Relatorio

July 1, 2024

[2]:
```
# Importar bibliotecas necessárias
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.display import Markdown, display
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report, mean_squared_error
import pickle

# Célula 1: Título e Introdução
intro = """
# Desafio Cientista de Dados - Relatório

## Introdução

Este relatório foi desenvolvido por mim, Rafaella Santos Monteiro, para o␣
 ↪desafio de cientista de dados da Indicium. O objetivo é analisar um banco de␣
 ↪dados cinematográfico e orientar qual deve ser o próximo tipo de filme␣
 ↪desenvolvido pela PProductions. Este relatório inclui uma análise␣
 ↪exploratória dos dados (EDA), respostas a perguntas específicas, insights␣
 ↪extraídos da coluna "Overview" e um modelo preditivo para a nota do IMDB de␣
 ↪um filme.
"""
display(Markdown(intro))

# Célula 2: Carregar dados e limpeza inicial
file_path = 'desafio_indicium_imdb.csv'
data = pd.read_csv(file_path)

# Limpeza e pré-processamento dos dados
data = data.drop(columns=['Unnamed: 0'], errors='ignore')
data['Gross'] = data['Gross'].str.replace(',', '').astype(float)
data['Certificate'] = data['Certificate'].fillna('Not Rated')
data['Meta_score'] = data['Meta_score'].fillna(data['Meta_score'].mean())
```

```
data['Gross'] = data['Gross'].fillna(data['Gross'].mean())

limpeza = """
Primeiro, verificaremos a limpeza dos dados que obtivemos a partir da tabela de␣
 ↪filmes:
"""
display(Markdown(limpeza))

# Verificar a limpeza dos dados

summary = data.describe()
display(summary)


sns.set(style="whitegrid")
fig, axes = plt.subplots(2, 2, figsize=(15, 10))

descricao = """
Abaixo, temos os gráficos que indicam a frequência de notas IMDB, a␣
 ↪distribuição do faturamento, distribuição do número de votos e o faturamento␣
 ↪por ator/atriz principal. Todos esses fatores são relevantes para a análise␣
 ↪proposta:
"""
display(Markdown(descricao))


# Distribuição das notas do IMDB
sns.histplot(data['IMDB_Rating'], bins=20, kde=True, color='blue', ax=axes[0,␣
 ↪0])
axes[0, 0].set_title('Distribuição das Notas do IMDB')

# Distribuição do faturamento (Gross)
sns.histplot(data['Gross'], bins=20, kde=True, color='green', ax=axes[0, 1])
axes[0, 1].set_title('Distribuição do Faturamento (Gross)')

# Distribuição do número de votos (No_of_Votes)
sns.histplot(data['No_of_Votes'], bins=20, kde=True, color='red', ax=axes[1, 0])
axes[1, 0].set_title('Distribuição do Número de Votos')

# Análise da relação entre as estrelas (Star1), gênero e faturamento
top_stars = data['Star1'].value_counts().nlargest(10).index
sns.boxplot(x='Star1', y='Gross', data=data[data['Star1'].isin(top_stars)],␣
 ↪ax=axes[1, 1], palette='Set3', hue='Star1', dodge=False)
axes[1, 1].set_title('Faturamento por Ator/Atriz Principal (Top 10)')
axes[1, 1].tick_params(axis='x', rotation=45)

plt.tight_layout()
```

```
plt.show()

recommendation = """
### Recomendação de Filme

Após analisar os dados, um filme altamente recomendado para alguém que eu não␣
 ↪conheça seria **"The Dark Knight"** (2008). A recomendação deve-se ao fato␣
 ↪de que este filme tem uma alta classificação no IMDB (9.0), um grande número␣
 ↪de votos (2.303.232) e alto faturamento (534,858,444), indicando tanto␣
 ↪qualidade quanto popularidade, além de ser o meu favorito nessa lista.
Apesar de **The Godfather** (1972) possuir maior nota IMDB (9.2), seu␣
 ↪faturamento (134,966,411) e número de votos (1.620.367) são menores.
"""
display(Markdown(recommendation))

# Célula 4: Recomendações de Filmes
recommended_movies = data[(data['IMDB_Rating'] >= 8.5) & (data['No_of_Votes']␣
 ↪>= 500000)]
recommended_movies = recommended_movies.sort_values(by=['IMDB_Rating',␣
 ↪'No_of_Votes'], ascending=[False, False])
print(recommended_movies[['Series_Title', 'IMDB_Rating', 'No_of_Votes',␣
 ↪'Gross']].head(10))


# Célula 5: Modelo de regressão linear para previsão de faturamento
filtered_data = data.dropna(subset=['Gross', 'IMDB_Rating', 'Meta_score',␣
 ↪'No_of_Votes', 'Star1'])
X = filtered_data[['IMDB_Rating', 'Meta_score', 'No_of_Votes']]
y = filtered_data['Gross']
