no contexto da indústria do cinema, possibilitando que a empresa PProductions tenha mais clareza na escolha dos filmes a serem lançados para obter maior faturamento e melhores notas da crítica.

2. Metodologia

Para guiar este projeto foi adotada a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining). Essa escolha visa padronizar a definição de objetivos e estabelecer de forma clara as etapas de análise necessárias para a resolução do problema de negócio proposto pelos Estúdios PProductions.

Foi fornecido uma base de dados com 999 amostras e 15 colunas referente aos seguintes atributos cinematográficos:

- Series_Title -> Nome do filme
- Released_Year -> Ano de lançamento
- Certificate -> Classificação indicativa
- Runtime -> Duração
- Genre -> Gênero
- IMDB_Rating -> Taxa da Nota IMDB
- Overview -> Sinopse
- Metascore -> Metacrítica
- Director -> Diretor
- Star1 -> Estrela 1
- Star2 -> Estrela 2
- Star3 -> Estrela 3
- Star4 -> Estrela 4
- No_of_Votes -> N° de votos
- Gross -> Faturamento

No projeto, foi utilizada a linguagem Python na sua versão 3.12.11, assim como os seguintes módulos e bibliotecas:

- Pandas: Para a manipulação, limpeza e estruturação dos dados no formato de DataFrames.
- NumPy: Para a realização de operações matemáticas e manipulação de arrays numéricos.
- Matplotlib, Seaborn e Plotly: Para a criação de visualizações de dados estáticas e interativas, auxiliando na análise exploratória.
- Re (Regular Expressions): Para a limpeza e extração de padrões em colunas de texto, como "Runtime" e "Gross".
- Scikit-learn: Para diversas etapas do fluxo de machine learning.
- TfidfVectorizer para processamento de texto (NLP).

- Scipy para inferência estatística.
- MultiLabelBinarizer e OneHotEncoder para transformar variáveis categóricas.
- ColumnTransformer para aplicar transformações específicas em colunas.
- RandomizedSearchCV para otimização de parâmetros dos modelos.
- StackingRegressor para criação do modelo final.

O desenvolvimento do projeto seguiu as seguintes fases:

- Entendimento do Negócio
- Entendimento dos Dados
- Preparação dos Dados
- Análise Exploratória
- Modelagem
- Avaliação e Deploy

3. Entendimento e Preparação dos Dados

Inicialmente, foi feita a verificação dos tipos de dados e foi observado que a base de dados possuia apenas 3 colunas de tipos numéricos: "IMDB_Rating", "Meta_score" e "No_of_Votes". Pela contagem de amostras, também havia valores ausentes na coluna "Meta_score". Além disso, as notas do IMDB variam em média por volta de 7.94 em que o filme mais bem avaliado tem nota IMDB 9.2 e o menos bem avaliado tem nota 7.6. As colunas "Certificate", "Meta_score" e "Gross" possuiam valores ausentes, assim como que havia exemplos de anos inválidos na coluna "Released_Year", como é o caso do dado "PG".

As colunas "Runtime" e "Gross" são numéricas, porém seus dados são do tipo object (string), então foi feita a conversão individual de cada coluna. Para a coluna "Runtime", a unidade dos minutos foi removida e os tipos dos dados convertidos para "int64". Já a coluna "Gross" teve as vírgulas removidas e a conversão para "int64".

A classificação indicativa de um filme parece não ser um bom preditor para a sua nota IMDB (mas pode ser um fator importante para a inferência do gênero, por isso, ela não foi descartada). Entretanto, a nota média das críticas e o faturamento do filme (apesar de ser influenciado pela divulgação, pode estar correlacionado) podem influenciar a previsão da nota. Considerando a proporção média de valores ausentes (aprox. 17% dos dados faltando), o método utilizado foi a imputação da mediana dos valores nas lacunas por ser uma medida menos sensível a outliers e a imputação da moda das classificações na coluna "Certificate". Assim, todos os valores ausentes foram preenchidos.

Ademais, foi feita a exclusão das linhas com anos inválidos na coluna “Released_Year”. As colunas “Star1”...”Star4” foram agrupadas em uma única coluna “Cast”. Em seguida, foi feita a conversão da coluna para o tipo numérico (int64). Para facilitar a extração de características para a inferência do gênero de um filme, os gêneros foram armazenados em uma lista para cada filme em vez de um texto único. Para identificar as palavras mais importantes da sinopse de cada filme (referente à coluna "Overview"), foi utilizado o módulo Tfidf Vectorizer do SK Learn como uma ferramenta de NLP (Processamento de Linguagem Natural) responsável por extrair características relevantes do conteúdo de cada filme. Foi utilizada a técnica de vetorização para extrair as 10 palavras mais relevantes, ou seja, com maior frequência relativa na coluna "Overview". Isso é importante para a inferência do gênero. Essas palavras foram armazenadas em um DataFrame com o mesmo número de linhas do fornecido pela empresa e com as 10 palavras mais relevantes como colunas: “family”, “life”, “love”, “man”, “new”, “story”, “war”, “woman”, “world” e “young”. As 5 primeira instâncias da base de dados pós tratamento estão representadas na Figura 01.

	Series_Title	Released_Year	Certificate	Genre	IMDB_Rating	Overview	Meta_score	Director	No_of_Votes	Gross	Runtime (minutes)	Cast
0	The Godfather	1972	A	[Crime, Drama]	9.2	An organized crime dynasty's aging patriarch t...	100.0	Francis Ford Coppola	1620367	134966411.0	175	Marlon Brando, Al Pacino, James Caan, Diane Ke...
1	The Dark Knight	2008	UA	[Action, Crime, Drama]	9.0	When the menace known as the Joker wreaks havo...	84.0	Christopher Nolan	2303232	534858444.0	152	Christian Bale, Heath Ledger, Aaron Eckhart, M...
2	The Godfather: Part II	1974	A	[Crime, Drama]	9.0	The early life and career of Vito Corleone in ...	90.0	Francis Ford Coppola	1129952	57300000.0	202	Al Pacino, Robert De Niro, Robert Duvall, Dian...
3	12 Angry Men	1957	U	[Crime, Drama]	9.0	A jury holdout attempts to prevent a miscarria...	96.0	Sidney Lumet	689845	4360000.0	96	Henry Fonda, Lee J. Cobb, Martin Balsam, John ...
4	The Lord of the Rings: The Return of the King	2003	U	[Action, Adventure, Drama]	8.9	Gandalf and Aragorn lead the World of Men agai...	94.0	Peter Jackson	1642758	377845905.0	201	Elijah Wood, Viggo Mortensen, Ian McKellen, Or...

Figura 01 - Base de dados após tratamento

4. Análise Exploratória

Como uma etapa de compreensão dos dados pós-tratamento, foi realizada uma análise estatística geral dos atributos, antes de dividir a base para cada objetivo específico. Observou-se que os filmes da base datam de 1920 até 2020 em seus anos de lançamento, possuem duração média de 122 minutos com desvio padrão de 28 minutos e faturamento de, em média, 122 milhões.

Ademais, foi realizada uma análise univariada dos atributos numéricos para verificar suas respectivas distribuições. Para isso, foram plotados histogramas e boxplots de cada atributo numérico. Os gráficos podem ser observados nas Figuras 02 e 03.

