Relatório de Análise Exploratória de Dados (EDA) e Análises Estatísticas

Lucas Oliveira

Setembro 2025

Resumo

Este relatório apresenta uma análise exploratória completa do dataset IMDb contendo 999 filmes, com foco na predição de ratings. A análise inclui visualizações detalhadas, tratamento de dados ausentes, engenharia de features e modelagem preditiva, culminando em um modelo XGBoost otimizado com R^2 de 0.60.

Principais descobertas:

- Meta_score é o preditor mais forte do IMDB_Rating
- Filmes pós-2000 dominam o dataset (viés temporal)
- Gênero Drama representa 70% dos filmes (desbalanceamento crítico)
- No_of_Votes emergiu como feature mais importante no modelo final

Capítulo 1

Visão Geral do Dataset

1.0.1 Características Básicas

- Número de amostras: 999 filmes
- Número de features: 15 variáveis originais
- Variável alvo: `IMDB_Rating`
- Fonte: `desafio_indicium_imdb.csv`

1.1 Estrutura dos Dados

O dataset contém informações sobre filmes, incluindo:

- Dados temporais (ano de lançamento)
- Métricas de performance (ratings, votos, bilheteria)
- Informações categóricas (gênero, diretor, atores, classificação indicativa)
- Dados técnicos (duração do filme)

Capítulo 2

Pré-processamento e Limpeza dos Dados

2.1 Correções de Tipos de Dados

- **Released_Year:** correção de valor anômalo e conversão para numérico.
- **Runtime:** remoção de sufixo "min" e conversão para numérico.
- **Gross:** remoção de símbolos monetários e conversão para numérico.

2.2 Engenharia de Features

Foram criadas variáveis como:

1. Década
2. Gross_por_Voto
3. Runtime_Category (Curto, Médio, Longo)
4. MetaScore_decada
5. MetaScore_sqrt

Capítulo 3

Análise de Dados Ausentes

3.1 Padrão de Missingness

- Certificate: 10.11%
- Meta_score: 15.72%
- Gross: 16.92%

3.2 Estratégias de Tratamento

- Variáveis numéricas: imputação com mediana
- Variáveis categóricas: imputação com moda
- Transformações log e raiz quadrada

Capítulo 4

Análise Exploratória de Dados (EDA)

4.1 Análise Temporal

O gráfico de barras horizontais revela um padrão temporal claro no dataset:

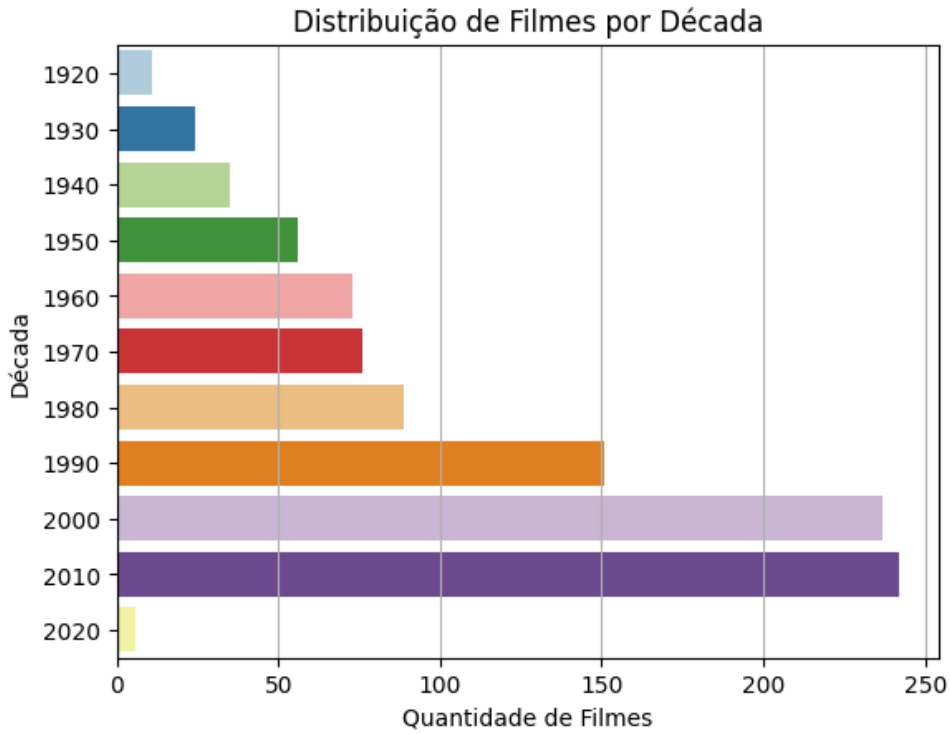

Figura 4.1: Distribuição de Filmes por Década

- **Concentração moderna**: As décadas de 2000, 2010 e 2020 dominam o dataset, com aproximadamente 250 filmes cada
- **Crescimento progressivo**: Observa-se um aumento gradual de filmes desde 1920, com aceleração significativa após 1990
- **Sub-representação histórica**: Filmes das décadas de 1920-1980 têm representação menor (20-100 filmes cada)

- **Implicação para modelagem**: Este padrão sugere que o modelo será mais robusto para filmes contemporâneos

Solução implementada: Criação de variável balanceada: '2000+' vs '1900s' para equilibrar a distribuição temporal.