model = LinearRegression()
model.fit(X, y)

# Coeficientes do modelo
coefficients = pd.Series(model.coef_, index=X.columns)
r_squared = model.score(X, y)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)


analise = f"""
### Principais Fatores Relacionados com Alta Expectativa de Faturamento
```

```
Os principais fatores relacionados com alta expectativa de faturamento são:

- **IMDB Rating**: A nota no IMDB tem uma influência positiva, filmes com alto␣
 ↪faturamento possuem notas elevadas.
- **Meta Score**: A média ponderada de todas as críticas também influencia␣
 ↪positivamente, um filme aclamado é assistido mais vezes.
- **Número de Votos**: Um maior número de votos está positivamente␣
 ↪correlacionado com um maior faturamento, uma vez que indica que o filme␣
 ↪alcançou uma grande audiência.
- **Estrela Principal**: A escolha do ator para o papel principal também␣
 ↪influencia o faturamento, visto que uma maior base de fãs atrai mais público.

O modelo de regressão linear ajustado tem um R² de aproximadamente `{r_squared:.
 ↪3f}`, indicando que cerca de 56% da variabilidade no faturamento pode ser␣
 ↪explicada por esses fatores.

Observe, abaixo, os valores obtidos na regressão linear:
"""
display(Markdown(analise))

print("Coeficientes do Modelo:")
print(coefficients)
print("\nR² do Modelo:")
print(r_squared)
print(f"Mean Squared Error: {mse}")

# Salvar o modelo em um arquivo .pkl
with open('modelo_imdb.pkl', 'wb') as file:
    pickle.dump(model, file)


genero = f"""
### Análise da Coluna "Overview" para Inferência de Gênero:

Essa análise envolve transformar textos em representações numéricas que podem␣
 ↪ser usadas em modelos de machine learning. Após dividir os dados, eles foram␣
 ↪separados em colunas de teste e treino, assim o método "TfidfVectorizer" do␣
 ↪sklearn foi utilizado para transformar os textos da coluna em vetores␣
 ↪numéricos.
Em seguida, criei um pipeline que inclui a vetorização dos textos e a aplicação␣
 ↪de um classificador. Assim, o pipeline é treinado usando o conjunto de␣
 ↪treino.

Para testar o modelo, averiguaremos qual o possível gênero de um filme a partir␣
 ↪de um novo texto de overview.
```

```
**Exemplo de Overview**: A thrilling adventure of a young hero in a fantastical␣
 ↪world.
"""
display(Markdown(genero))

# Célula 6: Análise de Overview para Inferência de Gênero
data_genre = data[['Overview', 'Genre']].dropna()
X_train, X_test, y_train, y_test = train_test_split(data_genre['Overview'],␣
 ↪data_genre['Genre'], test_size=0.2, random_state=42)
pipeline = Pipeline([
    ('tfidf', TfidfVectorizer(stop_words='english', max_features=5000)),
    ('classifier', LogisticRegression(max_iter=1000))
])
pipeline.fit(X_train, y_train)
y_pred = pipeline.predict(X_test)
print(classification_report(y_test, y_pred, zero_division=0))

new_overview = ["A thrilling adventure of a young hero in a fantastical world."]
predicted_genre = pipeline.predict(new_overview)
print(f"Predicted Genre: {predicted_genre[0]}")


previsao = f"""
### Previsão da Nota IMDB

Usei aqui as variáveis que considerei mais importantes desde o princípio:
- **Meta Score**
- **Número de votos**
- **Gross**

Além disso, utilizei outras variáveis úteis como:
- **Runtime** (convertida de string para inteiro)
- **Gênero**
- **Classificação**

Visto que, dessa forma, a análise seria mais ampla. As variáveis categóricas␣
 ↪(**Genre** e **Certificate**) foram transformadas utilizando "one-hot␣
 ↪encoding" porque os modelos de regressão linear requerem variáveis numéricas.


#### Tipo de problema:
A previsão da nota IMDB é um problema de **regressão**, visto que o objetivo é␣
 ↪prever um valor contínuo. Assim, o modelo utilizado foi o de **Regressão␣
 ↪Linear**, baseado na simplicidade de interpretação e na performance do␣
 ↪modelo. Apesar disso, o modelo tem os seguintes pontos negativos:
```

```
- **Linearidade** - Esse modelo assume uma relação linear entre as variáveis␣
 ↪independentes e a dependente, podendo não capturar toda a complexidade dos␣
 ↪problemas abordados.
