RELATÓRIO DESAFIO – CIENTISTA DE DADOS

CANDIDATO: Clériston Cláudio Carneiro Pereira de Albuquerque

RECIFE, 2025.

1. INTRODUÇÃO

1.1 Desafio

O desafio faz parte do programa Lighthouse da Indicium e tem como objetivo promover a imersão em projetos reais. O desafio proposto para a área de ciência de dados consiste em:

Você foi alocado em um time da Indicium contratado por um estúdio de Hollywood chamado *PProductions,* e agora deve fazer uma análise em cima de um banco de dados cinematográfico para orientar qual tipo de filme deve ser o próximo a ser desenvolvido. Lembre-se que há muito dinheiro envolvido, então a análise deve ser muito detalhada e levar em consideração o máximo de fatores possíveis (a introdução de dados externos é permitida - e encorajada).

1.2 Objetivos da análise

Os objetivos propostos pelo desafio são:

- Apresentar os principais insights na análise exploratória dos dados;
- Propor um método de recomendação de filmes para novo usuário que vai utilizar alguma plataforma streaming pela primeira vez;
- Identificar os principais fatores que estão relacionados com alta expectativa de faturamento;
- Propor modelo de previsão de notas IMDb;
- Propor insights na visão do filme;
- Propor o modelo de inferência entre a coluna visão geral do filme e do gênero.

2. ENTENDIMENTO DOS DADOS

A base de dados conta com 999 linhas e 15 colunas, a maioria das colunas são do tipo “String” e apenas três do tipo “Int” e “Float”. A figura a seguir retrata a visão geral dos dados.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 999 entries, 0 to 998
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Unnamed: 0     999 non-null    int64
 1   Series_Title   999 non-null    object
 2   Released_Year  999 non-null    object
 3   Certificate    898 non-null    object
 4   Runtime        999 non-null    object
 5   Genre          999 non-null    object
 6   IMDB_Rating    999 non-null    float64
 7   Overview       999 non-null    object
 8   Meta_score     842 non-null    float64
 9   Director       999 non-null    object
 10  Star1          999 non-null    object
 11  Star2          999 non-null    object
 12  Star3          999 non-null    object
 13  Star4          999 non-null    object
 14  No_of_Votes    999 non-null    int64
 15  Gross          830 non-null    object
dtypes: float64(2), int64(2), object(12)
memory usage: 125.0+ KB
```

Figura 01 – Visão geral dos dados

Fonte: Próprio autor

As colunas, “Released_Year” (ano de lançamento) e “Gross” (faturamento) estão do tipo string e foram convertidas para inteiro e ponto flutuante. A coluna ano de lançamento havia um dado errado, foi corrigido inserindo o ano correto do filme Apollo 13.

Foi identificado dados faltantes nas colunas: “Certificate” (faixa etária), “Meta_score” (número de críticas) e “Gross” (faturamento). A solução encontrada para preencher os dados faltantes nas colunas “Meta_score” e “Gross” foi obter a média por período, assim, preservando o número de críticas e faturamento para cada período. Já a coluna “Certificate” (faixa etária) foi utilizada como base coluna “Genre” (gênero), uma vez que, o gênero é um fator limitante para determinada faixa etária.

Foi obtido da base do FED (Federal Reserve) os dados do PIB (GDP) e inflação (CPI) com finalidade de realizar inferência com a coluna “Gross” (faturamento) da base de dados IMDb.

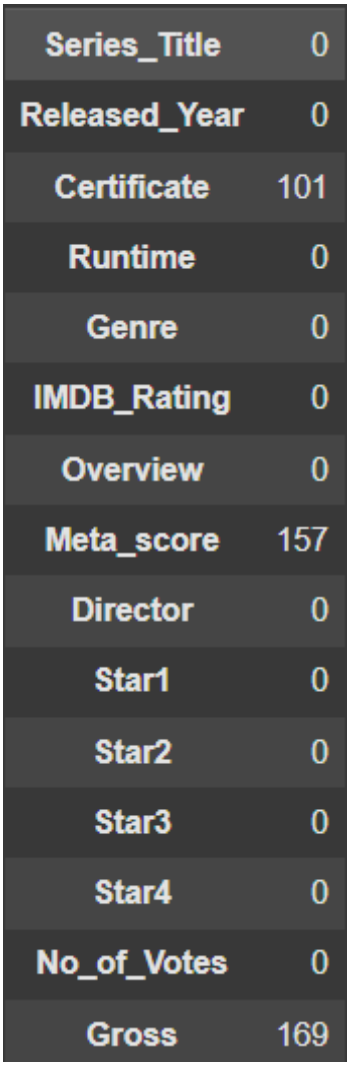

Figura 02 - Quantidade de dados faltantes por coluna

Fonte: Próprio autor

3. ANÁLISE EXPLORATÓRIA DE DADOS

3.1 Estatística descritiva

Foi empregado análise descritiva dos dados numéricos, conforme imagem apresentada abaixo:

	Released_Year	IMDB_Rating	Meta_score	No_of_Votes	Gross
count	999.00	999.00	999.00	999.00	9.990000e+02
mean	1991.19	7.95	77.98	271621.42	6.551194e+07
std	23.31	0.27	13.28	320912.62	1.022390e+08
min	1920.00	7.60	0.00	25088.00	0.000000e+00
25%	1976.00	7.70	72.00	55471.50	4.009348e+06
50%	1999.00	7.90	79.00	138356.00	2.694762e+07
75%	2009.00	8.10	87.00	373167.50	8.353862e+07
max	2020.00	9.20	100.00	2303232.00	9.366622e+08

Figura 03 – Estatística descritiva

Fonte: Próprio autor

Percebe-se claramente que as colunas “Gross” (faturamento) e “No_of_Votes” (número de votos) possuem desvio padrão alto, evidenciando alta variação dos dados.

Foi calculado o coeficiente de variação na variável “Gross” (faturamento), o coeficiente de variação foi de 156,06%, sinalizando que há maior dispersão dos dados, em outras palavras, os dados são heterogêneos.

3.2 Correlações

Antes de explicar a correlação, foi plotado o gráfico de barras da coluna “Gross” (faturamento em relação ao “Releazed_Year” (ano de lançamento) e verificar a evolução do faturamento em relação aos anos de lançamento.

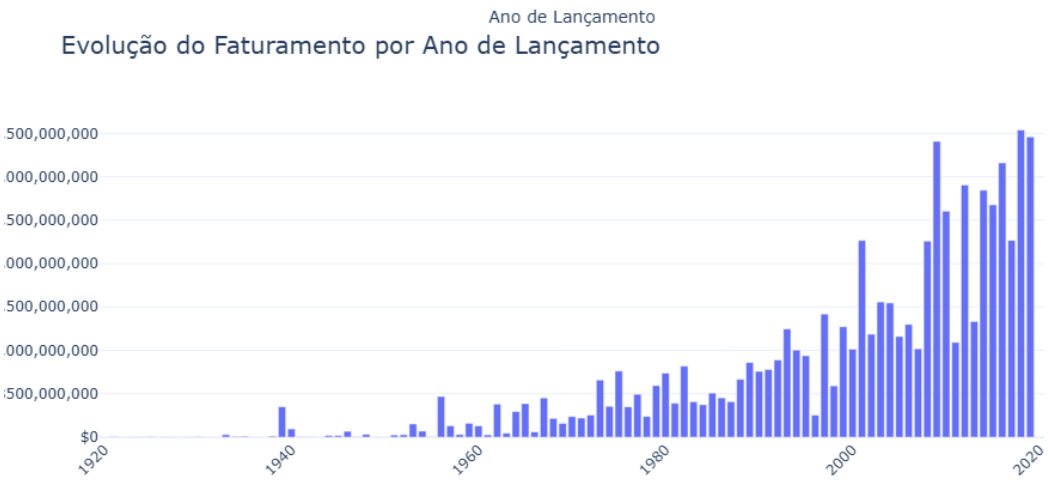

Figura 04 – Evolução do faturamento ao longo dos anos de lançamento

Fonte: Próprio autor

O gráfico evidencia que nos últimos anos, houve um aumento expressivo no faturamento, porém em determinados períodos, o faturamento apresentou queda, possivelmente pode está associado à eventos externos como por exemplo, o aumento da inflação nos Estados Unidos. Para investigar essa relação, foi empregada uma análise de correlação considerando os dados do PIB e da inflação norte-americana.

	Released_Year	IMDB_Rating	Meta_score	No_of_Votes	Gross	GDP	gdp_anual	CPIAUCSL	inflacao_anual
Released_Year	1.000000	-0.115700	-0.295667	0.205644	0.239983	0.953349	-0.482123	0.987194	-0.294948
IMDB_Rating	-0.115700	1.000000	0.263726	0.495986	0.098624	-0.093836	0.057128	-0.108606	0.012001
Meta_score	-0.295667	0.263726	1.000000	-0.011952	-0.039093	-0.206274	0.118587	-0.267865	0.024870
No_of_Votes	0.205644	0.495986	-0.011952	1.000000	0.561575	0.159837	-0.100269	0.196200	-0.049859
Gross	0.239983	0.098624	-0.039093	0.561575	1.000000	0.240305	-0.127308	0.234259	-0.045151
GDP	0.953349	-0.093836	-0.206274	0.159837	0.240305	1.000000	-0.524760	0.977588	-0.397094
gdp_anual	-0.482123	0.057128	0.118587	-0.100269	-0.127308	-0.524760	1.000000	-0.533719	0.611092
CPIAUCSL	0.987194	-0.108606	-0.267865	0.196200	0.234259	0.977588	-0.533719	1.000000	-0.369317
inflacao_anual	-0.294948	0.012001	0.024870	-0.049859	-0.045151	-0.397094	0.611092	-0.369317	1.000000

Figura 05 – Correlação das variáveis

Fonte: Próprio autor

A partir da análise da correlação tem-se:

- A coluna "Gross" possui correlação moderada e positiva com a coluna "No_of_Votes", o que sugere que a quantidade de votos de usuários está associada positivamente ao faturamento.
- A coluna "IMDB_Rating" possui correlação moderada e positiva com a coluna "No_of_Votes", indicando que a classificação no IMDB está associada positivamente à quantidade de votos dos usuários.
- A coluna "Meta_score" possui correlação fraca a moderada negativa com a coluna "Released_Year", sugerindo que a média ponderada das críticas está associada negativamente com o ano de lançamento.
- As colunas "No_of_Votes" e "Gross" possuem correlação fraca com a coluna "Released_Year", indicando que a quantidade de votos e o faturamento não sofrem grandes efeitos em função do ano de lançamento.
- As colunas "GDP" e "CPIAUCSL" possuem correlação fraca e positiva com a coluna "Gross", sugerindo que o faturamento dos filmes não sofre efeitos severos em relação ao PIB e à inflação dos Estados Unidos.
- As colunas "gdp_anual" e "inflacao_anual" apresentam correlação fraca a negativa com a coluna "Gross", sugerindo que as variações anuais do PIB e da inflação norte-americana não exercem efeitos relevantes sobre o faturamento dos filmes.

3.3 Visualizações

Foi empregado

O histograma foi aplicado na coluna “Gross” (faturamento) para ver como os dados estão distribuídos. A imagem é ilustrada a seguir.

Figura 06 – Gráfico de histograma da variável faturamento

Fonte: Próprio autor

Observa-se uma maior concentração de filmes com faturamento baixo, enquanto apenas poucos alcançam faturamentos muito elevados.

Outra variável que chamou atenção foi a variável "No_of_votes" (número de votos), na análise descritiva a amplitude é muito alta, percebe-se a confirmação de outliers na variável, indicando que certos filmes obtiveram um número de votos excepcionalmente alto em comparação com os demais.

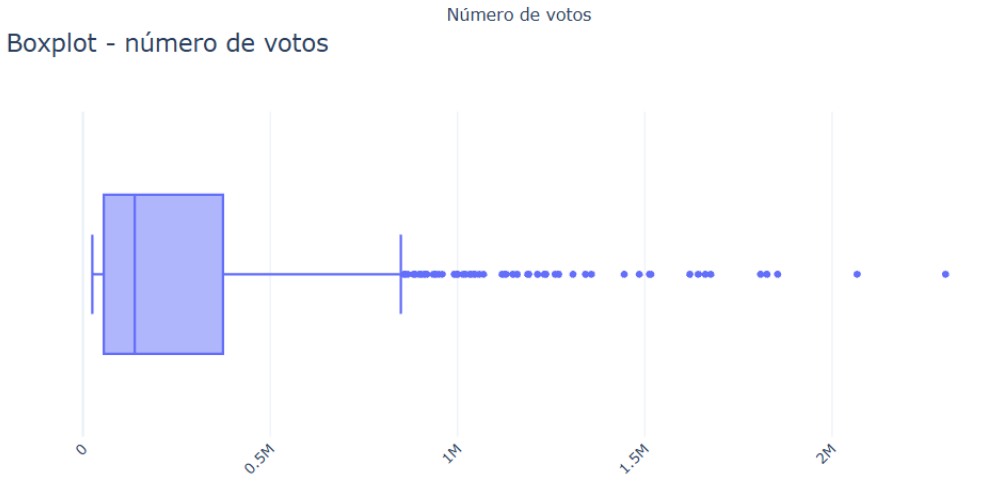

Figura 07 – Boxplot da variável número de votos

Fonte: Próprio autor

Foi analisada as notas IMDb por faturamentos, os maiores faturamentos se concentram em notas entre 7,5 e 8 no rating de classificação.

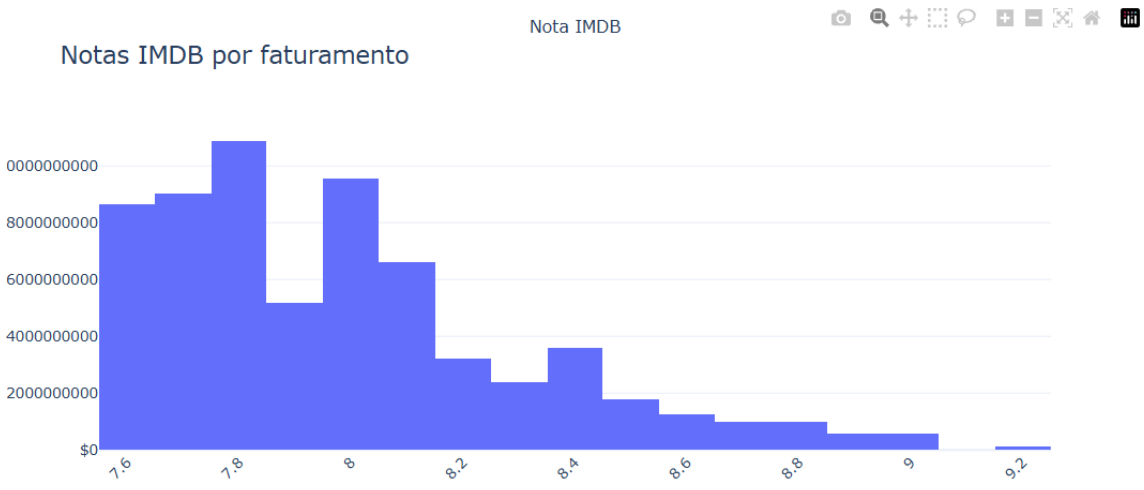

Figura 08 – Histograma de frequência da variável IMDB_Rating

Fonte: Próprio autor

Em seguida, foram analisados os 10 filmes com maior quantidade de votos por faturamento.

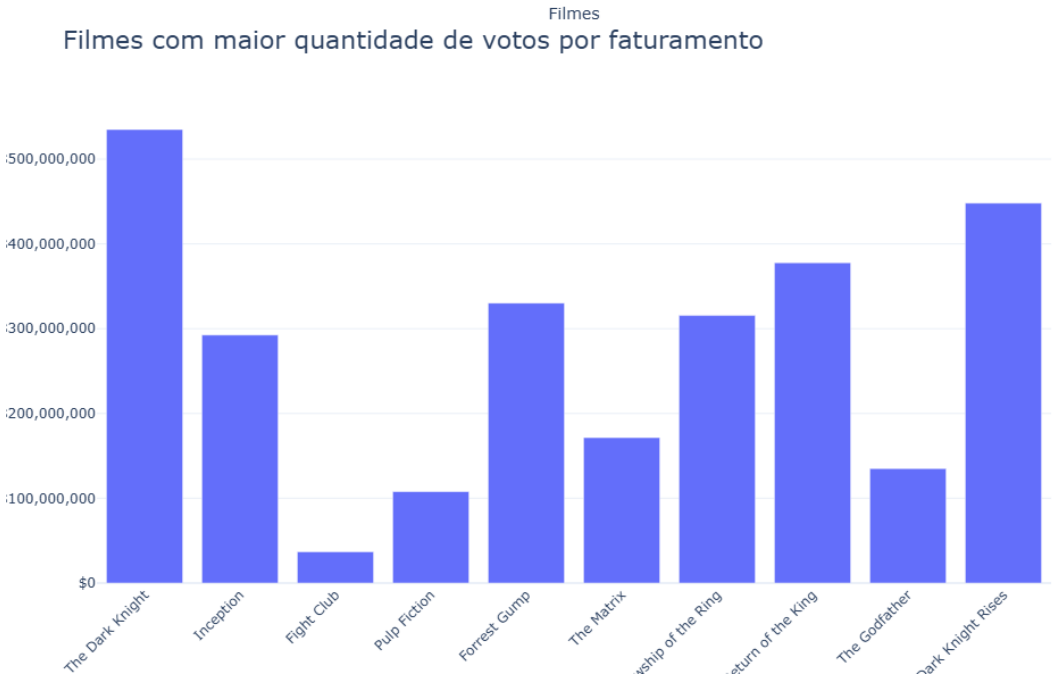

Figura 09 – Filmes com maior quantidade de votos por faturamento

Fonte: Próprio autor

A partir do gráfico, observa-se que há mais filmes populares com faturamento alto.

A próxima imagem, apresenta os 10 diretores com mais publicações de filmes.

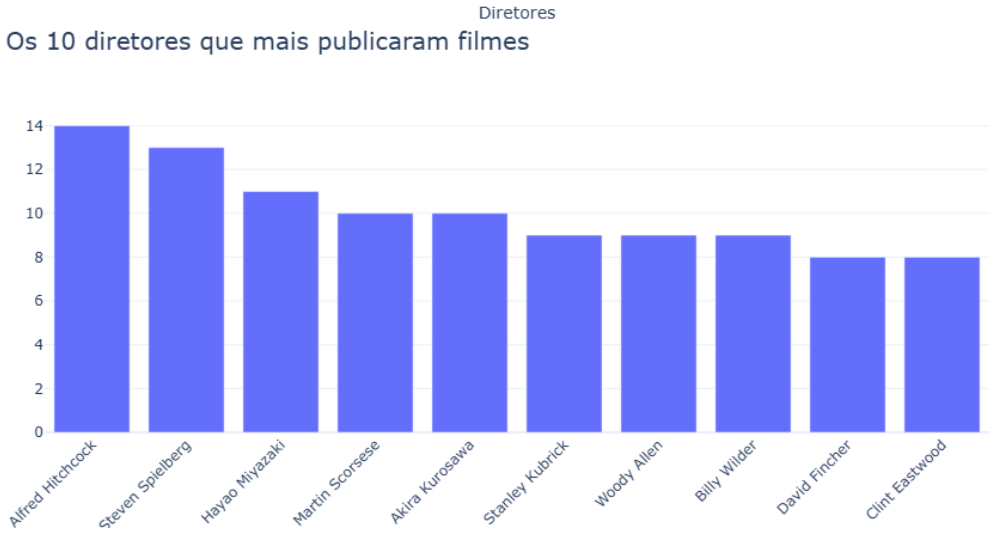

Figura 10 – Os 10 diretores com mais publicaram filmes

Fonte: Próprio autor

A maioria das publicações são do gênero “Drama”, em seguida, “Drama, Romance”, “Comedy, Drama” e etc.

Figura 11 – Os 10 gêneros mais publicados

Fonte: Próprio autor

Tem-se os 10 filmes mais faturados de todos os tempos.

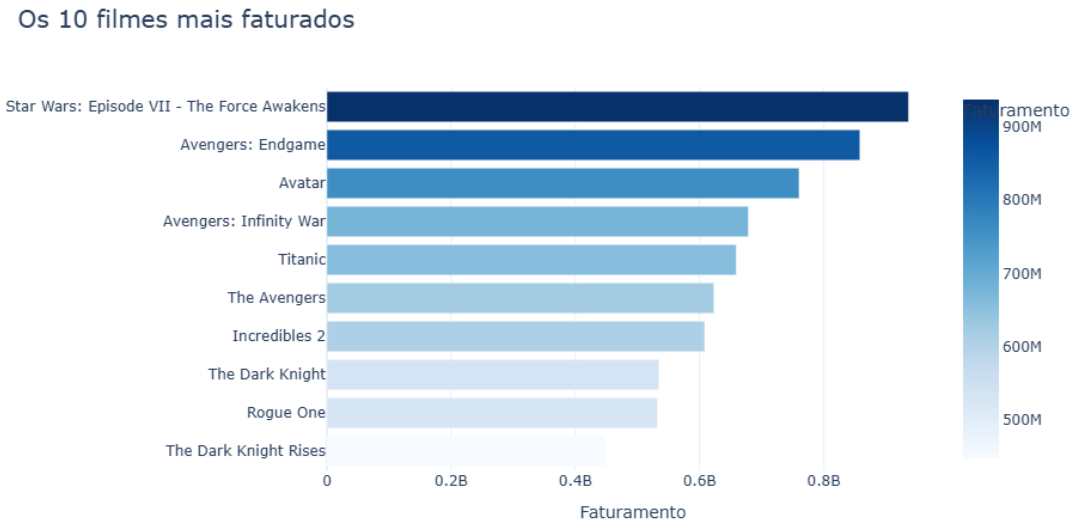

Figura 12 – Os 10 filmes mais faturados

Fonte: Próprio autor

A seguir, temos o gráfico com os 10 filmes com maiores notas de IMBd.

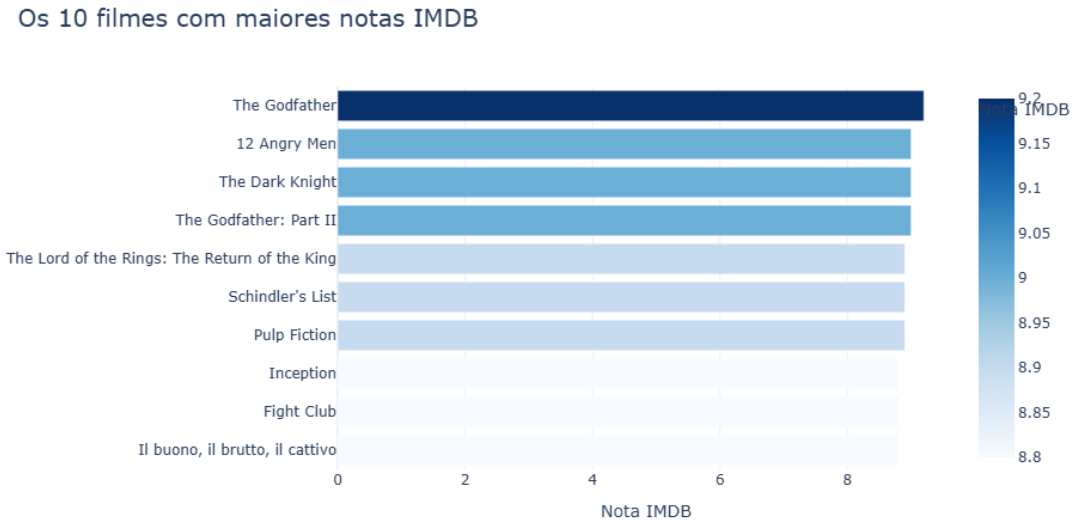

Figura 13 – Os 10 filmes com maiores notas IMDb

Fonte: Próprio autor

Os 10 filmes com maior número de votos.

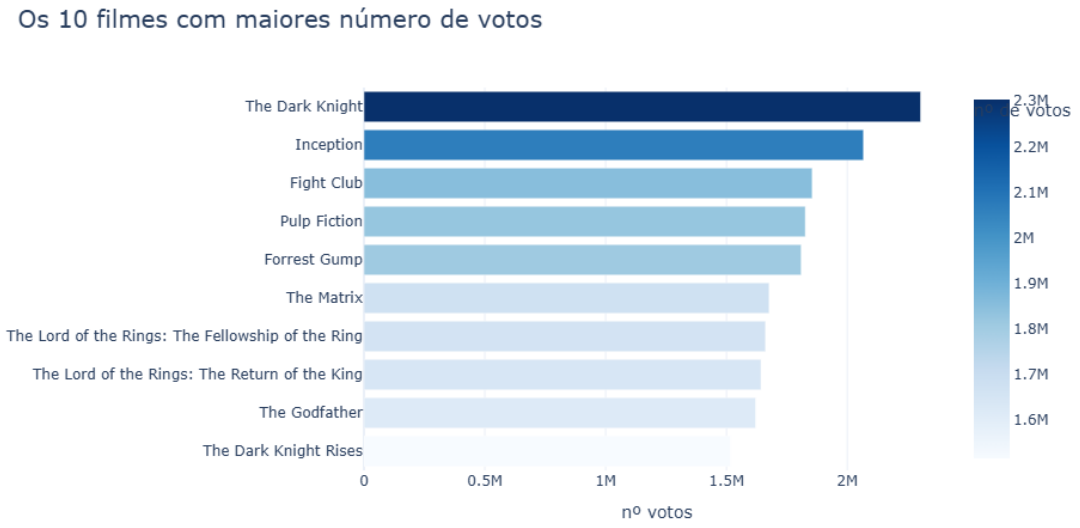

Figura 14 – Os 10 filmes com maiores número de votos

Fonte: Próprio autor

3.4 Hipóteses levantadas

As hipóteses levantadas na análise foram:

- Nos últimos anos, o faturamento aumentou vertiginosamente. Porém, houve períodos em que o faturamento foi baixo, sinalizando possíveis ciclos. No entanto, não se sabem os motivos que levaram a esses faturamentos reduzidos, uma vez que o PIB e a inflação norte-americana não apresentam forte associação com esse desempenho.
- Outra hipótese levantada foi se as notas altas estariam relacionadas a faturamentos elevados. Contudo, observou-se que os maiores faturamentos se concentram em filmes com notas entre 7,5 e 8 no rating do IMDb.
- Também foi levantada a hipótese de que filmes mais criticados poderiam apresentar maiores faturamentos. No entanto, conforme mostra a Figura 09, comprovou-se o contrário: as maiores críticas não estão associadas ao faturamento dos filmes.

4. PERGUNTAS DO DESAFIO

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

Para recomendar filmes para uma nova pessoa que ainda não possui histórico de perfil, pode-se adotar estratégia baseada em popularidade e qualidade. Os critérios de seleção são:

- Filmes com avaliação no IMDb igual ou superior a 7.5 (alta qualidade).

- Filmes com um número elevado de votos, o que indica que são amplamente conhecidos e assistidos (alta popularidade).

A regra de negócio apresentado gerou as seguintes recomendações:

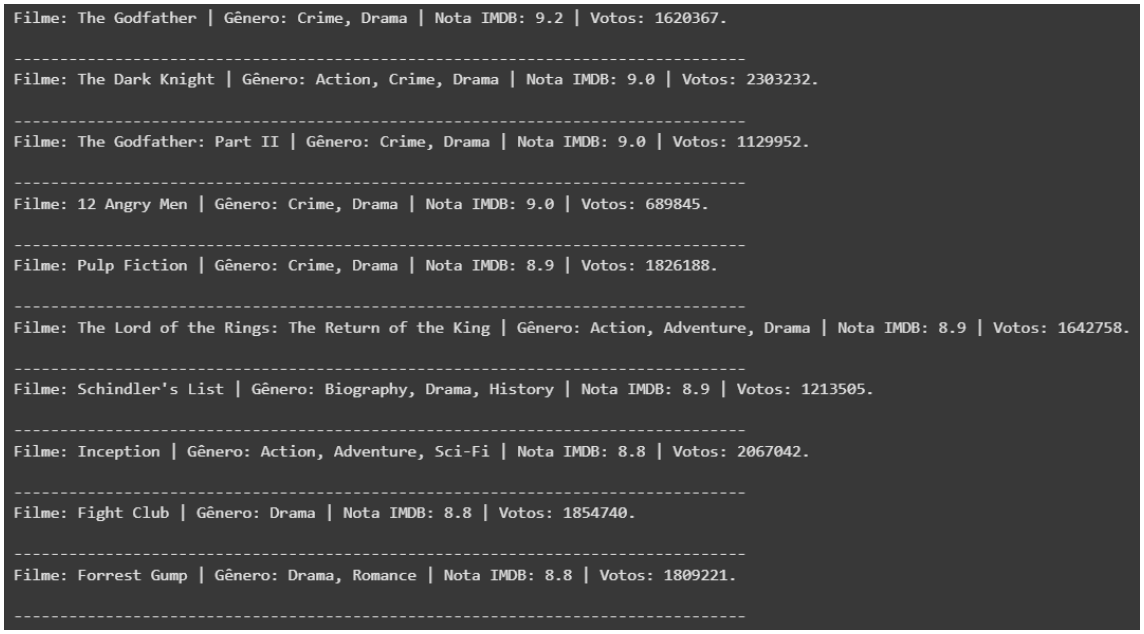

Figura 15 – Os 10 filmes recomendados pela regra do negócio

Fonte: Próprio autor

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

Os principais fatores que estão relacionados com alta expectativa de faturamento são:

- Ano de lançamento: observou-se um aumento expressivo no faturamento nos últimos anos, indicando uma tendência de crescimento.
- Quantidade de votos: verificou-se uma correlação moderada e positiva, sugerindo que a quantidade de votos de usuários está associada ao aumento do faturamento.
- PIB e CPI dos Estados Unidos: essas variáveis apresentam correlação fraca e positiva com o faturamento, o que indica que o desempenho dos filmes não sofre impactos significativos em função do PIB e da inflação norte-americana.

4.3 - Quais insights podem ser tirados com a coluna *Overview*?

Foi aplicada a técnica de nuvem de palavras à coluna *"Overview"* (visão geral) para os três gêneros mais recorrentes, sendo os resultados apresentados nas imagens a seguir.

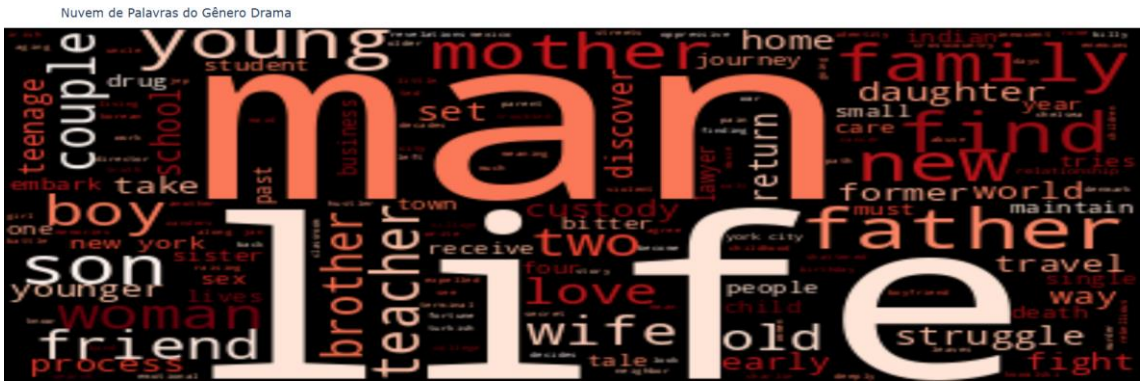

Figura 16 – Nuvem de palavras encontrada no gênero Drama

Fonte: Próprio autor

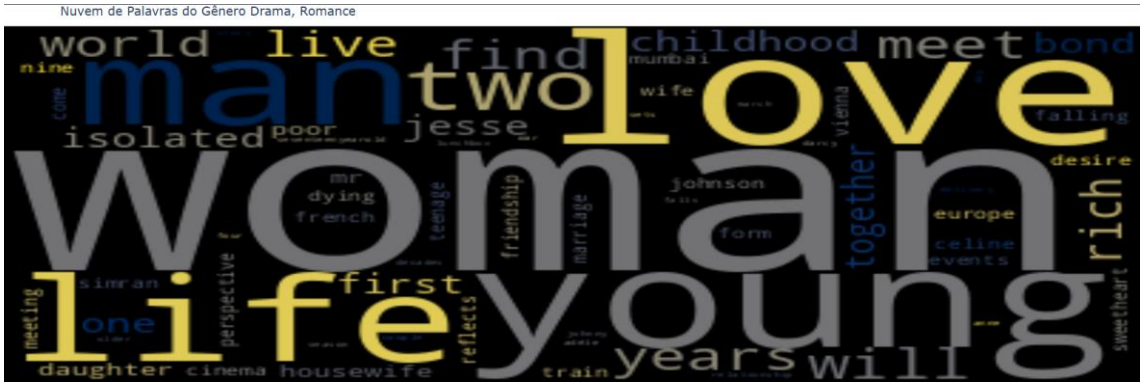

Figura 17 – Nuvem de palavras encontrada no gênero Drama, Romance

Fonte: Próprio autor

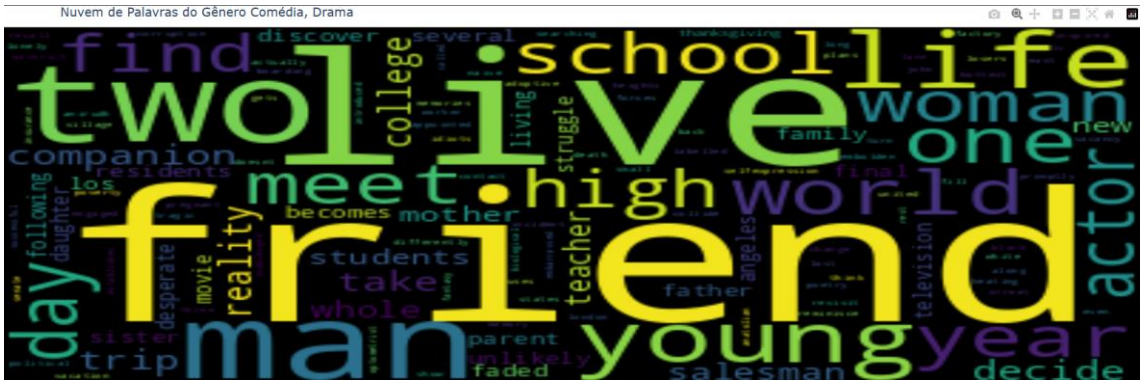

Figura 18 – Nuvem de palavras encontrada no gênero Comédia, Drama

Fonte: Próprio autor

4.4 - É possível inferir o gênero do filme a partir dessa coluna?

Sim. Foi empregada a técnica TF-IDF, que converte frases em vetores numéricos preservando a relevância semântica dos termos. Dessa forma, permite que algoritmos de machine learning consigam diferenciar textos por padrões de vocabulário.

O modelo de machine learning utilizado foi o MultinomialNB, um classificador probabilístico baseado no Teorema de Bayes. Esse modelo se adapta bem a variáveis discretas e busca representar a distribuição das palavras como ocorrências de uma distribuição multinomial.

Antes da vetorização e do treinamento com o modelo Naive Bayes, os dados passaram por oversampling, um método de rebalanceamento que aumenta o número de amostras dos gêneros minoritários, de modo a garantir uma representatividade mais equilibrada entre as classes. Essa etapa foi necessária porque o gênero Drama predominava na base, o que poderia enviesar o modelo durante o processo de treinamento.

Após o processo de oversampling, os dados foram divididos em dados de treinamento e dados de teste, foi atribuído 80% para dados de treino e 20% para os dados de teste. O modelo atingiu acurácia de 98%, valor expressivamente alto, sinalizando que o modelo poderia ter decorado, assim gerando overfitting.

4.5 - Como prever a nota do IMDB?

O modelo empregado para realizar a previsão de nota IMDb foi modelo de regressão. As variáveis utilizadas foram variáveis numéricas como: “Released_Year” (ano de lançamento), “Runtime” (tempo de duração), “Meta_score” (notas de crítica), “No_of_Votes” (número de votos) e “Gross” (faturamento) e variáveis categóricas como: “Certificate” (faixa etária), “Genre” (gênero).

Foi empregado a técnica One Hot Encode nas variáveis categóricas, esta técnica converte as variáveis categóricas em numéricas, criando uma coluna nova para cada categoria. Em cada nova coluna, é atribuído 1 se a categoria estiver presente nessa linha e 0 se estiver ausente.

Nas variáveis numéricas, foi empregada Standard Scaler do Scikit-Learn, esta ferramenta transforma os dados e os coloca na mesma escala, evita que os modelos deem maior prioridade com valores maiores. A ferramenta transforma os dados de forma que a média seja igual a zero com desvio padrão igual a um.

Foi aplicado técnica VIF (Variance Inflation Factor) ou Fator de Inflação de Variância, esta técnica visa detectar multicolinearidade nas variáveis, atua quantificando o quanto a variação de uma variável é inflada devido a correlações com outras variáveis.

A multicolinearidade ocorre quando duas ou mais variáveis em um modelo de regressão são altamente correlacionadas, dificultando a identificação do impacto individual de cada variável sobre a variável alvo.

Valores abaixo de 5 significa que a variável não possui multicolinearidade em relação às outras variáveis, porém valores altos, significa que a variável exerce impacto forte nas outras variáveis, caso ocorra, deve-se remover a variável ou aplicar técnicas de redução de dimensionalidade.

O resultado obtido na técnica VIF é apresentado na figura a seguir.

A partir dos resultados obtidos pelo VIF observa-se que não há multicolinearidade nas variáveis relacionadas ao gênero. No entanto, as variáveis associadas aos certificados apresentaram indícios de colineraridade que é natural na coluna de faixa etária.

	Feature	VIF
0	const	0.000000
1	Released_Year	2.018700
2	Meta_score	1.275376
3	No_of_Votes	1.735470
4	Gross	1.906826
5	Runtime_clean	1.427047
6	Action	1.658962
7	Adventure	1.870476
8	Animation	1.682340
9	Biography	1.351224
10	Comedy	1.676737
11	Crime	1.434966
12	Drama	2.233373
13	Family	1.400886
14	Fantasy	1.229276
15	Film-Noir	1.549122
16	History	1.250168
17	Horror	1.333904
18	Music	1.135962
19	Musical	1.164586
20	Mystery	1.246928
21	Romance	1.325874
22	Sci-Fi	1.323395
23	Sport	1.097731
24	Thriller	1.382215
25	War	1.174932
26	Western	1.152901
27	Cert_16	inf
28	Cert_A	inf
29	Cert_Action	inf
30	Cert_Approved	inf
31	Cert_Comedy	inf
32	Cert_Crime	inf
33	Cert_Drama	inf
34	Cert_Fantasy	inf
35	Cert_Film-Noir	inf
36	Cert_G	inf
37	Cert_GP	inf
38	Cert_PG	inf
39	Cert_PG-13	inf
40	Cert_Passed	inf
41	Cert_R	inf
42	Cert_TV-14	inf
43	Cert_TV-MA	inf
44	Cert_TV-PG	inf
45	Cert_Thriller	inf
46	Cert_U	inf
47	Cert_U/A	inf
48	Cert_UA	inf
49	Cert_Unrated	inf

Figura 19 – resultado da técnica VIF

Fonte: Próprio autor

Os modelos de regressão utilizados foram:

- Ridge Regression
- Lasso Regression

- ElasticNet
- Random Forest Regression
- Decision Tree Regression
- XGBoost Regression

Os modelos Ridge Regression, Lasso Regression e ElasticNet têm como objetivo aplicar penalizações aos coeficientes, a fim de reduzir o risco de overfitting e melhorar a generalização do modelo.

O modelo Regressão Ridge utiliza a penalização L2, que diminui a magnitude dos coeficientes, sendo especialmente útil em situações de multicolinearidade, pois reduz a influência excessiva de variáveis altamente correlacionadas, sem eliminá-las.

O modelo Regressão Lasso aplica a penalização L1, que atua de forma semelhante ao modelo anterior, mas com a característica adicional de poder zerar coeficientes. Dessa forma, além de reduzir a complexidade do modelo, também realiza uma seleção automática de variáveis.

Já o modelo ElasticNet combina as penalizações L1 e L2, unindo as vantagens de ambos: promove a seleção de variáveis, como o modelo Lasso, e mantém estabilidade em variáveis correlacionadas como o modelo Ridge.

Foi empregado o Erro Quadrático Médio como métrica para avaliar a precisão dos modelos de regressão. A técnica calcula a média das diferenças quadradas entre os valores previstos e os valores reais.

Após o treinamento dos modelos, o Random Forest apresentou desempenho satisfatório em comparação aos demais. Entretanto, para avaliar sua capacidade de generalização, foi empregada a validação cruzada. Esse método desempenha um papel fundamental na estimativa da habilidade do modelo em generalizar para dados não observados, oferecendo uma avaliação mais robusta e confiável de sua performance e assegurando previsões consistentes em diferentes subconjuntos do conjunto de dados.

O modelo Regressão Ridge se destacou por sua capacidade de generalização, apresentando uma média de 0,77 e desvio padrão de 0,05.

	Modelo	Scores	Desvio Padrão
0	Ridge Regression	0.77	0.05
1	Lasso Regression	0.76	0.06
2	ElasticNet	0.76	0.05
3	Random Forest	0.72	0.04
4	Decision Tree	1.02	0.09
5	XGBoost	0.75	0.05

Figura 20 – Resultado da Validação Cruzada

Fonte: Próprio autor

A otimização de hiperparâmetros foi empregada no modelo de regressão Ridge com o objetivo de encontrar o melhor valor de alpha, capaz de ajustar o modelo de forma mais adequada. A técnica utilizada foi o GridSearchCV, que combina diferentes valores de alpha em uma busca sistemática, ajustando o modelo até identificar aquele que maximiza o desempenho. O valor obtido para o parâmetro foi alpha = 10.

Finalmente, foi realizada a previsão da nota do IMDb a partir dos dados fornecidos no desafio, e o resultado obtido pelo modelo foi 4,91.