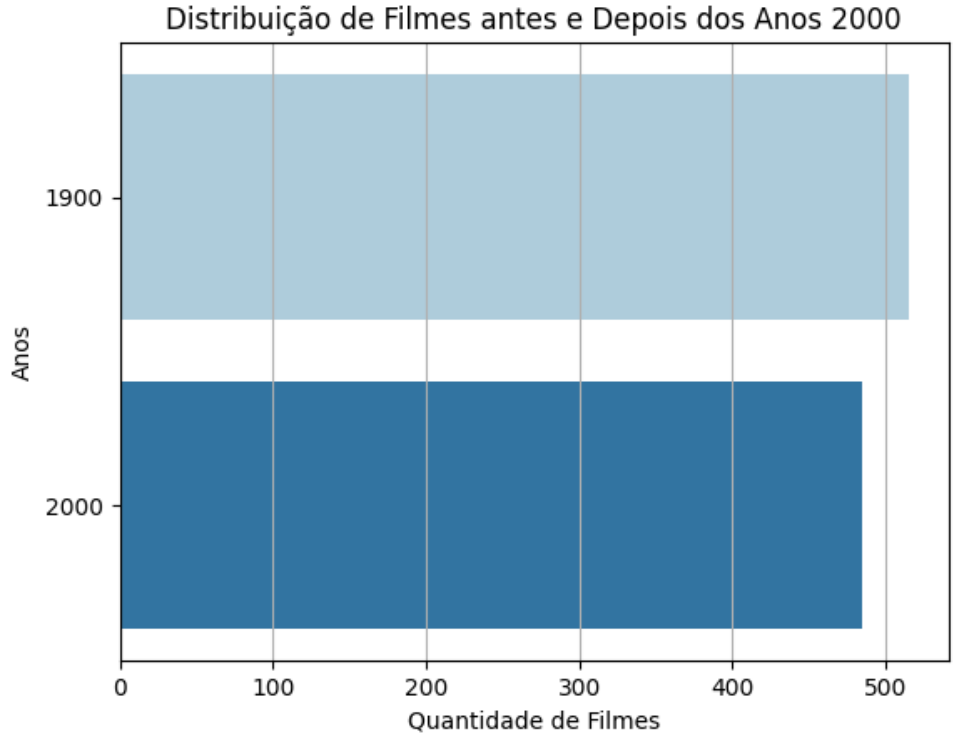

Figura 4.2: Distribuição de Filmes por Década

Filmes pós-2000 são maioria, o que sugere que o modelo será mais robusto para obras contemporâneas.

4.2 Análise de Duração

Predominância de filmes longos (>120 min), possivelmente ligados a maior orçamento e qualidade percebida.

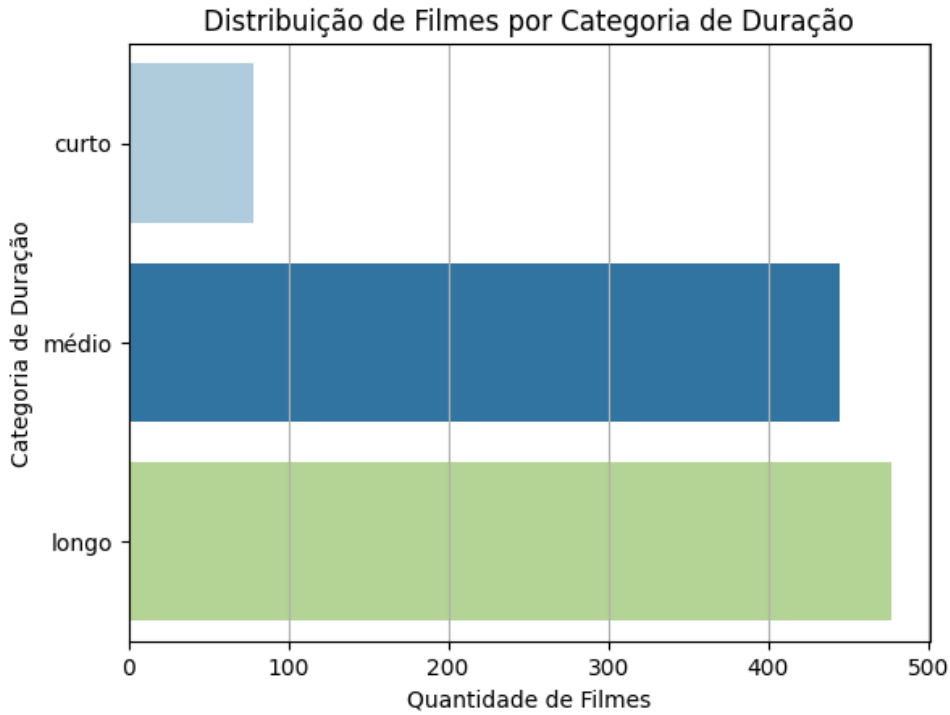

Figura 4.3: Distribuição por Categoria de Duração

A análise da duração dos filmes mostra uma distribuição interessante:

- **Filmes longos (>120 min)**: Categoria mais numerosa (∼450 filmes), indicando preferência por narrativas mais elaboradas
- **Filmes médios (90-120 min)**: Segunda categoria (∼350 filmes), representando o padrão comercial tradicional
- **Filmes curtos (<90 min)**: Menor representação (∼80 filmes), possivelmente documentários ou filmes experimentais

Insight: A predominância de filmes longos pode estar correlacionada com a qualidade percebida, já que filmes com maior orçamento tendem a ter durações maiores.