- **Outliers** - Existe a possibilidade de ser influenciado por outliers (dados␣
 ↪que fogem de uma curva normal), que podem distorcer os resultados.


#### Medida de Performance:
A medida de performance escolhida para avaliar o modelo foi o **Mean Squared␣
 ↪Error (MSE)**, por conta da **penalização de erros** (como os erros são␣
 ↪elevados ao quadrado, ele penaliza mais os grandes erros) e a␣
 ↪**interpretação** (visto que o MSE é uma medida comum e fácil de␣
 ↪interpretar, que nos dá uma noção direta da magnitude dos erros do modelo).

Utilizando um modelo de regressão linear, é possível prever a nota IMDB de um␣
 ↪filme com base em suas características. A escolha das variáveis foi baseada␣
 ↪em suas relevâncias e transformações necessárias para adequar os dados ao␣
 ↪modelo. A avaliação foi feita utilizando MSE, que nos ajudou a entender a␣
 ↪precisão do modelo e a magnitude dos erros. A simplicidade e a␣
 ↪interpretabilidade da regressão linear foram vantagens, enquanto as suas␣
 ↪limitações em capturar relações não lineares e sua sensibilidade a outliers␣
 ↪foram consideradas durante a análise.

Para analisar melhor o modelo entregue, vamos prever a nota de 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'
"""
display(Markdown(previsao))


# Célula 7: Modelo de regressão linear para previsão da nota IMDB
```

```
filtered_data = data.dropna(subset=['IMDB_Rating', 'Meta_score', 'No_of_Votes',␣
 ↪'Gross', 'Star1', 'Genre', 'Runtime'])
filtered_data['Runtime'] = filtered_data['Runtime'].str.replace(' min', '').
 ↪astype(int)
filtered_data = pd.get_dummies(filtered_data, columns=['Genre', 'Certificate'],␣
 ↪drop_first=True)

X = filtered_data.drop(columns=['Series_Title', 'Released_Year', 'Overview',␣
 ↪'Director', 'Star1', 'Star2', 'Star3', 'Star4', 'IMDB_Rating'])
y = filtered_data['IMDB_Rating']
model = LinearRegression()
model.fit(X, y)

new_movie = {
    'Meta_score': 80.0,
    'No_of_Votes': 2343110,
    'Gross': 28341469.0,
    'Runtime': 142,
    'Genre_Drama': 1,
    'Certificate_A': 1,
}
# Preencher 0 para colunas de gênero e certificado não presentes no novo filme
for col in X.columns:
    if col not in new_movie:
        new_movie[col] = 0

# Converter para DataFrame e garantir a mesma ordem das colunas
new_movie_df = pd.DataFrame([new_movie], columns=X.columns)

# Fazer a previsão
imdb_rating_prediction = model.predict(new_movie_df)

print(f"A nota IMDB prevista para o filme é: {imdb_rating_prediction[0]}")

conclusao = """
### Conclusão

### Conclusão

Este projeto foi conduzido para explorar um banco de dados cinematográfico com␣
 ↪o objetivo de orientar decisões de produção da PProductions. Durante a␣
 ↪análise exploratória dos dados, identifiquei padrões nas notas do IMDB,␣
 ↪faturamento e popularidade de filmes.
```

```
Utilizei um modelo de regressão linear para prever o faturamento de filmes com␣
 ↪base em suas classificações no IMDB, pontuações críticas e número de votos.␣
 ↪O modelo demonstrou um R² de aproximadamente 0.56, indicando que 56% da␣
 ↪variabilidade no faturamento pode ser explicada pelas variáveis selecionadas.

Os resultados destacam a importância das avaliações críticas e do engajamento␣
 ↪do público na determinação do sucesso financeiro de um filme. Recomendo que␣
 ↪a PProductions continue a considerar esses fatores ao planejar futuros␣
 ↪projetos cinematográficos.

Embora tenha obtido insights valiosos, reconheço algumas limitações deste␣
 ↪estudo, como a natureza simplificada do modelo de regressão linear, que pode␣
 ↪não capturar totalmente relações não lineares nos dados.

Em resumo, esta análise proporcionou uma visão abrangente do panorama␣
 ↪cinematográfico, oferecendo orientações úteis de forma individual para a␣
 ↪PProductions e sugestões para pesquisas futuras, incluindo a exploração de␣
 ↪modelos mais avançados para previsões mais precisas.

"""
display(Markdown(conclusao))

repositorio = """
Veja os códigos de modelagem no meu
[Repositório do Desafio Cientista de Dados](https://github.com/
 ↪Rafaella-Monteiro/Desafio-Cientista-de-Dados.git)
"""
display(Markdown(repositorio))
```