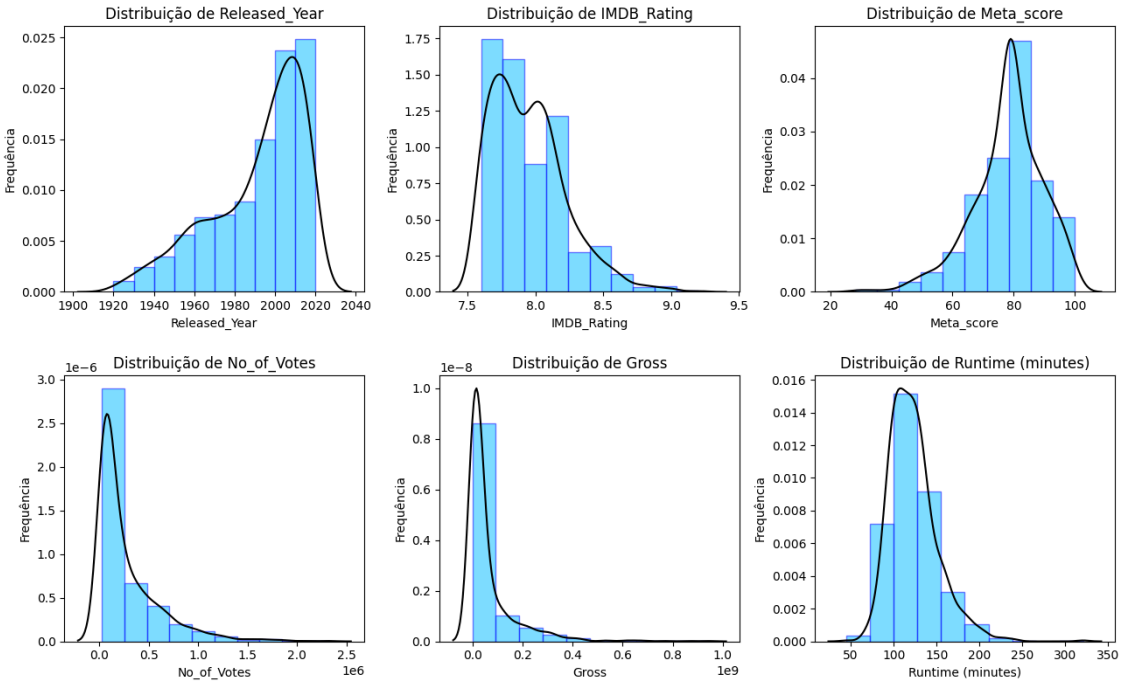

Figura 02 - Histogramas dos atributos numéricos

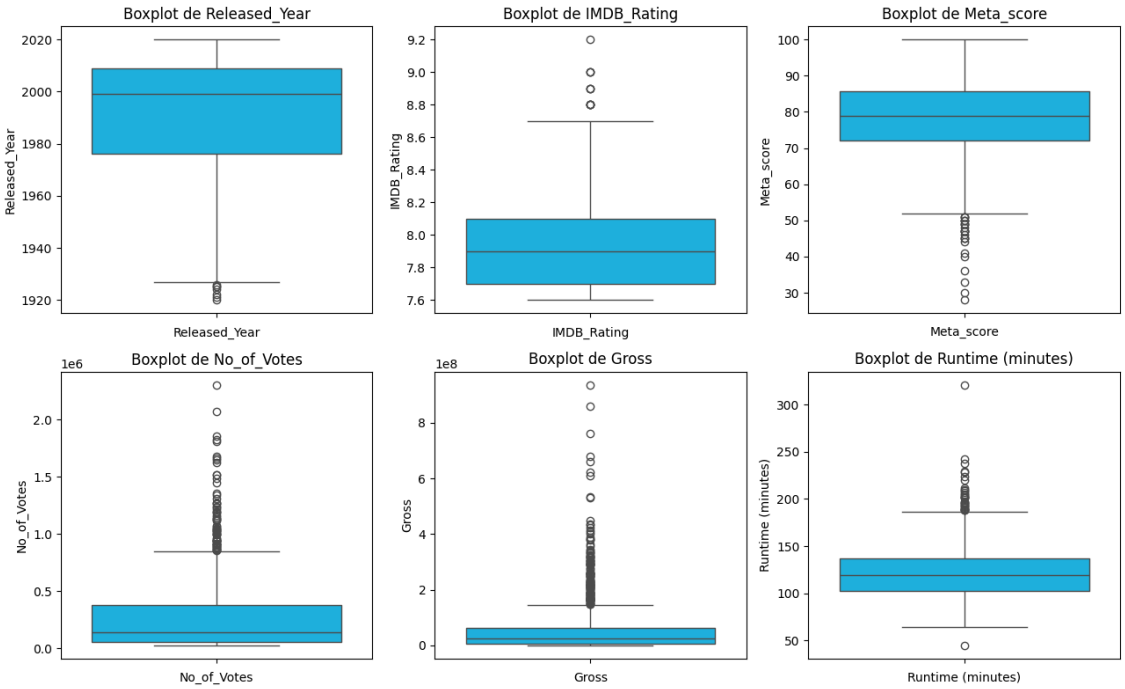

Figura 03 - Boxplots dos atributos numéricos

Observando os gráficos, percebe-se que todos os atributos numéricos possuem uma distribuição assimétrica. A distribuição da nota do IMDB é distorcida à direita, indicando que a maioria dos filmes possui nota entre 7.5 e 8.0, com poucos filmes variando em 9.0. A distribuição dos anos de lançamento é distorcida à esquerda, indicando que a maioria dos filmes são recentes, sendo lançados entre 2000 e 2020. A distribuição da duração dos filmes indica que a maioria deles têm entre 100 e 150 minutos.

Os atributos categóricos “Certificate”, “Genre”, “Director” e “Cast” também foram analisados por meio de gráficos de barra a fim de verificar suas distribuições de frequência. Dessa forma, foi observado que a maioria dos filmes têm classificação "U", referente à "Livre para todos os públicos", os 5 gêneros mais frequentes presentes nos filmes são: Drama, Comédia, Crime, Aventura e Ação; Alfred Hitchcock é o diretor mais frequente, tendo dirigido 14 filmes da base e Roberto de Niro, Al Pacino e Tom Hanks são os atores mais frequentes no elenco dos filmes. Essas observações são indicadas pelos gráficos abaixo.

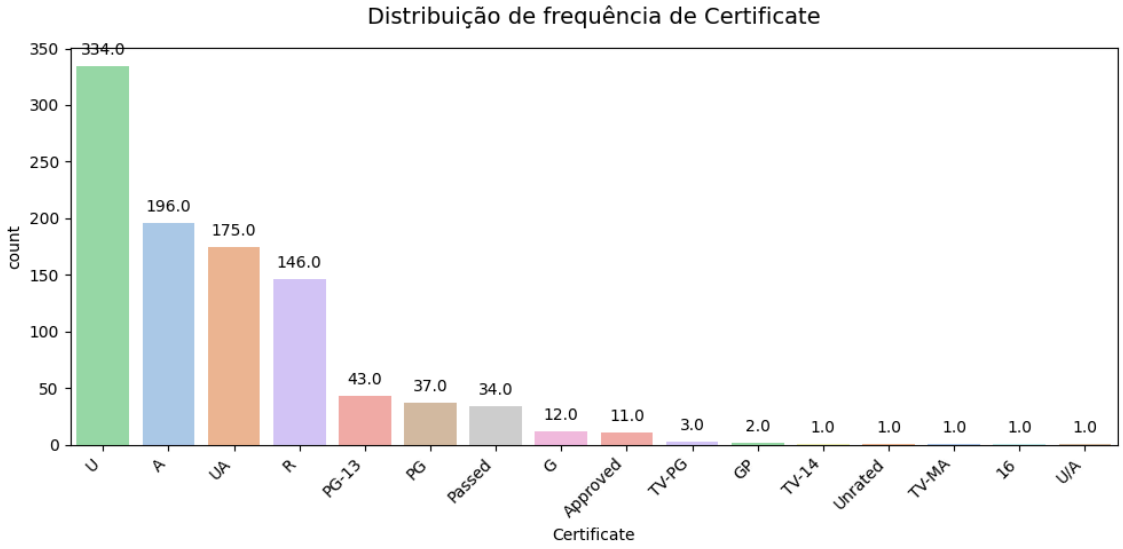

Figura 04 - Distribuição de frequência de “Certificate”