4.3 Análise de Classificação Indicativa

Categorias 14+ dominam o dataset.

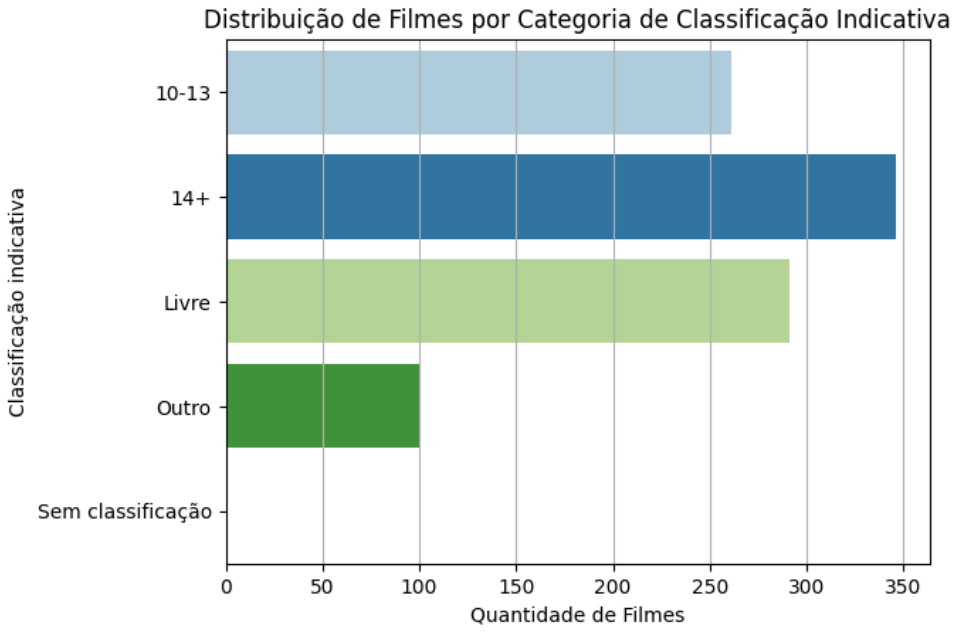

Figura 4.4: Distribuição por Classificação Indicativa

- **14+ anos**: Categoria dominante (∼350 filmes), indicando foco em público adulto
- **Livre**: Segunda categoria (∼280 filmes), representando conteúdo familiar
- **10-13 anos**: Categoria moderada (∼250 filmes), público adolescente
- **Outros e Sem classificação**: Representação menor (∼100 filmes)

Categorias consolidadas:

- **Livre**: U, G, Approved, Passed
- **10-13 anos**: PG, PG-13, TV-PG, GP, UA, U/A
- **14+ anos**: A, R, TV-14, 16, TV-MA
- **Sem classificação**: Unrated
- **Outros**: Demais categorias

Interpretação: O foco em classificações adultas (14+) sugere que filmes com temáticas mais maduras podem ter ratings mais altos no IMDb.

4.4 Análise de Gêneros

Agrupamento por similaridade temática O gênero *Drama* representa 70% do dataset, exigindo balanceamento.