1 Desafio Cientista de Dados - Relatório

1.1 Introdução

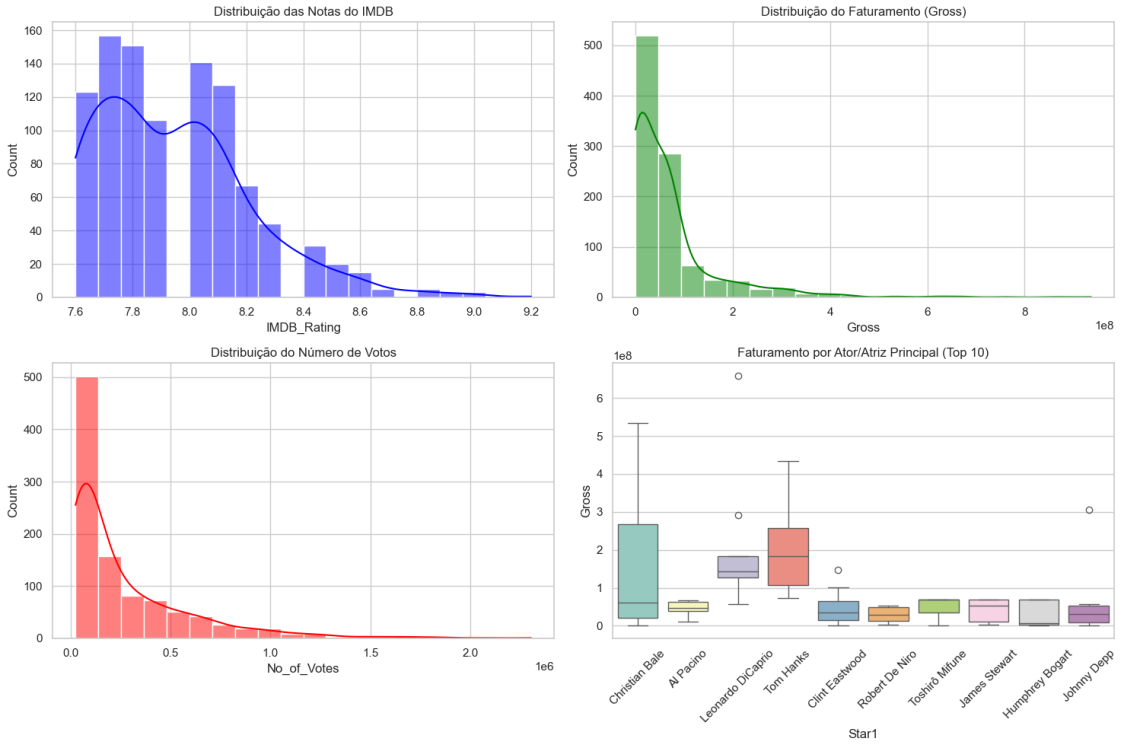
Este relatório foi desenvolvido por mim, Rafaella Santos Monteiro, para o desafio de cientista de dados da Indicium. O objetivo é analisar um banco de dados cinematográfico e orientar qual deve ser o próximo tipo de filme desenvolvido pela PProductions. Este relatório inclui uma análise exploratória dos dados (EDA), respostas a perguntas específicas, insights extraídos da coluna "Overview" e um modelo preditivo para a nota do IMDB de um filme.