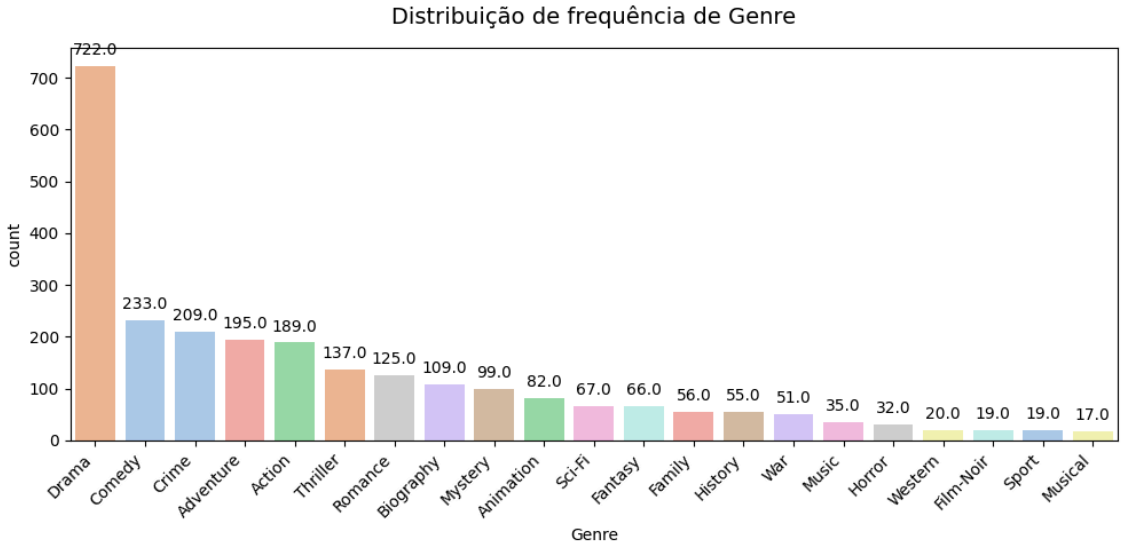

Figura 05 - Distribuição de frequência de “Genre”

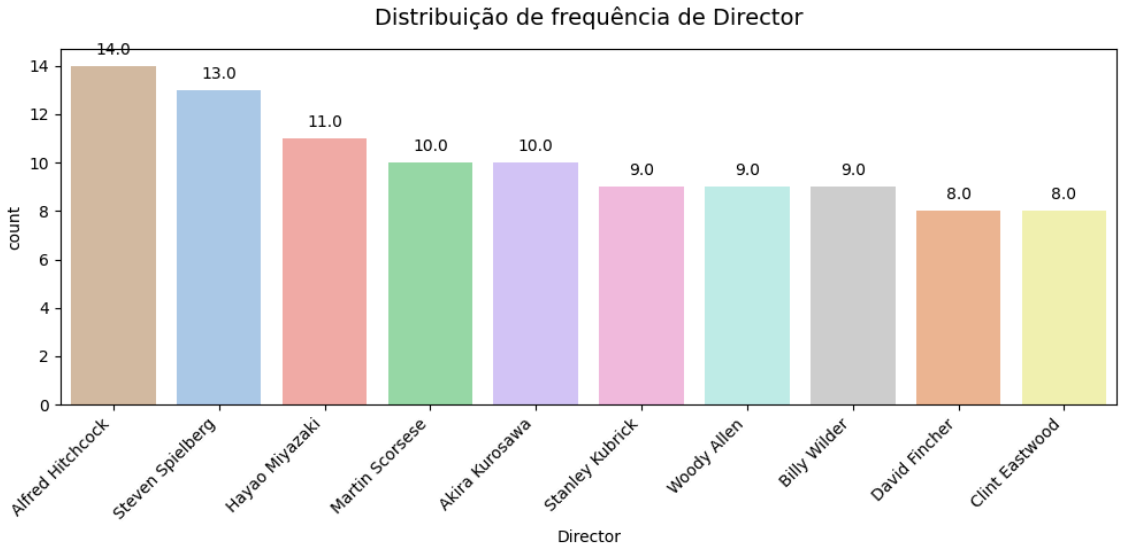

Figura 06 - Distribuição de frequência de “Director”

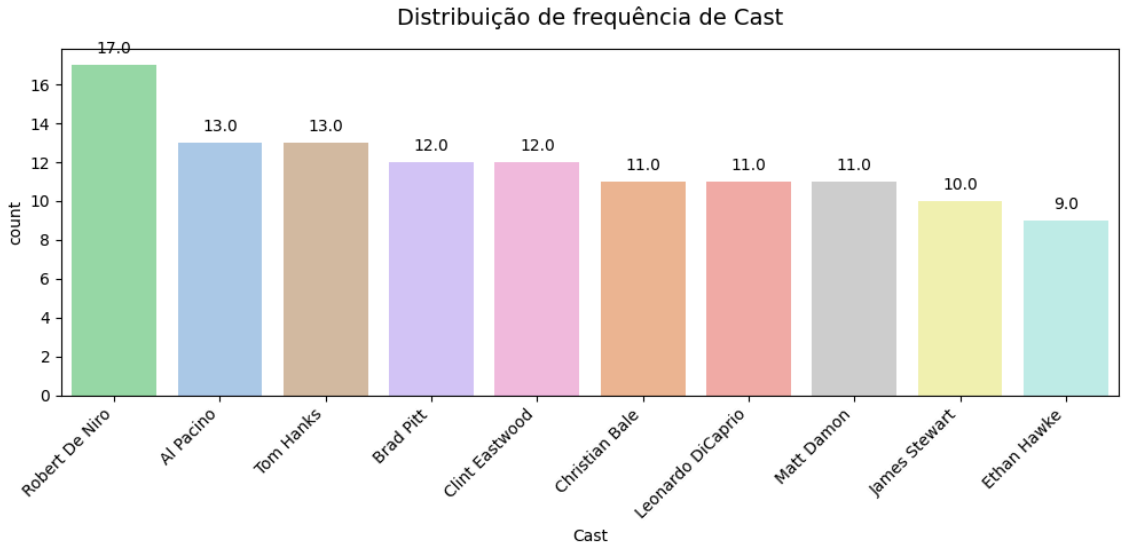

Figura 07 - Distribuição de frequência de “Cast”

Também foi verificada a distribuição de frequência das 10 palavras mais frequentes na sinopse dos filmes. Essa distribuição pode ser verificada na Figura 08:

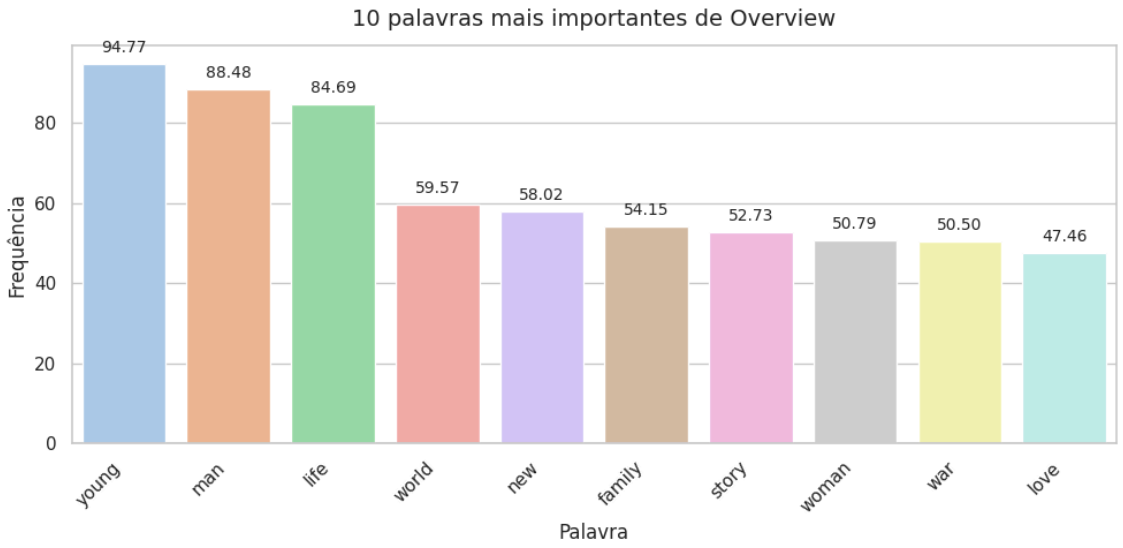

Figura 08 - 10 palavras mais importantes de “Overview”

Também foram plotadas nuvens de palavras a fim de explorar a coluna “Overview” dos filmes. A nuvem de palavras mostra que as palavras mais frequentes nas sinopses dos filmes são "life", "find", "two" e "man". Observando os dois gráficos abaixo, eles podem indicar uma recorrência de filmes centrados em protagonistas jovens, muitas vezes do sexo masculino, e em temas universais ligados à vida e à experiência humana.

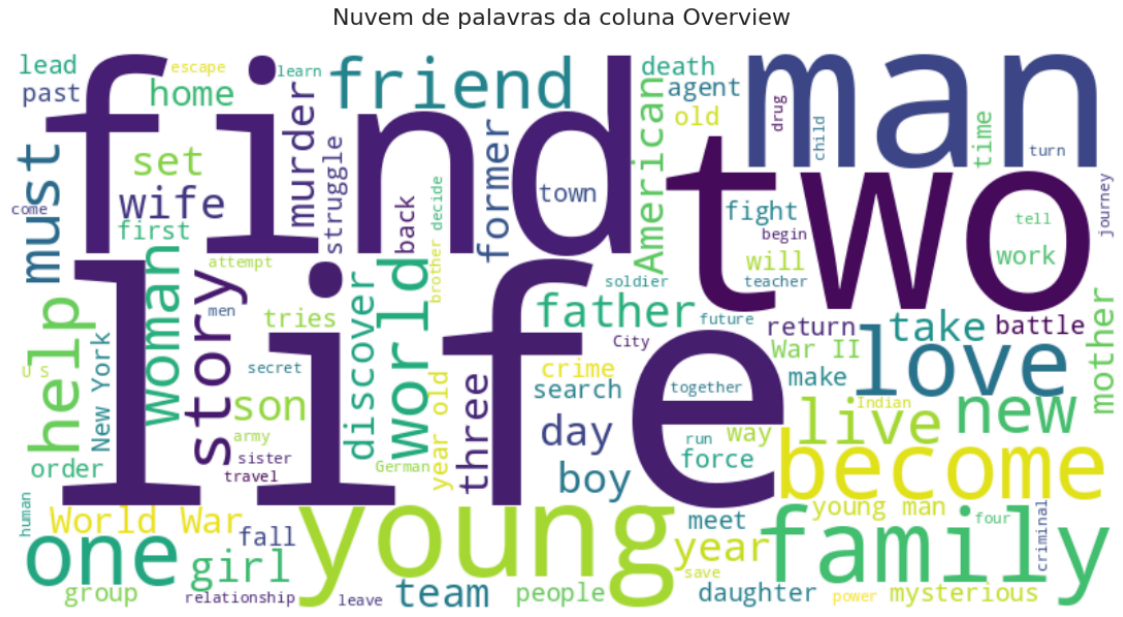

Figura 09 - Nuvem de palavras da coluna “Overview”

Fazendo uma análise por gênero, foram plotadas 3 nuvens de palavras para os gêneros mais frequentes na base de dados: Drama, Comédia e Crime, como observado nas figuras abaixos:

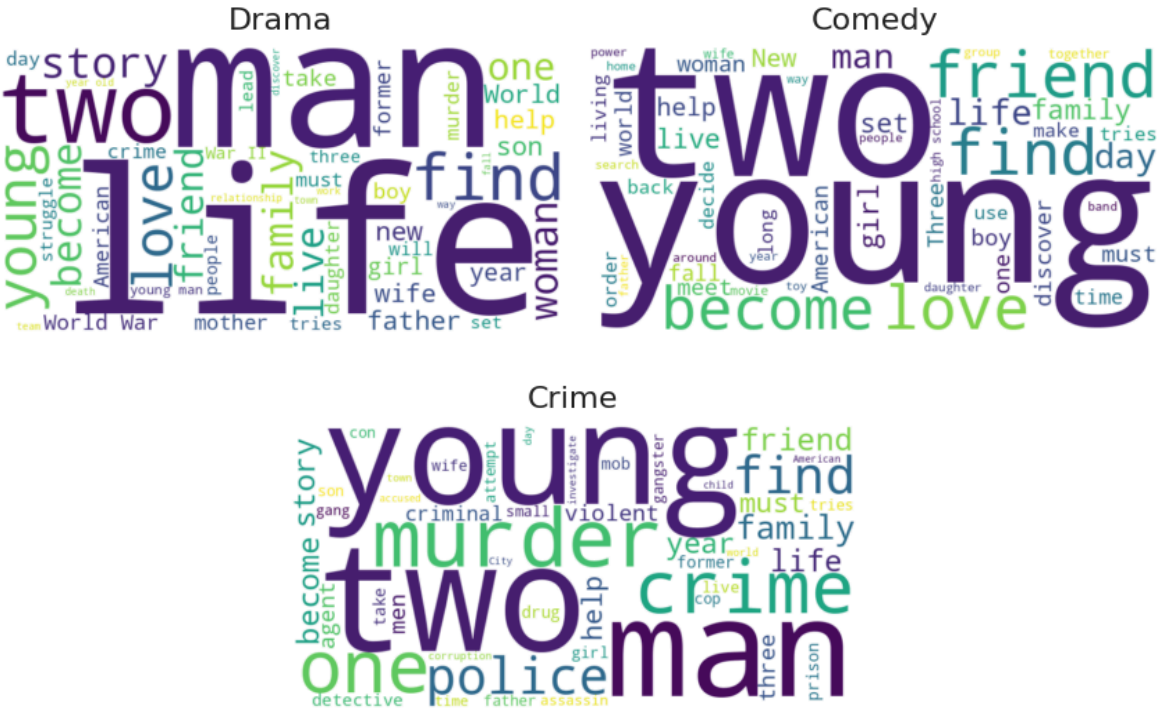

Figura 10 - Nuvens de palavras dos 3 gêneros mais frequentes

Observando as nuvens de palavras, é possível ver que as palavras "man" e "life" são mais comuns na descrição de filmes de drama, enquanto as palavras "two" e "young" são mais comuns em filmes de comédia e crime. Além disso, foi selecionada a palavra mais importante para a sinopse dos filmes de um gênero específico com base na vetorização que foi feita anteriormente. Observou-se que o termo "young" (jovem) predomina em gêneros como Action, Drama, Biography e Sci-Fi, sugerindo uma forte associação com narrativas de formação pessoal e histórias de origem. Já a palavra "man" (homem) é recorrente em Comedy, Crime, Horror e Romance, indicando protagonistas masculinos centrais nestas narrativas. Essas características indicam que é possível encontrar padrões de palavras em um gênero específico, trazendo a possibilidade de inferir o gênero de um filme pela sua descrição.

Tendo feito uma análise univariada (atributo por atributo), também foi realizada uma análise bivariada (comparando dois atributos por vez). Ao observar a matriz de correlação

abaixo, é perceptível que o número de votos de um filme é consideravelmente correlacionado com o seu faturamento, ou seja, filmes que arrecadam mais tendem a receber mais votos, logo, sendo melhor avaliados. O número de votos também é relacionado de forma moderada com a sua nota do IMDB.

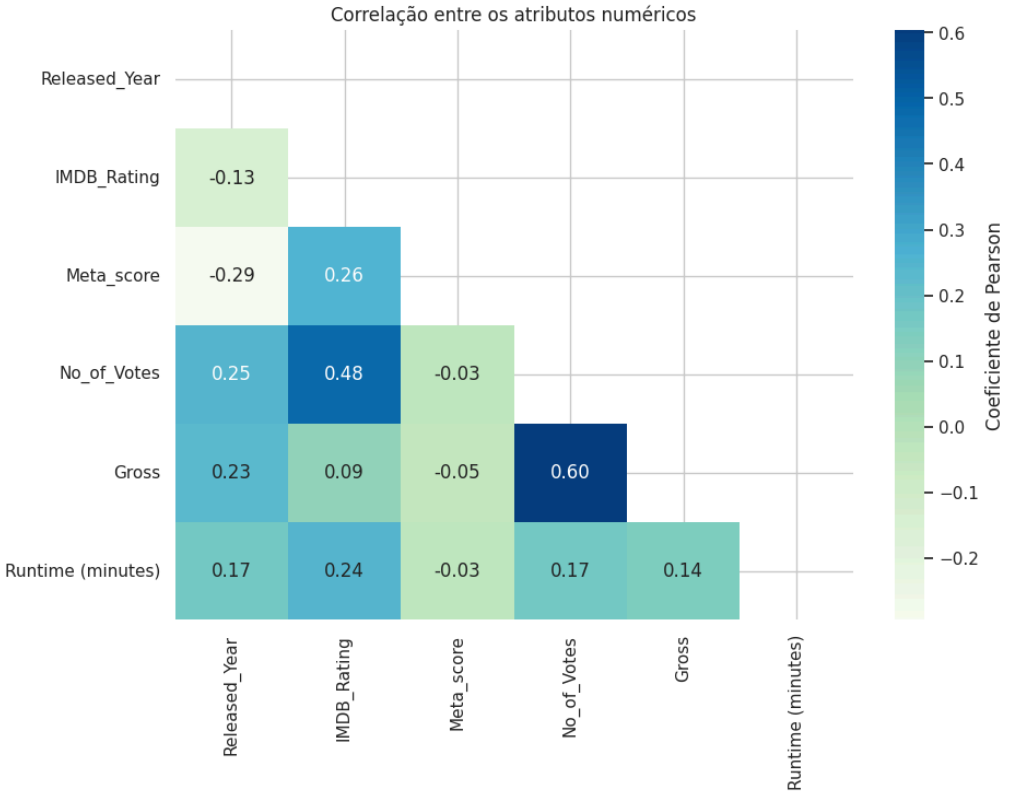

Figura 11 - Correlação entre os atributos numéricos

Ademais, foram gerados gráficos de dispersão dos atributos numéricos combinados dois a dois. Observando os gráficos abaixo, apesar de uma correlação diferente das observadas na matriz de correlação não ser perceptível, tem-se uma visão mais clara da correlação entre o faturamento de um filme e sua nota do IMDB.