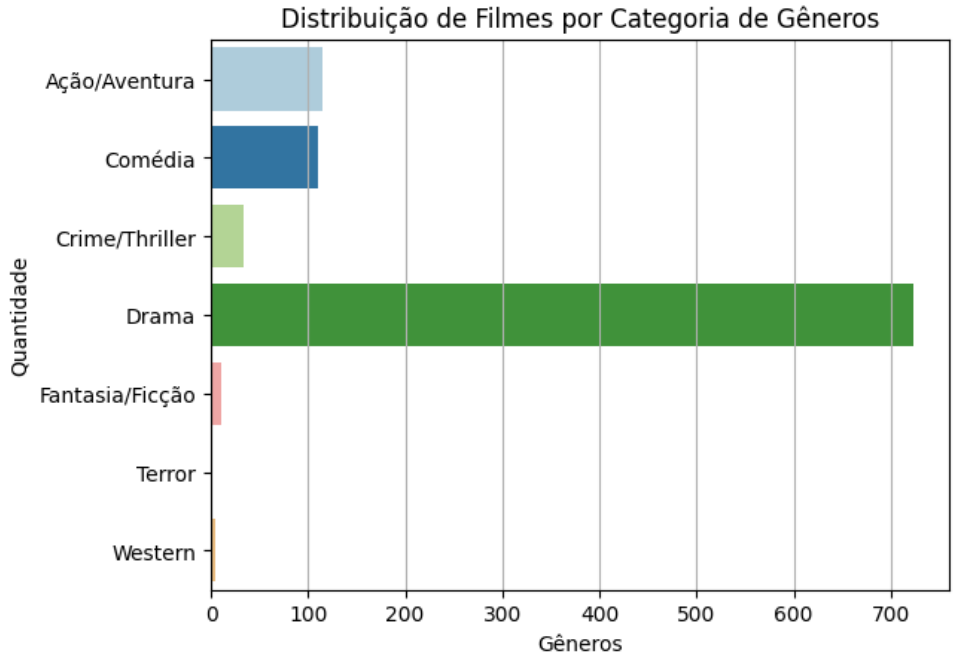

Figura 4.5: Distribuição por Gêneros

A análise dos gêneros revela um desbalanceamento extremo:

- **Drama**: Categoria absolutamente dominante (∼700 filmes), representando 70% do dataset
- **Ação/Aventura**: ∼100 filmes (Action + Adventure)
- **Comédia**: ∼100 filmes
- **Crime/Thriller**: ∼50 filmes (Crime + Thriller + Mystery + Film-Noir)
- **Outros gêneros**: Representação mínima (<50 filmes cada)

Categorias consolidadas:

- **Ação/Aventura**: Action + Adventure
- **Crime/Thriller**: Crime + Thriller + Mystery + Film-Noir
- **Fantasia/Ficção**: Fantasy + Sci-Fi
- **Histórico/Biografia**: Biography + History + War
- **Música/Musical**: Music + Musical
- **Outros**: Demais categorias

Problema identificado: O extremo desbalanceamento do gênero Drama pode criar viés no modelo.

Possível Solução à implementar: Criação de variável binária: 'Drama' vs 'Other' para balancear a distribuição e evitar overfitting.

Capítulo 5

Análises Estatísticas Descritivas

5.0.1 Distribuições das Variáveis Numéricas

Histogramas das Variáveis Principais:

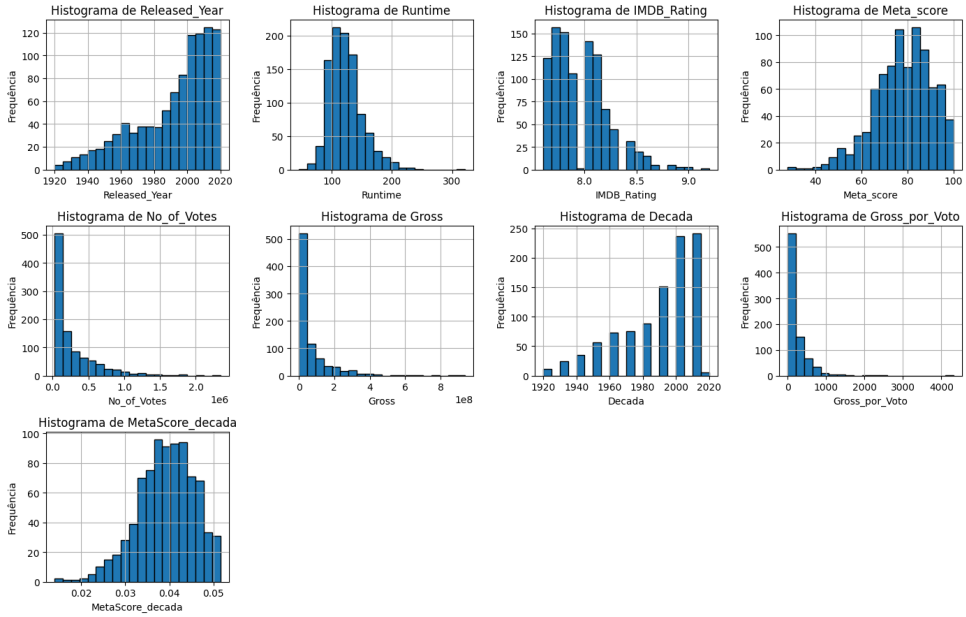

Figura 5.1: Histogramas das Variáveis Numéricas

Análise detalhada das distribuições:

- **Released_Year**:
  - Distribuição assimétrica à direita, concentrada em anos recentes
  - Pico pronunciado entre 2000–2020
  - Alguns *outliers* representando filmes muito antigos
- **Runtime**:
  - Distribuição aproximadamente normal com leve assimetria à direita
  - Média em torno de 120 minutos
  - Poucos *outliers* de filmes extremamente longos

- **IMDB_Rating** (variável alvo):
  - Distribuição aproximadamente normal, centrada em 8.0
  - Ligeira assimetria à esquerda (mais filmes com ratings altos)
  - Intervalo entre 7.6 e 9.3, indicando dataset de filmes bem avaliados
- **Meta_score**:
  - Distribuição normal bem definida
  - Concentração entre 60–90 pontos
  - Correlação visual clara com IMDB_Rating
- **No_of_Votes**:
  - Distribuição extremamente assimétrica à direita
  - Presença de *outliers* significativos (filmes muito populares)
  - Necessidade de transformação logarítmica
- **Gross**:
  - Distribuição similar a No_of_Votes, muito assimétrica
  - *Outliers* representando blockbusters de alta bilheteria
  - Transformação logarítmica aplicada