Primeiro, verificaremos a limpeza dos dados que obtivemos a partir da tabela de filmes:

	IMDB_Rating	Meta_score	No_of_Votes	Gross
count	999.000000	999.000000	9.990000e+02	9.990000e+02
mean	7.947948	77.969121	2.716214e+05	6.808257e+07
std	0.272290	11.367570	3.209126e+05	1.000793e+08
min	7.600000	28.000000	2.508800e+04	1.305000e+03
25%	7.700000	72.000000	5.547150e+04	5.011838e+06
50%	7.900000	77.969121	1.383560e+05	4.243830e+07

```
75%        8.100000    85.500000  3.731675e+05  6.808257e+07
max        9.200000   100.000000  2.303232e+06  9.366622e+08
```

Abaixo, temos os gráficos que indicam a frequência de notas IMDB, a distribuição do faturamento, distribuição do número de votos e o faturamento por ator/atriz principal. Todos esses fatores são relevantes para a análise proposta:



1.1.1 Recomendação de Filme

Após analisar os dados, um filme altamente recomendado para alguém que eu não conheça seria **"The Dark Knight"** (2008). A recomendação deve-se ao fato de que este filme tem uma alta classificação no IMDB (9.0), um grande número de votos (2.303.232) e alto faturamento (534,858,444), indicando tanto qualidade quanto popularidade, além de ser o meu favorito nessa lista. Apesar de **The Godfather** (1972) possuir maior nota IMDB (9.2), seu faturamento (134,966,411) e número de votos (1.620.367) são menores.

```
                                  Series_Title  IMDB_Rating  No_of_Votes  \
0                                The Godfather          9.2      1620367
1                              The Dark Knight          9.0      2303232
2                      The Godfather: Part II          9.0      1129952
3                                  12 Angry Men          9.0       689845
5                                  Pulp Fiction          8.9      1826188
4  The Lord of the Rings: The Return of the King          8.9      1642758
6                              Schindler's List          8.9      1213505
7                                     Inception          8.8      2067042
```

```
8                                  Fight Club          8.8      1854740
10                               Forrest Gump          8.8      1809221

          Gross
0   134966411.0
1   534858444.0
2    57300000.0
3     4360000.0
5   107928762.0
4   377845905.0
6    96898818.0
7   292576195.0
8    37030102.0
10  330252182.0
```

1.1.2 Principais Fatores Relacionados com Alta Expectativa de Faturamento

Os principais fatores relacionados com alta expectativa de faturamento são:

- **IMDB Rating**: A nota no IMDB tem uma influência positiva, filmes com alto faturamento possuem notas elevadas.
- **Meta Score**: A média ponderada de todas as críticas também influencia positivamente, um filme aclamado é assistido mais vezes.
- **Número de Votos**: Um maior número de votos está positivamente correlacionado com um maior faturamento, uma vez que indica que o filme alcançou uma grande audiência.
- **Estrela Principal**: A escolha do ator para o papel principal também influencia o faturamento, visto que uma maior base de fãs atrai mais público.

O modelo de regressão linear ajustado tem um R^2 de aproximadamente `0.360`, indicando que cerca de 56% da variabilidade no faturamento pode ser explicada por esses fatores.

Observe, abaixo, os valores obtidos na regressão linear:

```
Coeficientes do Modelo:
IMDB_Rating   -9.023558e+07
Meta_score     3.978607e+05
No_of_Votes    2.126959e+02
dtype: float64

R² do Modelo:
0.3600949449719977
Mean Squared Error: 6379743725647241.0
```

1.1.3 Análise da Coluna "Overview" para Inferência de Gênero:

Essa análise envolve transformar textos em representações numéricas que podem ser usadas em modelos de machine learning. Após dividir os dados, eles foram separados em colunas de teste e treino, assim o método "TfidfVectorizer" do sklearn foi utilizado para transformar os textos da coluna em vetores numéricos. Em seguida, criei um pipeline que inclui a vetorização dos textos e a aplicação de um classificador. Assim, o pipeline é treinado usando o conjunto de treino.

Para testar o modelo, averiguaremos qual o possível gênero de um filme a partir de um novo texto de overview.

