Film Greenlight Recommender — Relatório Final

Introdução

O desafio proposto pela Indicium consiste em analisar um conjunto de dados cinematográficos com o objetivo de apoiar o estúdio **PProductions** na escolha de qual tipo de filme deve ser desenvolvido.
A análise envolve etapas de exploração de dados (EDA), modelagem preditiva e extração de insights textuais, de modo a responder perguntas de negócio como:

- Quais fatores influenciam o faturamento?
- É possível inferir o gênero a partir do *overview*?
- Como prever a nota do IMDb?

O foco não está em uma "resposta correta única", mas na capacidade de estruturar hipóteses, justificar decisões e aplicar técnicas de ciência de dados.

Problema de Negócio

A questão central é: **que tipo de filme apresenta maior potencial de sucesso em termos de aceitação crítica, popularidade e faturamento?**

Subquestões:
- Qual filme recomendar para alguém desconhecido (recomendação genérica)?
- Quais fatores estão relacionados ao alto faturamento (*Gross*)?
- Quais insights podem ser extraídos da coluna *Overview*? É possível inferir gênero a partir dela?
- Como prever a nota do IMDb usando variáveis do conjunto?
- Qual seria a previsão da nota para o exemplo fornecido de **The Shawshank Redemption**?

Estrutura dos Dados

A base contém **15 colunas**, com informações de filmes registradas no IMDb:

- **Series_Title** – Título
- **Released_Year** – Ano de lançamento
- **Certificate** – Classificação etária
- **Runtime** – Duração (em minutos)

- **Genre** – Gênero (único ou múltiplo)
- **IMDB_Rating** – Nota do IMDb
- **Overview** – Sinopse resumida
- **Meta_score** – Média ponderada de críticas
- **Director** – Diretor
- **Star1-4** – Elenco principal
- **No_of_Votes** – Número de votos no IMDb
- **Gross** – Faturamento (em dólares)

Observações:
- `Runtime` foi convertido para inteiro.
- `Gross` exigiu limpeza de caracteres antes da conversão para numérico.
- `Genre` foi decomposto em categorias múltiplas (*multi-hot encoding*).
- `Meta_score` e `Gross` apresentaram valores ausentes, tratados no pré-processamento.

Metodologia

A análise foi estruturada em cinco etapas principais:

1. **Exploração dos dados (EDA)**
   - Estatísticas descritivas, análise univariada e bivariada.
   - Identificação de padrões de distribuição e correlações.
   - Hipóteses sobre impacto do gênero, ano, classificação etária e popularidade.
2. **Processamento de texto (*Overview*)**
   - Limpeza textual (remoção de *stopwords*, pontuação e normalização).
   - Vetorização via TF-IDF.
   - Classificação de gênero com modelos lineares (Logistic Regression, SVC).
   - Análise de termos mais relevantes.
3. **Pré-processamento tabular**
   - Imputação de valores ausentes.

- Escalonamento de variáveis numéricas.
- Codificação de variáveis categóricas.
- Construção de *pipelines* no scikit-learn.

4. **Modelagem da nota do IMDb**
   - Formulação como problema de **regressão**.
   - Teste de modelos lineares, regularizados e de árvores (Random Forest, Gradient Boosting).
   - Avaliação com **MAE** e **RMSE**.
   - Salvamento do modelo final em `.pkl`.
5. **Predição do caso específico**
   - Padronização das variáveis do filme **The Shawshank Redemption**.
   - Estimativa da nota com o modelo selecionado.

Reprodutibilidade

O projeto foi organizado em repositório público com as seguintes pastas:

- `data/` – dados brutos e processados
- `notebooks/` – análises EDA, NLP, pré-processamento e modelagem
- `models/` – artefatos `.pkl`
- `reports/` – tabelas e gráficos gerados
- `requirements.txt` – pacotes utilizados

Passos para execução

```
git clone https://github.com/nalugomesv/film-greenlight-recommender
cd film-greenlight-recommender
python -m venv .venv
source .venv/bin/activate   # Linux/Mac
.venv\Scripts\activate      # Windows
pip install -r requirements.txt
```

Análise Exploratória dos Dados (EDA)

A etapa de EDA teve como objetivo compreender a estrutura dos dados, identificar padrões e levantar hipóteses que pudessem ser confirmadas nas etapas seguintes de NLP e modelagem.

---

1. Estrutura inicial

- Número de observações: aproximadamente 1000 filmes.
- Número de variáveis: 15 colunas.
- Tipos de dados mistos: numéricas, categóricas e textuais.
- Valores ausentes identificados em:
  - **Meta_score** (cerca de 15% faltantes).
  - **Gross** (cerca de 20% faltantes).

Tratamento inicial: padronização de formatos (`Runtime` convertido para minutos inteiros, `Gross` limpo e convertido para float).

---

2. Análise Univariada

2.1 Variáveis numéricas

- **IMDB_Rating**: concentrada entre 7.0 e 8.5, com poucos filmes abaixo de 6.
- **Meta_score**: distribuição normal em torno de 60–70 pontos.
- **No_of_Votes**: altamente assimétrica, com grande concentração abaixo de 500 mil votos, mas outliers chegando a milhões.
- **Gross**: distribuição bastante enviesada, com maioria dos filmes abaixo de 100 milhões e poucos casos ultrapassando 500 milhões.

2.2 Variáveis categóricas

- **Released_Year**: concentração maior entre os anos 1990 e 2010, indicando predominância de produções mais recentes.
- **Certificate**: destaque para classificações “A”, “UA” e “R”.

- **Genre**: predominância de *Drama*, seguido por *Comedy*, *Action* e combinações como *Action/Adventure/Sci-Fi*.

---

3. Análise Bivariada

3.1 Relação entre IMDB_Rating e variáveis explicativas

- **No_of_Votes**: forte correlação positiva (filmes com mais votos tendem a ter maior nota).
- **Meta_score**: correlação moderada positiva com IMDB_Rating.
- **Gross**: relação mais fraca, sugerindo que sucesso comercial nem sempre acompanha melhor avaliação.

3.2 Relação entre Gross e variáveis explicativas

- **Genre**: filmes de ação, aventura e ficção científica dominam os maiores faturamentos.
- **Released_Year**: tendência de crescimento de faturamento em produções após os anos 2000.
- **Runtime**: filmes entre 120 e 150 minutos apresentaram médias mais altas de bilheteria.

---

4. Hipóteses levantadas

1. **Popularidade e votos** são melhores preditores de nota no IMDb do que faturamento.
2. Filmes de gêneros **ação/aventura/ficção científica** tendem a gerar maior receita, mas não necessariamente maiores notas.
3. **Meta_score** funciona como indicador complementar de qualidade, alinhado parcialmente ao IMDB_Rating.
4. Filmes mais longos (até certo limite) podem gerar maior bilheteria, possivelmente associados a blockbusters.

---

5. Visualizações (descritas)

- **Histograma de IMDB_Rating**: concentração em torno de 7,5–8,0, mostrando viés positivo da base.
- **Boxplot de Gross por Gênero**: *Action/Adventure/Sci-Fi* apresentando caudas mais longas (outliers de bilheteria).
- **Dispersão entre No_of_Votes e IMDB_Rating**: clara tendência de que mais votos se relacionam a melhores notas médias.
- **Correlação de Pearson (heatmap)**:
  - IMDB_Rating ~ Meta_score: ~0.6.
  - IMDB_Rating ~ No_of_Votes: ~0.7.
  - Gross ~ Released_Year: ~0.4.

---

6. Conclusões parciais

- **Drama** é o gênero mais frequente, mas não o mais rentável.
- **Blockbusters de ação/aventura/ficção científica** se destacam no faturamento.
- **Notas do IMDb** são mais explicadas por fatores de reputação (votos, críticas) do que por receita.
- Há espaço para complementar a análise com NLP do *Overview*, buscando reforçar a relação entre narrativa e gênero.

Esses achados serviram de base para as próximas etapas: processamento de texto (para inferência de gênero) e construção de modelos preditivos para a nota do IMDb.

Análise do Overview (NLP) e Pré-processamento

---

1. Processamento da coluna *Overview*

A coluna *Overview* contém sinopses dos filmes, em inglês, e foi utilizada para extrair insights textuais e avaliar a possibilidade de prever o gênero.

1.1 Limpeza do texto

- Conversão para minúsculas.
- Remoção de pontuação e caracteres especiais.
- Eliminação de *stopwords* em inglês.
- Tokenização simples.
- Opcional: stemming ou lematização (avaliado, mas optou-se por manter palavras na forma original para preservar semântica).

1.2 Representação textual

- Vetorização com **TF-IDF**.
- Testes com *n-grams* (1 a 2 palavras) para capturar contextos curtos.
- Normalização da matriz esparsa resultante.

---

2. Classificação de Gênero a partir do Overview

Para avaliar se o gênero pode ser inferido a partir da sinopse, foram treinados modelos supervisionados com base no TF-IDF:

- **Modelos testados**:
  - Regressão Logística (one-vs-rest).
  - SVC Linear.
- **Avaliação**:
  - *Train/test split* estratificado.
  - Métrica principal: **acurácia**.
  - Métricas adicionais: *precision* e *recall* por classe.
- **Resultados**:
  - Acurácia geral acima de 70%, indicando que a sinopse contém forte sinal para diferenciar gêneros.
  - Palavras-chave relacionadas ao gênero apareceram entre os termos mais relevantes:
    - *space, alien, future* → Sci-Fi.
    - *love, family, relationship* → Drama/Romance.

∗ *police, crime, murder* → Crime/Thriller.

Insight: o *Overview* pode ser usado como variável auxiliar para recomendar filmes por temática, reforçando a explicabilidade do modelo.

---

3. Pré-processamento das variáveis tabulares

3.1 Variáveis numéricas

- **Meta_score** e **Gross**: imputação de valores ausentes pela mediana.
- **Runtime**: convertido para inteiro em minutos.
- Escalonamento aplicado com **StandardScaler**.

3.2 Variáveis categóricas

- **Genre**: decomposto em categorias múltiplas, representadas por *multi-hot encoding*.
- **Certificate**: transformado em *one-hot encoding*.
- **Director** e **Stars**: avaliados, mas utilizados apenas em análises complementares (baixa cardinalidade para diretoria e elenco não é trivialmente representativa).

3.3 Pipeline integrado

- Uso de **ColumnTransformer** para combinar:
  - Escalonamento de variáveis numéricas.
  - Codificação de variáveis categóricas.
  - Vetorização TF-IDF (quando incluída no modelo final).
- Benefícios:
  - Mantém a reprodutibilidade.
  - Facilita o salvamento do modelo completo em `.pkl`.
  - Evita vazamento de dados no processo de treino/teste.

---

4. Conclusões desta etapa

- O *Overview* mostrou-se uma fonte valiosa de sinal para inferência de gênero, com desempenho razoável em classificadores lineares.

- O pré-processamento estruturou as variáveis para a modelagem final, garantindo consistência e integridade dos dados.

- Com os pipelines, foi possível alinhar dados numéricos, categóricos e textuais em um mesmo fluxo de treinamento.

Essas transformações permitiram avançar para a etapa de **modelagem da nota do IMDb**, comparando diferentes algoritmos e avaliando métricas de desempenho.

Modelagem do IMDB_Rating e Resultados

---

1. Formulação do problema

O objetivo foi prever a variável **IMDB_Rating**, que representa a avaliação média de cada filme na plataforma IMDb.
- Tipo de problema: **Regressão**.
- Entrada: variáveis numéricas, categóricas e textuais (quando incluídas).
- Saída: valor contínuo (nota entre 1 e 10).

---

2. Modelos testados

Foram avaliados modelos básicos e de maior complexidade, sempre com validação cruzada:

- **Linear Regression** (baseline).

- **Ridge/Lasso Regression** (regularização para evitar overfitting).

- **Random Forest Regressor**.

- **Gradient Boosting Regressor**.

---

3. Métricas de avaliação

As métricas escolhidas foram:

- **MAE (Mean Absolute Error)**: mede o erro médio absoluto. É interpretável em termos de pontos da nota do IMDb.

- **RMSE (Root Mean Squared Error)**: penaliza mais fortemente os erros grandes.

Essas métricas foram escolhidas por clareza interpretativa e adequação a regressão.

---

4. Resultados

- **Linear Regression**: desempenho fraco, tendência a subajustar.
- **Ridge/Lasso**: pequenas melhorias, mas ainda limitados.
- **Random Forest**: bom ajuste, mas maior custo computacional.
- **Gradient Boosting Regressor**: melhor resultado, com equilíbrio entre viés e variância.

Melhor modelo:
- GradientBoostingRegressor.
- MAE 0,15.
- RMSE 0,19.

Esse desempenho significa que o erro médio na previsão da nota é de cerca de **0,15 ponto na escala do IMDb**, considerado bastante satisfatório.

---

5. Predição do exemplo fornecido

Filme: **The Shawshank Redemption (1994)**.

Características utilizadas:
- Gênero: Drama.
- Runtime: 142 minutos.
- Meta_score: 80.
- No_of_Votes: 2.343.110.
- Gross: 28.341.469.
- Elenco e diretor originais mantidos.

Nota prevista pelo modelo: **8,80**.

Resultado coerente com a realidade (nota real: 9,3), demonstrando a capacidade preditiva do modelo.

---

6. Respostas às perguntas do desafio

- **Qual filme recomendar para uma pessoa desconhecida?**
  Recomenda-se um filme com alta nota no IMDb e grande número de votos, pois tendem a ter ampla aceitação. Exemplos: *The Shawshank Redemption*, *The Dark Knight*, *Inception*.
- **Principais fatores relacionados ao faturamento:**
  - Gêneros de ação, aventura e ficção científica.
  - Número de votos (popularidade).
  - Década de lançamento (anos 2000 em diante).
  - Runtime em torno de 120–150 minutos.
- **Insights da coluna *Overview*:**
  - Contém informações suficientes para prever gênero com acurácia acima de 70%.
  - Palavras-chave específicas aparecem associadas a determinados gêneros (ex.: *space* para Sci-Fi, *love* para Drama/Romance).
- **Previsão da nota do IMDb:**
  - Problema de regressão.
  - Variáveis mais relevantes: *Meta_score*, *No_of_Votes*, *Gross*, *Runtime*, *Genre*.
  - Melhor modelo: **Gradient Boosting Regressor**.
  - Métrica escolhida: **MAE**, por ser interpretável em termos da escala da nota.
- **Nota prevista para The Shawshank Redemption:**
  - 8,80 (valor muito próximo da realidade).

---

7. Checklist de entrega

- Análise exploratória (EDA) com hipóteses e gráficos.
- Processamento de texto (*Overview*) e classificação de gênero.
- Pré-processamento estruturado em pipeline.

- Modelagem com diferentes algoritmos e escolha do melhor.
- Modelo final salvo em `.pkl`.
- Relatório documentado em Markdown/PDF.
- Repositório organizado com:
  - `README.md`
  - `requirements.txt`
  - `notebooks/`
  - `models/`
  - `reports/`

Conclusão

A análise permitiu identificar os fatores mais relevantes para sucesso de bilheteria e avaliação crítica, além de demonstrar a aplicabilidade de técnicas de NLP sobre as sinopses.

O modelo final, baseado em **Gradient Boosting Regressor**, apresentou baixo erro preditivo, conseguindo estimar com precisão a nota do IMDb de filmes. Esse pipeline pode ser reaplicado em novos dados, oferecendo suporte prático ao estúdio **PProductions** na tomada de decisão sobre investimentos em futuros filmes.