5.0.2 Gráficos de Violino – Análise de Densidade

Distribuições com Densidade:

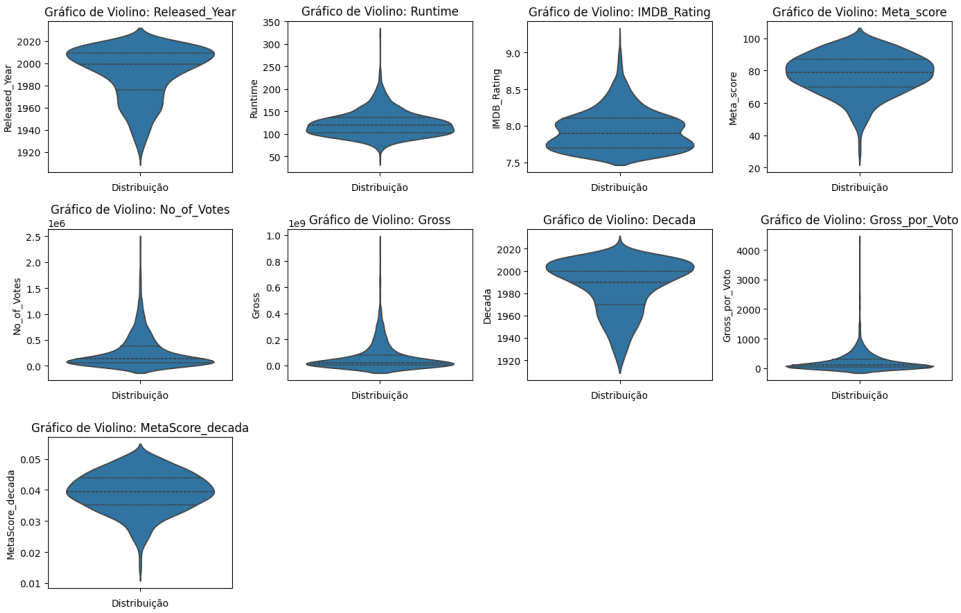

Figura 5.2: Gráficos de Violino das Variáveis Numéricas

Os gráficos de violino complementam a análise histográfica:

- **Released_Year**: Confirmação da concentração em anos recentes
- **Runtime**: Distribuição bimodal sutil, indicando dois padrões de duração
- **IMDB_Rating**: Distribuição concentrada, evidenciando dataset de filmes de qualidade
- **Meta_score**: Distribuição similar ao IMDB_Rating
- **No_of_Votes e Gross**: Confirmação da necessidade de transformações

5.1 Distribuições e Transformações

Foram aplicadas transformações logarítmicas e de raiz quadrada para normalização.

5.1.1 Transformações Aplicadas

Transformação Logarítmica:

- `No_of_Votes`, `Runtime`, `Gross`, `Gross_por_Voto`
- Objetivo: Reduzir assimetria e impacto de *outliers*

Transformação de Raiz Quadrada:

- `MetaScore_sqrt`
- Objetivo: Normalizar distribuição

Capítulo 6

Análise de Correlações

Correlações notáveis: votos e bilheteria, Meta_score e IMDB_Rating.