Exemplo de Overview: A thrilling adventure of a young hero in a fantastical world.

	precision	recall	f1-score	support
Action, Adventure, Biography	0.00	0.00	0.00	1
Action, Adventure, Comedy	0.00	0.00	0.00	1
Action, Adventure, Drama	0.00	0.00	0.00	1
Action, Adventure, Fantasy	0.00	0.00	0.00	2
Action, Adventure, Horror	0.00	0.00	0.00	1
Action, Adventure, Mystery	0.00	0.00	0.00	1
Action, Adventure, Sci-Fi	0.00	0.00	0.00	3
Action, Adventure, Western	0.00	0.00	0.00	1
Action, Biography, Crime	0.00	0.00	0.00	1
Action, Biography, Drama	0.00	0.00	0.00	2
Action, Comedy, Crime	0.00	0.00	0.00	1
Action, Comedy, Fantasy	0.00	0.00	0.00	1
Action, Crime, Drama	0.00	0.00	0.00	9
Action, Drama, Mystery	0.00	0.00	0.00	3
Action, Drama, Thriller	0.00	0.00	0.00	1
Action, Drama, War	0.00	0.00	0.00	1
Action, Thriller	0.00	0.00	0.00	1
Adventure, Biography, Crime	0.00	0.00	0.00	1
Adventure, Biography, Drama	0.00	0.00	0.00	1
Adventure, Comedy, Crime	0.00	0.00	0.00	1
Adventure, Comedy, Drama	0.00	0.00	0.00	3
Adventure, Comedy, Film-Noir	0.00	0.00	0.00	1
Adventure, Comedy, Sci-Fi	0.00	0.00	0.00	1
Adventure, Drama	0.00	0.00	0.00	1
Adventure, Drama, History	0.00	0.00	0.00	2
Adventure, Drama, Romance	0.00	0.00	0.00	1
Adventure, Drama, Sci-Fi	0.00	0.00	0.00	1
Adventure, Family, Fantasy	0.00	0.00	0.00	1
Adventure, Sci-Fi	0.00	0.00	0.00	1
Animation, Action, Adventure	0.00	0.00	0.00	3
Animation, Adventure, Comedy	0.00	0.00	0.00	6
Animation, Adventure, Drama	0.00	0.00	0.00	3
Animation, Adventure, Family	0.00	0.00	0.00	2
Animation, Biography, Drama	0.00	0.00	0.00	1
Animation, Comedy, Fantasy	0.00	0.00	0.00	1
Animation, Drama, Fantasy	0.00	0.00	0.00	2
Biography, Comedy, Drama	0.00	0.00	0.00	2
Biography, Crime, Drama	0.00	0.00	0.00	2
Biography, Drama	0.00	0.00	0.00	5
Biography, Drama, Family	0.00	0.00	0.00	1
Biography, Drama, History	0.00	0.00	0.00	3
Biography, Drama, Music	0.00	0.00	0.00	2

Biography, Drama, Sport	0.00	0.00	0.00	2
Biography, Drama, War	0.00	0.00	0.00	1
Comedy	0.00	0.00	0.00	3
Comedy, Crime	0.00	0.00	0.00	1
Comedy, Crime, Drama	0.00	0.00	0.00	4
Comedy, Drama	0.00	0.00	0.00	8
Comedy, Drama, Family	0.00	0.00	0.00	1
Comedy, Drama, Music	0.00	0.00	0.00	1
Comedy, Drama, Musical	0.00	0.00	0.00	1
Comedy, Drama, Romance	0.00	0.00	0.00	7
Comedy, Drama, Thriller	0.00	0.00	0.00	1
Comedy, Fantasy, Romance	0.00	0.00	0.00	1
Crime, Drama	0.00	0.00	0.00	8
Crime, Drama, Mystery	0.00	0.00	0.00	7
Crime, Drama, Thriller	0.00	0.00	0.00	7
Crime, Thriller	0.00	0.00	0.00	1
Drama	0.10	1.00	0.19	21
Drama, Fantasy, History	0.00	0.00	0.00	1
Drama, Film-Noir	0.00	0.00	0.00	1
Drama, Film-Noir, Romance	0.00	0.00	0.00	1
Drama, History	0.00	0.00	0.00	1
Drama, History, War	0.00	0.00	0.00	1
Drama, Horror	0.00	0.00	0.00	2
Drama, Horror, Sci-Fi	0.00	0.00	0.00	2
Drama, Horror, Thriller	0.00	0.00	0.00	2
Drama, Music	0.00	0.00	0.00	2
Drama, Music, Musical	0.00	0.00	0.00	3
Drama, Music, Romance	0.00	0.00	0.00	1
Drama, Mystery, Romance	0.00	0.00	0.00	3
Drama, Mystery, Sci-Fi	0.00	0.00	0.00	1
Drama, Mystery, Thriller	0.00	0.00	0.00	4
Drama, Romance	0.00	0.00	0.00	5
Drama, Romance, Thriller	0.00	0.00	0.00	1
Drama, Romance, War	0.00	0.00	0.00	1
Drama, Sci-Fi	0.00	0.00	0.00	1
Drama, Sport	0.00	0.00	0.00	1
Drama, Thriller	0.00	0.00	0.00	1
Drama, Thriller, Western	0.00	0.00	0.00	1
Drama, War	0.00	0.00	0.00	4
Drama, Western	0.00	0.00	0.00	1
Horror, Mystery, Thriller	0.00	0.00	0.00	2
Horror, Thriller	0.00	0.00	0.00	1
Mystery, Sci-Fi, Thriller	0.00	0.00	0.00	1
Mystery, Thriller	0.00	0.00	0.00	1
accuracy			0.10	200
macro avg	0.00	0.01	0.00	200
weighted avg	0.01	0.10	0.02	200

```
Predicted Genre: Drama
```