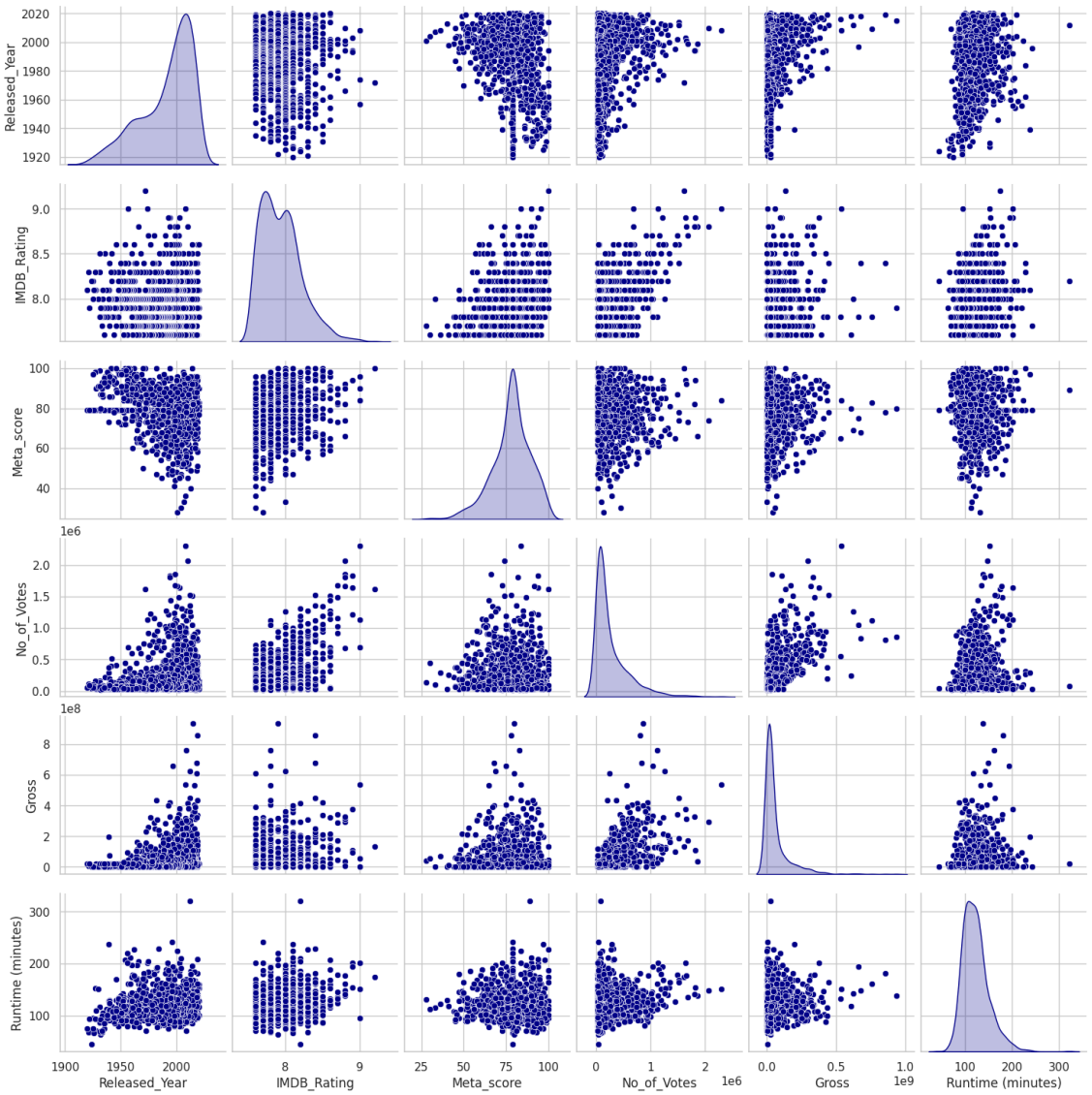

Figura 12 - Relação gráfica dos atributos numéricos dois a dois

Sabendo dessas informações, a recomendação de um filme para uma pessoa desconhecida, sem conhecer suas preferências pessoais, se torna mais viável. Um critério válido é escolher um filme popular e com maior relevância para o público (com a maior nota do IMDB ou com um número de votos considerável). Dessa forma, para obter o máximo de confiabilidade, foram realizadas análises inferenciais. Foi realizado um teste de hipótese sobre a média da nota IMDB de um filme para verificar se os filmes mais bem avaliados têm média de notas significativamente maior do que os demais. Isso foi feito pelo teste t de Student para comparação de médias.

Estabelecendo a hipótese nula de que os 10% dos filmes mais bem avaliados não possuem notas IMDBs médias significativamente maiores que os demais, as médias de nota foram calculadas e a média dos filmes mais bem avaliados foi de 8.46, enquanto que a média dos restantes foi de 7.87 com o p-valor tendendo a zero. Dessa forma, pode-se rejeitar a hipótese nula de que as médias são iguais. Portanto, isso sustenta a hipótese de que a recomendação de filmes com maior nota IMDB, são mais seguros para agradar uma pessoa desconhecida.

Entretanto, outro atributo relevante é o número de votos. É mais provável que filmes populares sejam mais bem vistos por pessoas desconhecidas. Nesse caso, é relevante considerar esse atributo também. Foi calculada uma nota ponderada com base no número de votos de um filme e na sua nota IMDB normalizadas, tendo essa última um peso maior, por maior confiabilidade como observado no teste t de Student. As fórmulas em questão são a representadas abaixo:

$$taxaIMDB = \frac{notaIMDB - IMDBmenor}{IMDBmaior - IMDBmenor}$$

$$taxaVotos = \frac{numeroVotos - menorNumeroVotos}{maiorNumeroVotos - menorNumeroVotos}$$

$$notaRecomendação = 2taxaIMDB + taxaVotos$$

Dessa forma, com nota de recomendação 9.0 e 2.303.232 votos, o melhor filme para recomendar a uma pessoa desconhecida é “The Dark Night” (“Batman O Cavaleiro das Trevas”).

Para identificar os principais fatores que estão relacionados com alta expectativa de faturamento de um filme, é possível verificar quais variáveis se relacionam significativamente com o faturamento. Para isso, foi calculada a correlação de Pearson entre o faturamento e os outros atributos numéricos.

	Gross
Gross	1.000000
No_of_Votes	0.603091
Released_Year	0.232659
Runtime (minutes)	0.138057
IMDB_Rating	0.089700
Meta_score	-0.051787

Figura 13 - Coeficiente de correlação de Pearson entre os atributos numéricos e o faturamento.

Dessa forma, o número de votos segue o fator mais relacionado com o faturamento de um filme. Para identificar a relação com os atributos categóricos, foi realizado o teste ANOVA (análise de variância) sobre eles. Esse teste foi escolhido visando o objetivo de comparar a média do faturamento nos grupos de gênero e classificação indicativa dos filmes, visto que são os atributos que possuem menos classes. Assim, foi definida a hipótese nula de que as médias do faturamento entre as categorias de gênero são iguais. O objetivo é verificar se a média do faturamento difere significativamente entre as categorias gênero e classificação indicativa. Os resultados podem ser observados abaixo, sendo "sum_sq" a soma dos quadrados e "PR(>F)" o p-valor:

	sum_sq	df	F	PR(>F)
C(Genre)	3.460732e+18	20.0	17.433305	1.732284e-57
Residual	2.497289e+19	2516.0	NaN	NaN

Figura 14 - Resultados do teste ANOVA para os gêneros

	sum_sq	df	F	PR(>F)
C(Certificate)	3.109592e+18	15.0	20.637265	1.813788e-53
Residual	2.532403e+19	2521.0	NaN	NaN

Figura 15 - Resultados do teste ANOVA para as classificações indicativas

Como o p-valor é menor que 0.05 para ambas as categorias Gênero e Classificação Indicativa, rejeita-se a hipótese nula e podemos afirmar que são fatores importantes que influenciam a expectativa de faturamento. No gráfico abaixo, pode-se observar a diferença das médias do faturamento em cada categoria em ambos os atributos:

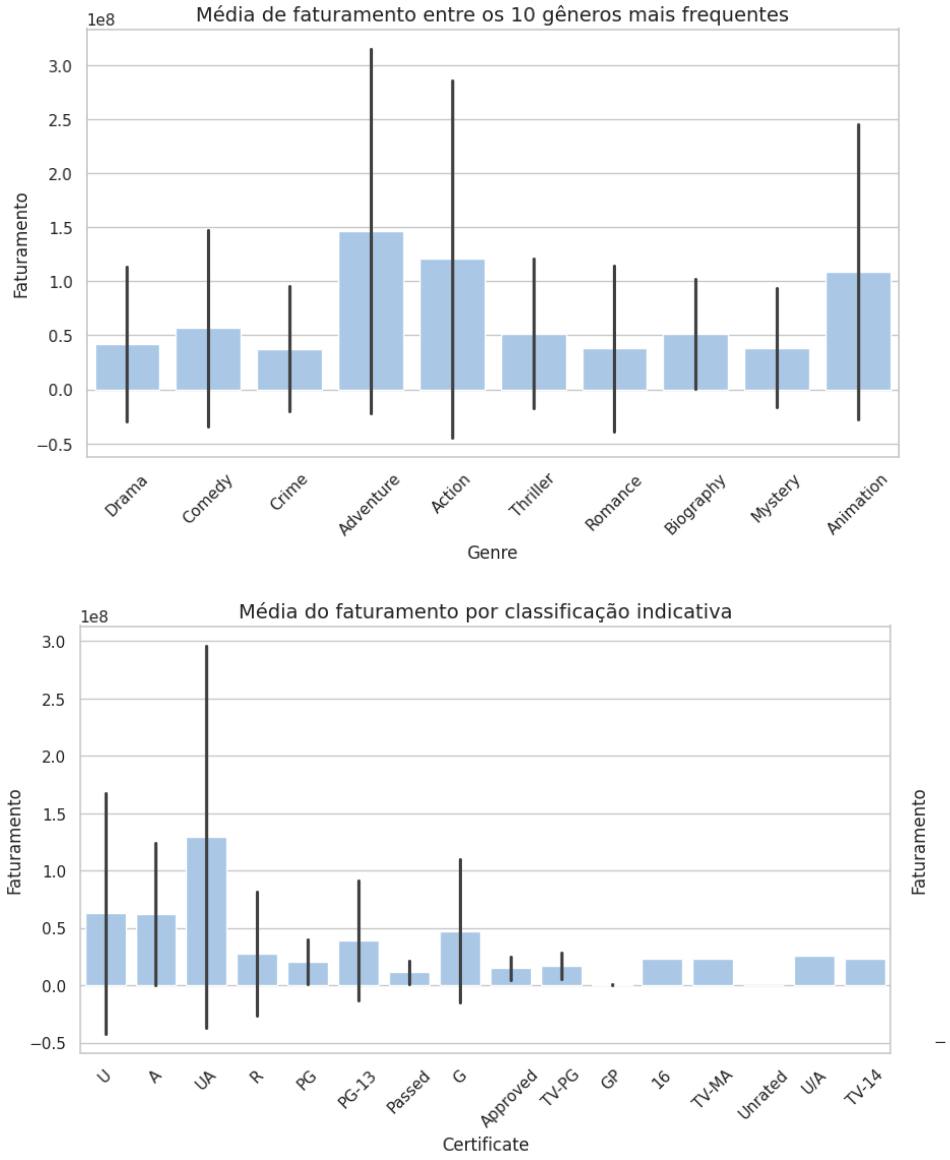

Figura 16 - Representação gráfica das médias de faturamento entre os gêneros e classificações indicativas.

Portanto, como observado pela análise de correlação, os atributos: número de votos e ano de lançamento são os mais relacionados significativamente com o faturamento. Da mesma forma, o teste ANOVA indicou que o gênero e a classificação indicativa também influenciam significativamente a expectativa de faturamento. Assim, tanto o engajamento do público, quanto o ano de lançamento, tipo de filme e classificação indicativa estão entre os fatores mais relevantes. Uma observação interessante é que média ponderada de todas as críticas ("meta_score") possui uma correlação negativa com o faturamento. Uma possível razão para isso é a curiosidade que o público tem de conferir um filme que recebeu críticas negativas, ocasionando na popularidade das vendas de ingresso, aumentando o faturamento consequentemente.

5. Modelagem

Na etapa de modelagem, dois cenários serão abordados: a previsão dos filmes antes do lançamento e depois. Com os dados tratados, é importante definir quais atributos serão relevantes para cada modelo. Para o modelo preditor da nota do IMDB de um filme antes do seu lançamento, não é viável incluir as colunas "Gross" e "No_of_Votes", já que esses dados só existem após a estreia de um filme e incluí-los na base causaria vazamento de informações (data leakage). Para o modelo preditor da nota do IMDB de um filme após o seu lançamento, todos os atributos podem ser considerados.

5.1 - Previsão da nota IMDb após do lançamento de um filme

Inicialmente, foi necessário realizar o pré-processamento dos atributos preditores utilizados no regressor. Os atributos numéricos "Released_Year", "Runtime (minutes)", "Meta_Score", "No_of_Votes" e "Gross" foram mantidos brutos sem normalização. O atributo categórico"Certificate" foi codificado utilizando One-Hot-Encoding e os atributos "Genre", "Cast" e "Director" foram codificados utilizando Multi-Label Encoding, pois possuem mais rótulos. Para simplificar o processamento, foram utilizados os 10 gêneros, atores e diretores mais frequentes. As 10 palavras mais importantes em frequência relativas obtidas anteriormente foram utilizadas como atributos no modelo também.

Como o objetivo é prever a nota IMDB que é um dado contínuo, o problema em questão é de regressão supervisionada. Dessa forma, foram escolhidos 9 modelos candidatos de regressão, sendo 5 deles lineares: Regressão Linear, Ridge, Lasso e Elastic Net por serem mais simples e interpretáveis e 4 baseados em árvores: Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, XGBoost e CatBoost Regressor para capturarem relações mais complexas. Os modelos foram treinados sem ajustes de hiperparâmetros e depois comparados utilizando métricas de regressão. Como método de validação cruzada, foi utilizado K-Fold com 10 folds e as métricas de avaliação RMSE (Root Mean Square Error), MAE (Mean Absolute Error) e R2 (Coeficiente de Determinação). O RMSE foi escolhido para destacar impactos de erros graves, MAE para entender o erro médio esperado e R^2 para contextualizar a qualidade do modelo em relação a um baseline simples. Após o primeiro treinamento, os resultados obtidos foram os seguintes:

	model	rmse_mean	rmse_std	mae_mean	mae_std	r2_mean	r2_std
6	GradientBoosting	0.180501	0.010149	0.144467	0.007941	0.531308	0.106185
8	CatBoost	0.180914	0.013356	0.143699	0.008832	0.530530	0.103468
5	RandomForest	0.184347	0.011161	0.147628	0.007628	0.512512	0.101791
7	XGBoost	0.191396	0.009488	0.152953	0.007759	0.472397	0.118998
1	Ridge	0.202124	0.008575	0.160508	0.008163	0.413006	0.129046
0	Linear	0.202753	0.008964	0.161078	0.008019	0.409673	0.128515
3	ElasticNet	0.212550	0.011724	0.172400	0.010886	0.351094	0.145662
2	Lasso	0.222573	0.011546	0.182181	0.011884	0.288163	0.160752
4	DecisionTree	0.258449	0.018164	0.199385	0.018027	0.043158	0.214306

Figura 17 - Ranking de desempenho dos modelos candidatos sem o ajuste dos parâmetros

Observando os resultados, percebe-se que os modelos com melhor performance foram CatBoost, Random Forest e GradientBoosting. XGBoost e Ridge tiveram um desempenho moderado para baixo. Os modelos lineares não obtiveram bons resultados. Com o intuito de ajustar os parâmetros desses 4 melhores modelos, foi realizado uma busca com RandomizedSearchCV considerando seu custo computacional menor e uma quantidade menor de parâmetros. Assim, CatBoost, RandomForest, GradientBoosting e XGBoost tiveram seus parâmetros ajustados e foram treinados novamente utilizando KFold com 10 splits. Os resultados após o ajuste fino podem ser observados abaixo:

	modelo	rmse_medio	rmse_std	mae_medio	r2_medio
0	CatBoost	0.178152	0.010505	0.142144	0.542110
2	GradientBoost	0.187885	0.011375	0.150517	0.494186
3	XGBoost	0.191541	0.011633	0.154769	0.475676
1	RandomForest	0.205821	0.012507	0.167204	0.399947

Figura 18 - Ranking de desempenho dos melhores modelos após o ajuste dos parâmetros

Com os novos resultados, pode-se notar que o CatBoost seguiu com a melhor performance e seus resultados não mudaram muito com os novos parâmetros. O modelo Random Forest, mesmo com os novos parâmetros, ainda teve o pior desempenho, portanto,

não foi considerado no modelo final. Finalmente, os modelos serão combinados através do método Stacking para agregar no desempenho das predições. O metamodelo simples escolhido foi o Ridge devido ao controle de overfitting da sua regularização L2 e maior estabilidade numérica. O modelo final foi formado a partir do CatBoost, XGBoost e GradientBoosting.

Por fim, o modelo final foi salvo como “stacking_reg.pkl” e testado inicialmente com as notas IMDB já conhecidas. A tabela abaixo mostra a comparação entre as notas reais e as notas preditas pelo modelo:

	Titulo	Valor Real	Valor Previsto
0	The Godfather	9.2	8.427209
1	The Dark Knight	9.0	8.452012
2	The Godfather: Part II	9.0	8.362055
3	12 Angry Men	9.0	8.385568
4	The Lord of the Rings: The Return of the King	8.9	8.259949
...	...	...	...
993	A Hard Day's Night	7.6	7.853207
994	Breakfast at Tiffany's	7.6	7.713533
995	Giant	7.6	7.910589
996	From Here to Eternity	7.6	7.769472
997	Lifeboat	7.6	7.705789

Figura 19 - Comparação entre as notas reais e as notas preditas

Embora com uma margem de erro de aproximadamente 0.5-0.8, é notável que o modelo de stacking combinando CatBoost, XGBoost e Gradient Boosting obteve um bom desempenho na previsão das notas, visto que as previsões são consistentemente próximas dos valores reais. É relevante testar o modelo com dados de teste que não estão na base de dados, para isso, foram usados exemplos de 4 novos filmes: “Parasite”, “Joker”, “La La Land”, “The Shape of Water” e o requisitado pelo cliente “The Shawshank Redemption”, este último com os seguintes atributos:

{

'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']

}

Assim, o modelo já treinado realizou as predições e obteve os seguintes resultados com os dados não vistos:

	Título	Nota Prevista
0	The Shawshank Redemption	8.312107
1	Parasite	8.422018
2	Joker	8.281824
3	La La Land	7.881096
4	The Shape of Water	7.973118

Figura 20 - Comparação entre as notas dos dados não vistos e a nota prevista pelo modelo

Os resultados previstos pelo modelo são bem próximos das notas IMDB reais, indicando uma boa capacidade preditiva. O filme “The Shawshank Redemption” tem nota 9.3 e foi subestimado em mais de 1 ponto, o que sugere que o modelo tende a ser conservador em filmes com avaliações muito altas. Entretanto, o filme “Parasite”, com nota 8.6, Joker, com nota 8.4, “La La Land”, com nota 8.0 e “The Shape of Water”, com nota 7.3, tiveram previsões bem alinhadas com as notas reais, apresentando diferenças inferiores a 0.7. Isso demonstra que, apesar de certa limitação em capturar outliers de altíssimo desempenho, o modelo consegue manter previsões consistentes e próximas da realidade para a maioria dos casos.

5.2 - Previsão da nota IMDb antes do lançamento de um filme

Para essa abordagem, será feita a predição do desempenho de um filme antes do lançamento, quando o número de votos e o faturamente ainda são desconhecidos. Para isso, um novo modelo de pré-lançamento foi treinado de forma similar ao primeiro cenário. Também foi utilizada a técnica de Stacking com os melhores modelos do primeiro cenário. (CatBoost, XGBoost e GradientBoosting). As mesmas métricas e mesmas técnicas de pré-processamento foram utilizadas nos dados e no modelo. O comparativo das métricas de avaliação de ambos os cenários podem ser observados abaixo:

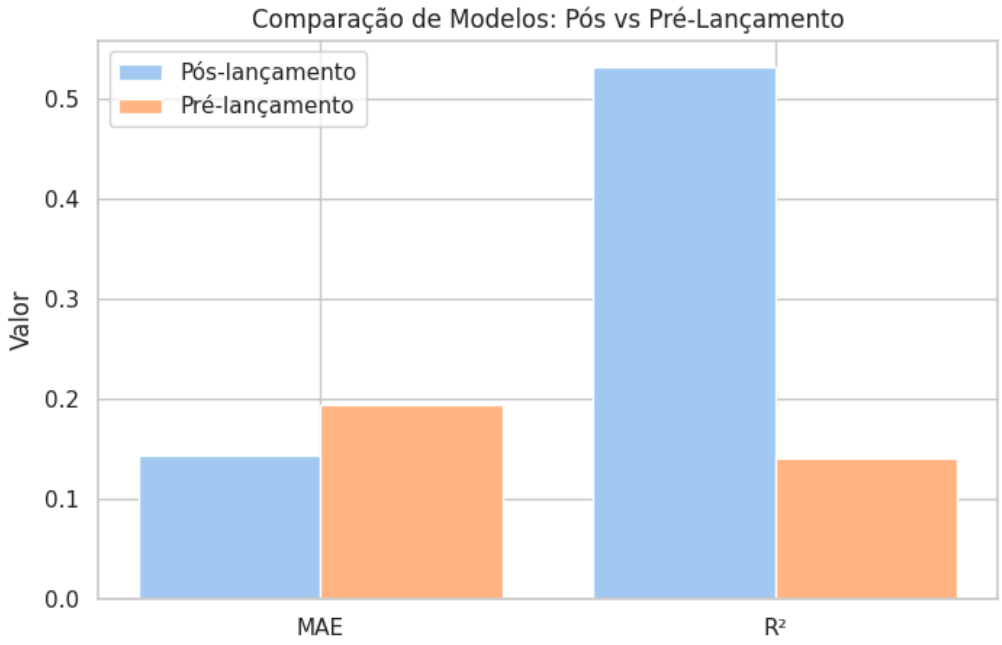

Figura 21 - Comparação de performance dos modelos pré e pós lançamento de um filme

Os resultados evidenciam uma diferença marcante entre os dois cenários avaliados: enquanto o modelo pós-lançamento apresentou desempenho aceitável, o modelo pré-lançamento obteve métricas bem inferiores, indicando baixa capacidade preditiva quando variáveis cruciais como número de votos e bilheteria ainda não estão disponíveis. Apesar de não ter alcançado o resultado esperado no contexto pré-lançamento, o projeto abre espaço para melhorias futuras, como o uso de técnicas mais avançadas de NLP para sinopses ou variáveis externas que possam enriquecer as previsões antes da estreia de um filme.

6. Conclusão

O projeto seguiu de forma estruturada todas as etapas do ciclo CRISP-DM, começando pelo entendimento do negócio, que definiu como objetivo prever as notas de filmes no IMDb para apoiar análises pré e pós-lançamento. Em seguida, o entendimento dos dados permitiu mapear as variáveis mais relevantes, como gênero, elenco, direção, certificação e atributos numéricos.

Na etapa de preparação, os dados foram tratados com as técnicas one-hot encoding, binarização de variáveis multilabel e seleção de palavras-chave, resultando em um conjunto consistente para modelagem.

A fase de modelagem testou diferentes algoritmos, culminando na construção de um modelo de stacking com CatBoost, XGBoost e Gradient Boosting, que apresentou excelente desempenho no cenário pós-lançamento. Foi observado tanto o sucesso nesse caso quanto as limitações em predições pré-lançamento, reforçando a importância de dados adicionais.

Por fim, a fase de implantação foi simulada com novos exemplos, demonstrando a aplicabilidade do modelo. Assim, o ciclo CRISP-DM foi concluído de forma prática, deixando espaço para avanços futuros, sobretudo na predição de filmes ainda não lançados.


indicium