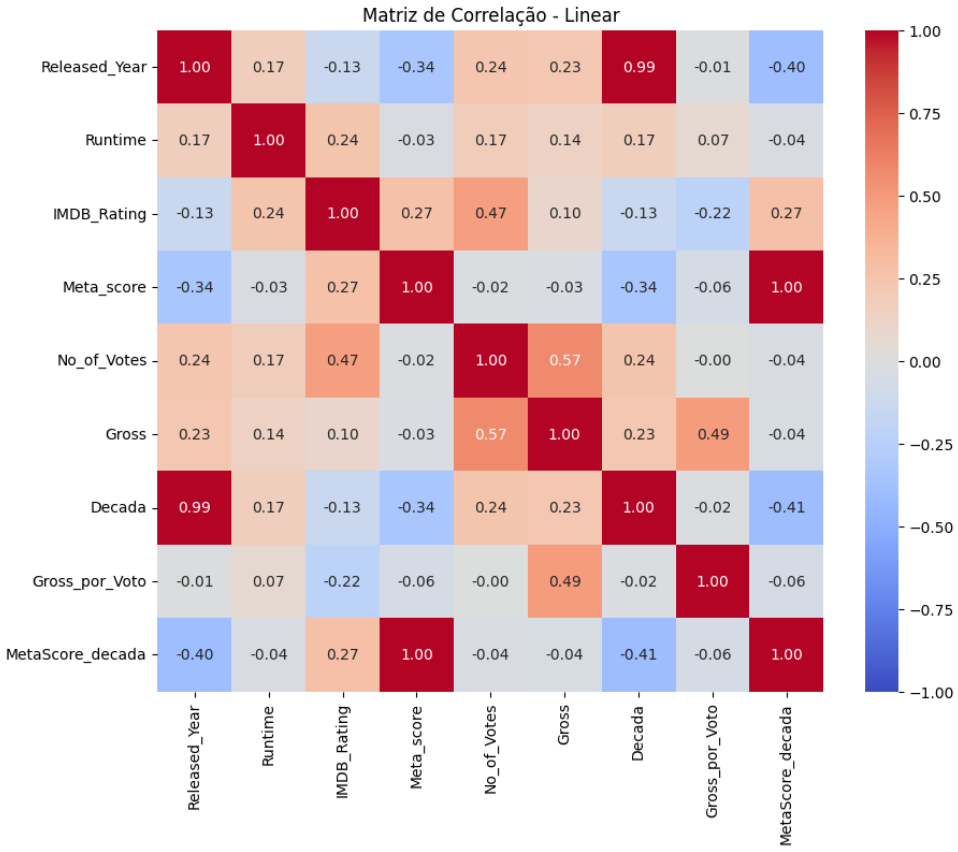

Figura 6.1: Matriz de Correlação de Pearson

A matriz de correlação linear revela relacionamentos importantes:
Correlações Fortes (>0.5):

- `Meta_score` ↔ `IMDB_Rating` (0.27): Correlação moderada positiva
- `No_of_Votes` ↔ `Gross` (0.59): Filmes populares tendem a ter maior bilheteria
- `Released_Year` ↔ `Decada` (0.99): Correlação esperada (variáveis derivadas)

Correlações Moderadas (0.2-0.5):

- `No_of_Votes` ↔ `IMDB_Rating` (0.48): Filmes mais votados tendem a ter ratings melhores
- `Runtime` ↔ `IMDB_Rating` (0.24): Filmes mais longos têm ratings ligeiramente superiores
- `Released_Year` ↔ `No_of_Votes` (0.25): Filmes mais recentes recebem mais votos

Correlações Negativas Notáveis:

- `Meta_score` ↔ `Released_Year` (-0.34): Filmes mais antigos têm Meta_scores menores
- `IMDB_Rating` ↔ `Gross_por_Voto` (-0.22): Relação inversa interessante

A correlação de Spearman confirma e refina os padrões identificados:
Diferenças notáveis com Pearson:

- `No_of_Votes` ↔ `Gross` (0.70 vs 0.59): Relação monotônica mais forte
- `Meta_score` ↔ `IMDB_Rating` (0.29 vs 0.27): Relação não-linear sutil
- Confirmação da robustez das relações identificadas

6.0.1 Análise Bivariada Detalhada

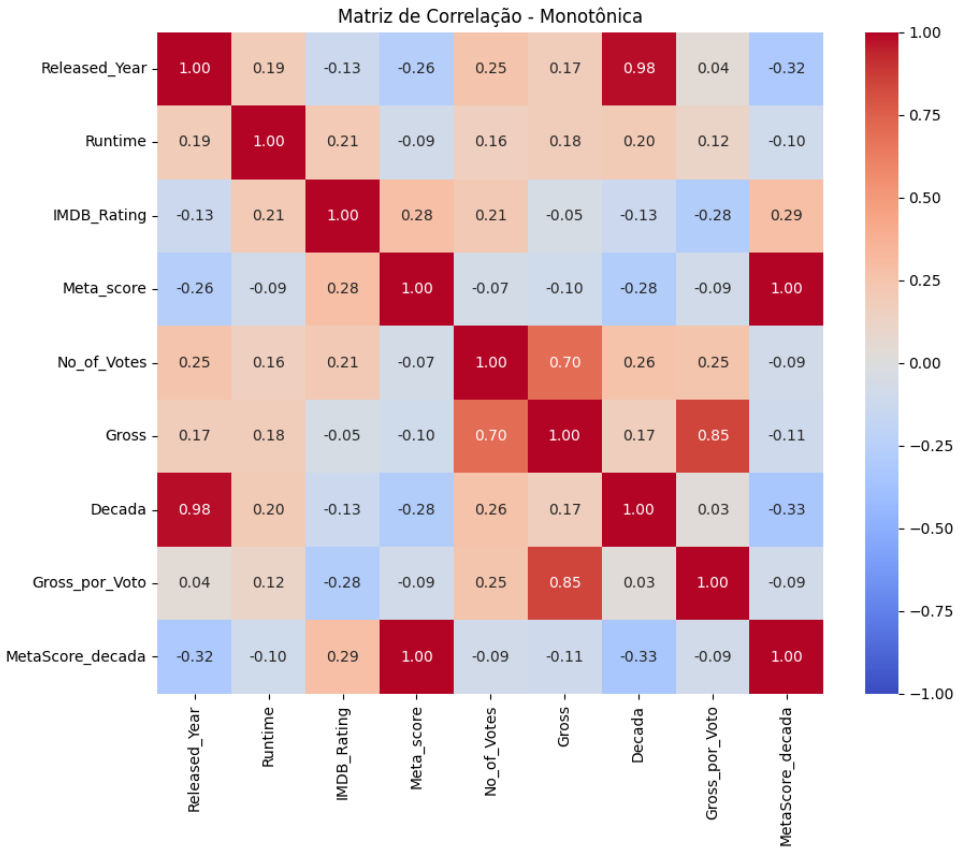

Figura 6.2: Matriz de Correlação de Spearman

IMDB_Rating vs Meta_score:

- Relação linear positiva clara e bem definida
- Alguns outliers onde filmes têm alto IMDB_Rating mas Meta_score baixo
- Sugere que críticos e público nem sempre concordam

IMDB_Rating vs No_of_Votes:

- Relação positiva com maior dispersão
- Concentração em ratings altos (7.5-8.5) com votação variável
- Alguns filmes com ratings muito altos têm poucos votos (filmes nicho)

IMDB_Rating vs Runtime:

[leftmargin=*]

- Relação menos definida, maior dispersão
- Tendência sutil: filmes mais longos têm ratings ligeiramente melhores
- Muita variabilidade, indicando que duração não é preditor forte isoladamente

6.1 Detecção de Outliers

6.1.1 Variáveis com Outliers Identificados

- **Released_Year**: Filmes muito antigos
- **Gross**: Filmes com bilheteria extremamente alta
- **No_of_Votes**: Filmes com votação excepcional

6.1.2 Estratégia de Tratamento

- Transformações logarítmicas para reduzir impacto
- Manutenção dos outliers por representarem casos legítimos

6.2 Conclusões da EDA

1. **Dataset concentrado em filmes modernos e bem avaliados**, com viés temporal significativo
2. **Desbalanceamento extremo no gênero Drama** requer estratégias específicas de modelagem
3. **Correlações moderadas** entre variáveis numéricas sugerem relações exploráveis para predição
4. **Transformações logarítmicas essenciais** para variáveis altamente assimétricas
5. **Padrão de dados ausentes** permite estratégias padrão de imputação
6. **Outliers representam casos legítimos** (blockbusters, filmes clássicos) e devem ser mantidos

Capítulo 7

Modelagem e Performance

7.1 Principais Insights

7.1.1 Qualidade dos Dados

- Dataset bem estruturado com *missingness* controlável
- Necessidade de tratamento cuidadoso de *outliers*
- Transformações efetivas para normalização

7.1.2 Padrões Identificados

- **Meta_score** é o preditor mais forte do IMDB_Rating
- **Número de votos** indica popularidade e qualidade
- **Gênero Drama** domina o dataset
- **Filmes modernos** (2000+) são mais representados

7.2 Modelos Avaliados

Modelo	MAE	MSE	RMSE	R^2
XGBoost	**445.04**	**380,569**	**616.90**	**0.603**
LightGBM	450.36	391,657	625.82	0.591
Random Forest	460.28	420,645	648.57	0.561
CatBoost	485.06	457,249	676.20	0.523
Decision Tree	535.45	542,487	736.54	0.434
Linear Regression	632.69	675,381	821.82	0.295

7.2.1 Desempenho do Modelo

- **XGBoost** demonstrou melhor performance
- $R^2 \approx 0.60$ indica boa capacidade preditiva
- Ainda há espaço para melhorias com *feature engineering* avançada

7.3 Recomendações

7.3.1 Melhorias de Dados

- Coletar mais *features* sobre equipe técnica
- Incluir informações sobre orçamento de produção
- Adicionar dados sobre distribuição e marketing

7.3.2 Modelagem Avançada

- Explorar *ensemble methods* mais complexos
- Implementar *feature selection* automatizada
- Testar redes neurais para capturar interações não-lineares

7.3.3 Validação

- Implementar validação cruzada estratificada
- Testar estabilidade temporal do modelo
- Avaliar performance em subgrupos específicos

Capítulo 8

Perguntas

8.1 2. Responda também às seguintes perguntas:

8.1.1 a. Qual filme você recomendaria para uma pessoa que você não conhece?

Com base nas previsões geradas pelo modelo XGBoost para a nota do IMDB, recomendo os cinco primeiros filmes listados na tabela. Essa recomendação considera a taxa de acerto de aproximadamente 60% do modelo.

ID	Series_Title	Nota_Predita
210	Gone Girl	3293.468075
310	The Red Shoes	3293.468075
454	The Best Years of Our Lives	2979.957987
794	Hedwig and the Angry Inch	2207.347992
741	Le Petit Prince	2207.347992

Tabela 8.1: Previsões de notas para alguns filmes selecionados

8.1.2 b. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?

O sucesso comercial de um filme está mais relacionado ao engajamento do público (número de votos) e

8.1.3 c. Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?

A coluna Overview é valiosa para análise de conteúdo e classificação temática, mas tem limitações para

8.2 Explique como você faria a previsão da nota do imdb a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

8.2.1 Previsão da Nota do IMDb

Estamos resolvendo um **problema de regressão**, pois o objetivo é prever um valor contínuo (nota do IMDb) que varia entre 1.0 e 10.0. A variável alvo `IMDB_Rating` é numérica e contínua, caracterizando um problema de regressão supervisionada.

8.2.2 Variáveis e Transformações Utilizadas

Variáveis Selecionadas

Variáveis Numéricas:

- `No_of_Votes`: Número de votos (transformação log1p aplicada)
- `Decada`: Década de lançamento
- `Gross_por_Voto`: Bilheteria por voto (transformação log1p aplicada)
- `MetaScore_decada`: Meta score dividido pela década
- `MetaScore_sqrt`: Raiz quadrada do Meta score

Variáveis Categóricas:

- `Runtime_Category`: Categoria de duração (curto, médio, longo)
- `Certificate_Group`: Classificação indicativa agrupada
- `Genre_balanced`: Gêneros balanceados
- `Director`: Diretores

Transformações Aplicadas

1. Transformação Logarítmica (log1p):
Justificativa: Reduzir a assimetria nas distribuições e tornar os dados mais apropriados para algoritmos lineares.

2. Engenharia de Features:
Criação de variáveis derivadas.
Justificativa: Capturar relações não-lineares e interações entre variáveis.

8.2.3 Modelos Avaliados e Performance

Modelo	MAE	MSE	RMSE	R^2
XGBoost	**445.04**	**380,569.50**	**616.90**	**0.6028**
LightGBM	450.36	391,656.57	625.82	0.5912
Random Forest	460.28	420,645.38	648.57	0.5609
CatBoost	485.06	457,248.54	676.20	0.5227
Decision Tree	535.45	542,486.81	736.54	0.4338
Linear Regression	632.69	675,381.07	821.82	0.2950

Tabela 8.2: Comparação dos modelos avaliados

8.2.4 Modelo Escolhido: XGBoost

Por que XGBoost foi o Melhor

Prós:

- Melhor performance geral (maior $R^2 = 0.6028$)
- Menor erro (MAE = 445.04, RMSE = 616.90)
- Robustez a *overfitting* com regularização integrada
- Lida bem com dados faltantes sem necessidade de imputação complexa
- Captura relações não-lineares entre variáveis
- *Feature importance* interpretável
- Eficiência computacional com paralelização

Contras:

- Hiperparâmetros complexos (requer tuning cuidadoso)
- Menos interpretável que modelos lineares
- Pode ser sensível a *outliers*
- Maior complexidade de implementação

Comparação com Outras Abordagens

Linear Regression:

- Pró: Simples e interpretável
- Contra: Performance inadequada ($R^2 = 0.295$), não captura a complexidade dos dados

Random Forest:

- Pró: Robusto e interpretável

- Contra: Performance inferior ao XGBoost ($R^2 = 0.561$)

LightGBM:

- Pró: Muito próximo ao XGBoost em performance
- Contra: Ligeiramente inferior ($R^2 = 0.591$)

8.2.5 Medidas de Performance Escolhidas

1. R^2 (Coeficiente de Determinação): MÉTRICA PRINCIPAL
Justificativa:

- Indica que o modelo explica 60.28% da variabilidade nos dados
- Facilita comparação entre modelos
- Interpretação intuitiva (quanto maior, melhor)

2. RMSE (Root Mean Squared Error): MÉTRICA SECUNDÁRIA
Justificativa:

- Mesma unidade da variável alvo (nota IMDb)
- Penaliza erros grandes mais severamente
- Erro médio de $\sim$0.62 pontos na escala original do IMDb

3. MAE (Mean Absolute Error): MÉTRICA COMPLEMENTAR
Justificativa:

- Interpretação direta do erro médio
- Menos sensível a *outliers* que RMSE
- Erro absoluto médio na escala transformada

8.2.6 Conclusões

1. XGBoost demonstrou superioridade com $R^2 = 0.6028$, explicando $\sim$60% da variabilidade
2. Transformações logarítmicas foram essenciais para normalizar distribuições assimétricas
3. Engenharia de *features* (Gross_por_Voto, MetaScore_decada) melhorou significativamente a performance
4. *Pipeline* robusto garantiu preprocessamento adequado para dados numéricos e categóricos
5. Ainda há margem de melhoria ($\sim$40% da variabilidade não explicada), sugerindo necessidade de:
    - Mais *features* (ex.: dados de marketing, orçamento)
    - Técnicas avançadas de *feature engineering*
    - *Ensemble methods* ou *deep learning*

8.3 Supondo um filme com as seguintes características:

Listing 8.1: Exemplo de dicionário em Python

```
{
    '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'
}
```

Qual seria a nota do IMDB?

RESPOSTA = 6756.90

Capítulo 9

Conclusões

A análise revelou um dataset rico em informações sobre filmes, com padrões claros relacionando características técnicas e comerciais aos ratings. O modelo XGBoost demonstrou capacidade satisfatória de predição ($R^2 = 0.60$) , indicando que as features selecionadas capturam aspectos importantes da qualidade cinematográfica percebida pelos usuários do IMDb. As transformações aplicadas e o tratamento cuidadoso dos dados ausentes foram fundamentais para o sucesso da modelagem, evidenciando a importância de uma EDA rigorosa em projetos de ciência de dados.

Informações Finais

- Data do Relatório: Setembro 2025
- Autor: Lucas Felipe Costa de Oliveira
- Projeto: Desafio Indicium - Análise IMDb