Relatório de Análise Preditiva e Exploratória em Dados Cinematográficos

Análise Exploratória de Dados (EDA)

A questão central foi:
“Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?”

Para responder essa questão segui o passo a passo abaixo:

1. Limpeza de Dados

Após utilizar o comandos df.head( ) e df.info( ) para ter uma noção melhor das características do meu dataframe, notei a necessidade de:

- Remoção da coluna **Unnamed: 0** que não continha nenhum dado a ser usado na análise

- Conversão dos dados das colunas **Gross, Runtime e Released_Year** para valores numéricos dos tipos *float* e *int*

  A coluna **Released_Year** tinha um único dado do tipo *str* com valor **PG**, que foi substituído pelo dado real **1995** após uma busca rápida no próprio site do IMDB

- Tratamento de valores ausentes:
  - Uso de mediana para variáveis numéricas (colunas **Meta_score** e **Gross**)

  - Criação da categoria **“Not Rated”** para valores ausentes na coluna **Certificate**

2. Distribuição de Faturamento

Alguns insights foram descobertos ao investigar mais a fundo a correlação da coluna de faturamento com as demais variáveis dos filmes. São eles:

- A maioria dos filmes tem faturamento modesto
- Blockbusters criam uma **cauda longa** que distorce a média, mas são cruciais para o lucro do estúdio
- **IMDB Rating vs. Faturamento:** uma boa nota não garante sucesso financeiro
- **Número de Votos vs. Faturamento:** forte correlação positiva → filmes populares tendem a arrecadar mais
- **Gêneros mais lucrativos:** Ação, Animação e Aventura
- **Diretores e atores renomados**: Estão associados a filmes de maior faturamento

---

Recomendação de Filme

Uma outra questão proposta foi **"Qual filme você recomendaria para uma pessoa que você não conhece?"**

Para responder essa pergunta, utilizei os insights já mencionados e me vali da seguinte linha de raciocínio:

1. Aplicação de filtros:

- Popularidade alta (**top 25% em votos**)
- Qualidade elevada (**IMDb > 8.5**)
- Gêneros mais lucrativos
- Diretores e atores renomados
- Bons números de bilheteria

```
if not top_recommendation.empty:

          print(top_recommendation[['Series_Title',   'IMDB_Rating',
'Genre', 'Director', 'Star1', 'No_of_Votes', 'Gross']].to_string())
```

Recomendação principal: *The Departed (2006, Dir. Martin Scorsese)*, que atualmente está na posição 38 do top 250 do IMDB e foi vencedor do Oscar de Melhor Filme.

Análise de Texto

Uma outra questão proposta foi **"Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?"**

A coluna Overview contém texto não estruturado, o que exige uma abordagem diferente da análise de números. O objetivo é extrair significado de todas as sinopses. A técnica mais comum para começar é a Análise de Frequência, que dirá quais são os temas mais recorrentes nos filmes.

Para isso, é necessário seguir três etapas de pré-processamento de texto:

Etapa 1: Preparação do Ambiente e Limpeza de Texto
Instalar os pacotes necessários e preparar o texto de forma a manter apenas caracteres e palavras que contêm significado

Etapa 2: Análise de Frequência e Visualização
Determinar quantas e quais palavras observar

Etapa 3: Inferir o Gênero a partir da Sinopse
É possível, mas com limitações. A tarefa de prever uma categoria (como

o gênero) a partir de texto é um problema de **classificação** em Machine Learning. Para construir um modelo capaz de fazer essa previsão, precisaria de técnicas mais avançadas, como Vectorização de Texto e a construção de um Modelo de Classificação com um algoritmo apropriado para essa tarefa.

Na prática, a precisão do modelo dependeria de quão "únicos" os vocabulários de cada gênero são. Por exemplo, filmes de "guerra" provavelmente teriam palavras como "soldier" e "battle" em suas sinopses, enquanto filmes de "terror" teriam "fear" e "ghost". No entanto, muitos gêneros, como Drama e Romance, podem compartilhar um vocabulário similar, o que tornaria a inferência mais difícil.

De modo geral, embora o problema seja complexo, as ferramentas e a lógica para resolvê-lo estão disponíveis e provavelmente seriam a próxima fase do projeto.

Modelagem Preditiva

Metodologia e resultados

Objetivo: prever a nota do IMDB de um filme (regressão). Foram escolhidas como variáveis preditoras No_of_Votes, Gross, Meta_score, Genre, Director e Star1, sendo IMDB_Rating a variável alvo.

O pré-processamento agrupa as transformações em um pipeline: as variáveis categóricas (Genre, Director, Star1) são codificadas por One-Hot Encoding (com handle_unknown='ignore') via ColumnTransformer, enquanto as variáveis numéricas (No_of_Votes, Gross, Meta_score) são mantidas inalteradas. Como modelo, utilizou-se um Random Forest Regressor (100 árvores, random_state=42) por sua robustez a outliers e habilidade de capturar relações não lineares.

Os dados foram divididos em treino e teste (80/20, random_state=42) e o pipeline completo foi treinado e serializado em imdb_rating_predictor.pkl para uso em inferência. Ao carregar o pipeline e aplicar ao registro de The Shawshank Redemption (com conversão de Gross de string para float), a previsão gerada foi 8.77.

Observações e recomendações: validar o modelo com k-fold cross-validation, avaliar métricas (RMSE, MAE, R^2), inspecionar importâncias de variáveis (e/ou usar SHAP) e investigar possíveis problemas de target leakage (por exemplo, quando features como No_of_Votes só existem ou mudam após o lançamento).

Para produção, manter a serialização do pipeline evita divergências de pré-processamento entre treino e previsão.