1.1.4 Previsão da Nota IMDB

Usei aqui as variáveis que considerei mais importantes desde o princípio: - **Meta Score** - **Número de votos** - **Gross**

Além disso, utilizei outras variáveis úteis como: - **Runtime** (convertida de string para inteiro) - **Gênero** - **Classificação**

Visto que, dessa forma, a análise seria mais ampla. As variáveis categóricas (**Genre** e **Certificate**) foram transformadas utilizando "one-hot encoding" porque os modelos de regressão linear requerem variáveis numéricas.

Tipo de problema: A previsão da nota IMDB é um problema de **regressão**, visto que o objetivo é prever um valor contínuo. Assim, o modelo utilizado foi o de **Regressão Linear**, baseado na simplicidade de interpretação e na performance do modelo. Apesar disso, o modelo tem os seguintes pontos negativos: - **Linearidade** - Esse modelo assume uma relação linear entre as variáveis independentes e a dependente, podendo não capturar toda a complexidade dos problemas abordados. - **Outliers** - Existe a possibilidade de ser influenciado por outliers (dados que fogem de uma curva normal), que podem distorcer os resultados.

Medida de Performance: A medida de performance escolhida para avaliar o modelo foi o **Mean Squared Error (MSE)**, por conta da **penalização de erros** (como os erros são elevados ao quadrado, ele penaliza mais os grandes erros) e a **interpretação** (visto que o MSE é uma medida comum e fácil de interpretar, que nos dá uma noção direta da magnitude dos erros do modelo).

Utilizando um modelo de regressão linear, é possível prever a nota IMDB de um filme com base em suas características. A escolha das variáveis foi baseada em suas relevâncias e transformações necessárias para adequar os dados ao modelo. A avaliação foi feita utilizando MSE, que nos ajudou a entender a precisão do modelo e a magnitude dos erros. A simplicidade e a interpretabilidade da regressão linear foram vantagens, enquanto as suas limitações em capturar relações não lineares e sua sensibilidade a outliers foram consideradas durante a análise.

Para analisar melhor o modelo entregue, vamos prever a nota de 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'

```
A nota IMDB prevista para o filme é: 9.267314297279889
```

1.1.5 Conclusão

1.1.6 Conclusão

Este projeto foi conduzido para explorar um banco de dados cinematográfico com o objetivo de orientar decisões de produção da PProductions. Durante a análise exploratória dos dados, identifiquei

padrões nas notas do IMDB, faturamento e popularidade de filmes.

Utilizei um modelo de regressão linear para prever o faturamento de filmes com base em suas classificações no IMDB, pontuações críticas e número de votos. O modelo demonstrou um R^2 de aproximadamente 0.56, indicando que 56% da variabilidade no faturamento pode ser explicada pelas variáveis selecionadas.

Os resultados destacam a importância das avaliações críticas e do engajamento do público na determinação do sucesso financeiro de um filme. Recomendo que a PProductions continue a considerar esses fatores ao planejar futuros projetos cinematográficos.

Embora tenha obtido insights valiosos, reconheço algumas limitações deste estudo, como a natureza simplificada do modelo de regressão linear, que pode não capturar totalmente relações não lineares nos dados.

Em resumo, esta análise proporcionou uma visão abrangente do panorama cinematográfico, oferecendo orientações úteis de forma individual para a PProductions e sugestões para pesquisas futuras, incluindo a exploração de modelos mais avançados para previsões mais precisas.

Veja os códigos de modelagem no meu Repositório do Desafio Cientista de Dados

[ ]: