RELATÓRIO TÉCNICO - PROJETO LIGHTHOUSE

Resumo técnico

Este trabalho investiga quais atributos de um filme estão associados a desempenho comercial e crítico, e desenvolve protótipos de modelos para duas **tarefas centrais**: (i) **inferir gêneros** a partir da sinopse (H9) e (ii) **prever a nota do IMDb** (H11).

A análise começa com a base Lighthouse (999 linhas, 15 colunas) e evolui para um **dataset ampliado do Kaggle** (+44 mil filmes), permitindo EDA robusta e modelos mais estáveis. O fluxo de dados, a engenharia de atributos e o desenho de validação priorizam clareza, reprodutibilidade e comparabilidade.

Em negócios, os achados indicam que **orçamento** e **engajamento** (popularidade, votos) explicam boa parte da **receita bruta**, enquanto **gênero** e **janela** modulam o resultado e o risco. Em portfólio, animações e títulos familiares sustentam bilheteria, enquanto horror, crime e Mystery costumam oferecer **ROI** superior por requererem menos investimento.

1. Entendimento do problema e escopo

A PProductions precisa decidir **que filmes produzir** reduzindo incerteza em um mercado de alto risco financeiro. O objetivo técnico é transformar dados históricos em **insumos quantitativos** que ajudem a responder:

a. Que filme recomendar a um público genérico;
b. Quais fatores se associam a **alta expectativa de faturamento**;
c. O que a **sinopse** revela (e se permite inferir **gêneros**);
d. Como **prever a nota do IMDb**, com variáveis e métricas justificadas;
e. Qual seria a **nota prevista** para *The Shawshank Redemption*;
f. Como **entregar o modelo** salvo e reprodutível.

Do ponto de vista analítico, tratamos os resultados como **associações** — não causalidade. O foco está em **comparabilidade** (mesma preparação para treino/validação/teste), **robustez** (evitar sobre ajuste, especialmente em base pequena) e **utilidade** (entregáveis claros: notebook final, relatórios, modelos salvos).

2. Dados e decisões de fonte

A base Lighthouse, embora curada e didática, é **pequena** para treinar modelos preditivos confiáveis, especialmente com variáveis categóricas de alta cardinalidade (diretores, elencos) e texto (sinopses). Por isso, após higienização e um **match exploratório** com títulos Kaggle (cobertura final ~82%), o projeto passa a **trabalhar diretamente no Kaggle** para EDA e modelagem, conforme registrado nos notebooks K01 (consolidação), K02 (EDA) e K03 (modelagem).

- **Lighthouse (999 × 15)**: leitura, pré-limpeza mínima, criação de chaves (title_norm, Year) e derivados simples (Runtime_min, Gross_USD).
- **Kaggle (~45k filmes)**: consolidação de múltiplas tabelas (metadata, credits, links/ratings), normalização de tipos e criação de uma versão **otimizada (leve)** para análise/modelagem, com overview, genres_list, original_language, vote_average, vote_count, budget, revenue, runtime, popularity, além de diretor e principais atores extraídos de credits.
- **Decisão de fonte para modelagem**: usar **Kaggle otimizado** como base principal. A motivação é reduzir **viés de amostra** e aumentar **variabilidade**, mitigando sobre ajuste e tornando métricas de validação mais informativas.

```
25
26 print(f"[MATCH 2] Ganho via fallback (original_title_norm, Year): {match_via_fallback:.1%}")
27 print(f"[TOTAL ] Cobertura final (qualquer chave): {match_rate_total:.1%}")
✓ 0.2s
[MATCH 2] Ganho via fallback (original_title_norm, Year): 8.7%
[TOTAL ] Cobertura final (qualquer chave): 82.2%
```

3. Preparação e engenharia de atributos

A preparação busca equilíbrio entre **simplicidade** e **poder preditivo**, com as mesmas regras aplicadas a treino, validação e teste por meio de Pipeline/ColumnTransformer.

Normalizações e derivados:

Os campos financeiros (budget, revenue) são convertidos para numéricos coerentes; criam-se log_budget e log_revenue para suavizar escala e roi = revenue / budget para capturar eficiência.

Datas alimentam year e release_season (estações do Hemisfério Norte), apoiando análise de janela.

O texto overview é padronizado e gera overview_len como proxy simples de conteúdo.

Países e gêneros viram listas limpas; derivam n_countries, n_genres e genre_primary.

Categóricas com cardinalidade controlada:

Para original_language, genre_primary, director e star1..star4, usa-se **top-K** frequente com classe **Other**, reduzindo ruído e aliviando explosão no One-Hot. Essa decisão melhora estabilidade e tempo de treino sem perder os sinais dominantes (ex.: "star power" como efeito agregado dos nomes mais recorrentes).

Texto (overview):

O texto é representado com **TF-IDF** (1–2-gram), vocabulário limitado — opção deliberada pela **interpretação simples** e custo computacional baixo, além de integrar bem com modelos lineares e de árvores.

Partição de dados:

Quando disponível, o conjunto é dividido **temporalmente** por year (aprox. 70% treino | 15% validação | 15% teste via quantis), respeitando o fluxo do tempo e reduzindo *look-ahead bias*. Na ausência de year, usa-se particionamento estratificado/aleatório com as mesmas proporções. As métricas de referência em regressão são **RMSE** (prioritária), **MAE** e **R^2**; em multilabel, **F1-micro** e **F1-macro** (além de *subset accuracy* e *hamming loss* quando conveniente).

4. EDA — Achados-chave para negócios (H1–H8)

A exploração confirma padrões esperados e revela nuances relevantes para o portfólio. O quadro a seguir foi registrado em *reports/hypotheses_summary.csv* e reapresentado no notebook final.

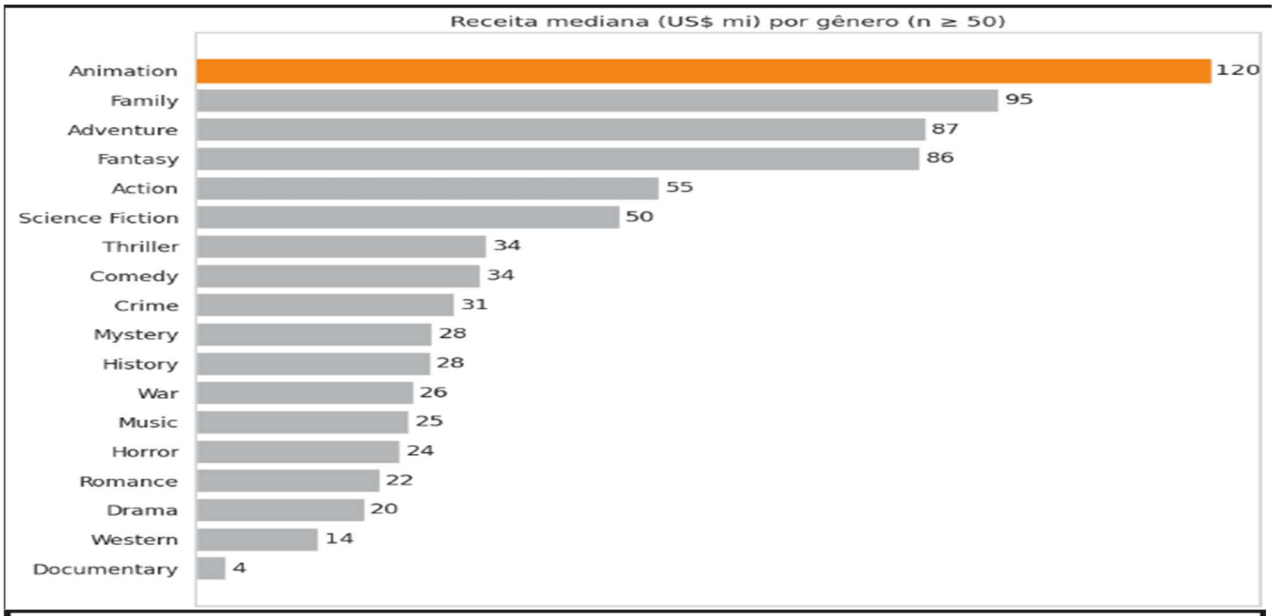
Hipótese 1 - mostra associação clara entre **orçamento** e **receita bruta**: à medida que o investimento cresce, também cresce a arrecadação média. Esse efeito, porém, **não garante ROI** — projetos caros podem falhar, enquanto propostas enxutas podem multiplicar capital investido.

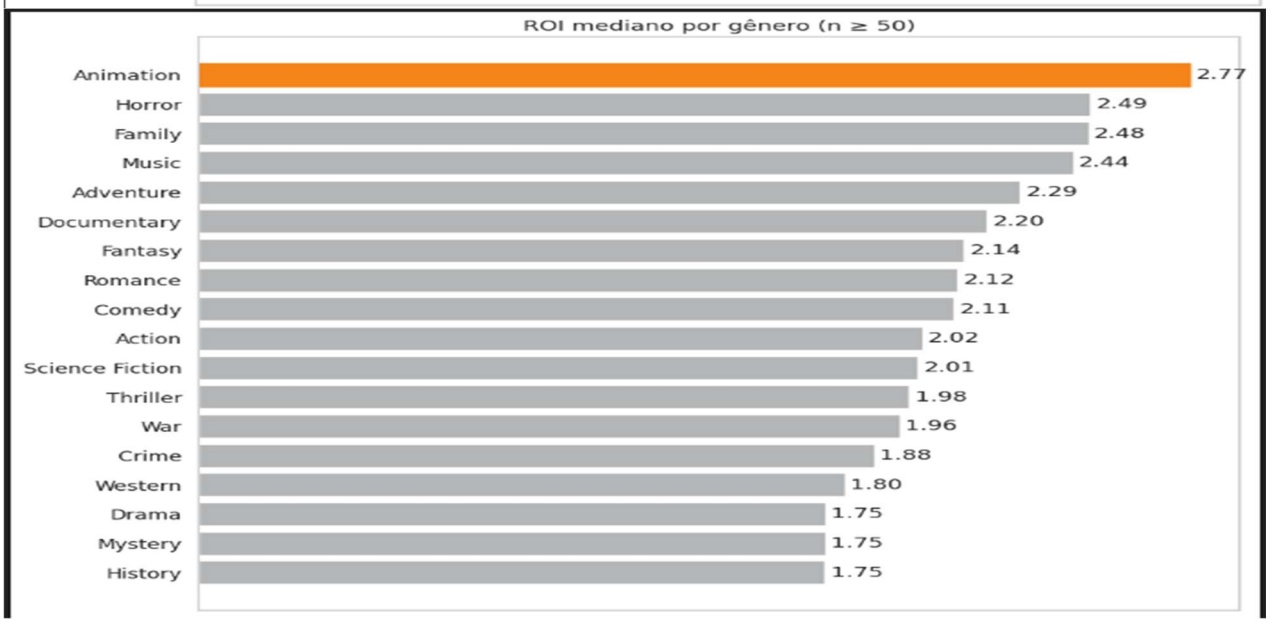
Hipótese 2 - indica que **popularidade** (e votos) funciona como **indicador antecedente** de performance: títulos com tráfego e engajamento mais altos tendem a converter melhor comercialmente. Esse sinal é útil para calibrar marketing e distribuição antes e durante a janela de lançamento.

Hipótese 3 - sugere que notas **IMDb** mais altas se associam a maior **sobrevida** do filme, mas **não asseguram** grande bilheteria. O público consome vários perfis de produto; reputação e faturamento caminham juntos em alguns nichos, mas divergem em outros.

Hipótese 4 – não se encontra evidência robusta de que **metragens muito longas** reduzam a nota de forma consistente; há efeito fraco/indefinido, dependente do gênero e do apelo do conteúdo.

Hipótese 5 - confirma que **gênero importa, mas com algumas considerações**. **Animation/Family/Adventure** lideram **receita** e exibem **consistência**, sustentadas por orçamentos elevados e apelo global. **Horror/Crime/Mystery** costuma entregar **ROI** superior por exigirem menos capital, embora com **maior variância**: há sucessos extremos e também muitos medianos. A escolha ótima depende do objetivo: **volume de bilheteria**, **retorno proporcional** ou **prestígio** crítico.



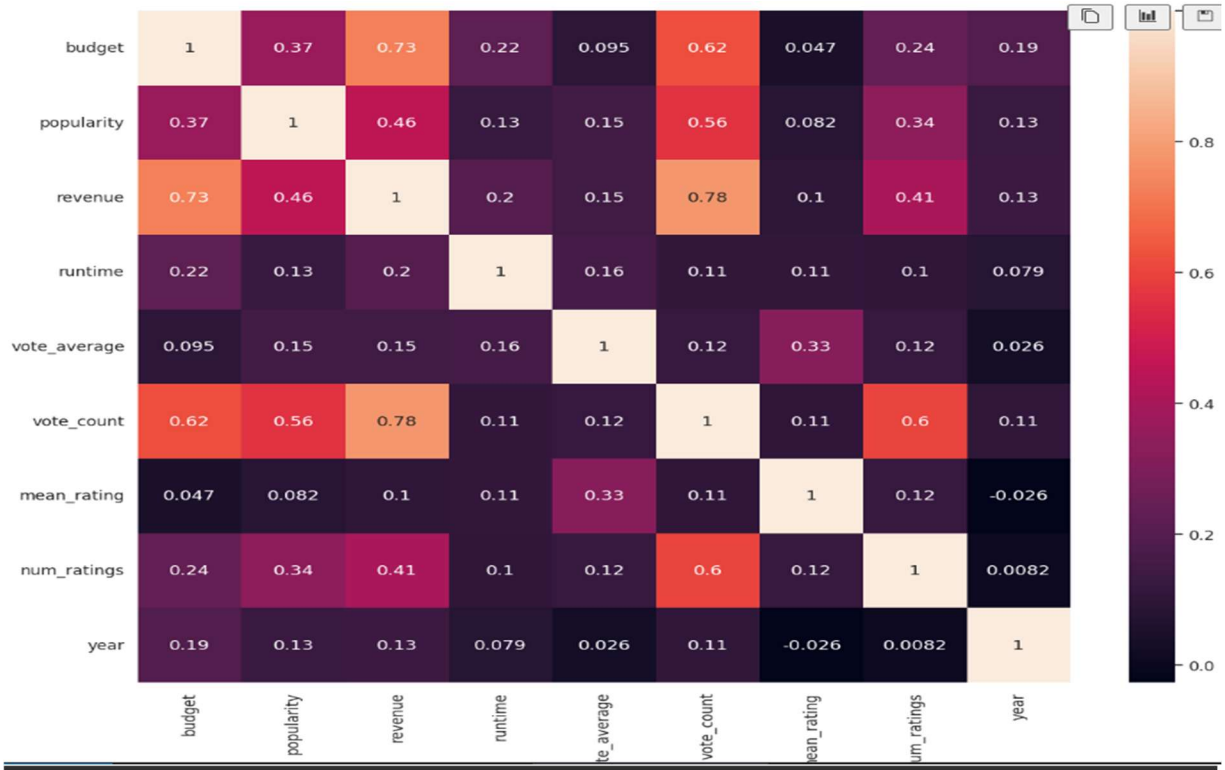


Hipótese 6 - aponta que **diretores e elenco** estão associados a melhor desempenho, sobretudo quando presentes entre os mais frequentes/renomados. O efeito, no entanto, é condicionado por custo e gênero.

Hipótese 7 - sugere, para recomendação a público desconhecido, priorizar títulos com **alta nota e muitos votos**, isto é, consenso social. Exemplos clássicos como *Dilwale Dulhania Le Jayenge* (1995) e *The Shawshank Redemption* (1994) aparecem naturalmente nessa regra.

Hipótese 8 - mostra que, **após 2010**, houve aumento de **engajamento** e queda da **mediana de receita**, sinalizando **fragmentação** do consumo e competição do **streaming**.

Para referência executiva, os trechos numéricos consolidados por gênero (receita, ROI, nota) e por variáveis críticas (popularidade, votos) foram exportados do notebook final como figuras e tabelas.



	H	Status	Resumo
0	H1	⚠ parcial	Orçamento ↑ = Receita ↑; ROI não garantido.
1	H2	✅	Popularidade antecipa bilheteria (bom leading indicator).
2	H3	❌	Nota IMDb ↑ ajuda longevidade/atração, mas não garante bilheteria.
3	H4	⚠ fraco/indefinido	Runtime alto (>150m) com evidência não robusta sobre nota.
4	H5	✅ com nuances	Animation/Family/Adventure sustentam receita/consistência (alto orçamento/ROI moderado); Horror/Crime/Mystery tendem a maior ROI (mais voláteis). Escolha depende de receita × ROI × prestígio.
5	H6	⚠ parcial	Diretor/Elenco ajudam receita; ROI depende do custo.
6	H7	✅	Recomendação p/ desconhecido: alta nota + muitos votos (ex.: DDLJ (1995), Shawshank (1994)).
7	H8	✅	Pós-2010: engajamento ↑, mediana de receita ↓ (fragmentação/streaming).

5. Inferência de gêneros a partir da sinopse (overview) – H9

O objetivo desta seção é avaliar a viabilidade de prever o(s) gênero(s) de um filme usando exclusivamente o texto da sinopse. A motivação de negócio é dupla: acelerar a curadoria de catálogo com um pré-tagueamento automático e criar um insumo que ajude a classificar propostas em estágios iniciais, quando ainda não existem outros metadados confiáveis.

A base de trabalho reúne mais de quarenta mil títulos com sinopse e rótulos de gênero. Foram removidos registros sem overview ou sem gêneros, padronizados os rótulos como listas e excluídas classes muito raras para reduzir ruído (frequência < 50).

O texto foi transformado em representações numéricas via **TF-IDF** com unigramas e bigramas, limitando o vocabulário para controlar tempo e memória. Os rótulos foram binarizados em formato **multilabel**.

Testamos alternativas. A simplificação “multiclasse” (usar apenas o primeiro gênero da lista) apresentou viés e baixo desempenho macro. Abordagens com balanceamento (class_weight/oversampling) trouxeram ganhos marginais. **LSA** reduziu dimensionalidade, mas

diminuiu *recall* nas classes médias/raras. A melhor combinação veio com **One-Vs-Rest + Logistic Regression** sobre TF-IDF, calibrando **threshold** e **top-k** para cada instância, solução que preserva interpretabilidade (pesos por termo), tem custo computacional razoável e generaliza melhor entre classes.

Os resultados globais foram **F1-micro = ~0,568**, **F1-macro = ~0,303**, **Subset Accuracy = 0,12** e **Hamming Loss = 0,06**. Em termos práticos, o modelo acerta pelo menos um gênero relevante em pouco mais da metade dos filmes, com melhor desempenho nas classes mais representadas (**Drama, Comedy, Documentary, Horror, Action**) e dificuldades nas classes raras (**Fantasy, History, Foreign, TV Movie**).

```
F1-micro: 0.568
F1-macro: 0.303
Subset accuracy: 0.102
Hamming loss: 0.066

Relatório por classe:
                                    precision    recall  f1-score   support

                            Action       0.55      0.55      0.55      1275
                         Adventure       0.49      0.22      0.31       687
                         Animation       0.81      0.24      0.37       372
                           Aniplex       0.00      0.00      0.00         0
                        BROSTA TV        0.00      0.00      0.00         0
                Carousel Productions     0.00      0.00      0.00         0
                            Comedy       0.51      0.76      0.61      2557
                             Crime       0.56      0.35      0.43       872
                       Documentary       0.71      0.74      0.73       759
                             Drama       0.58      0.90      0.70      4036
                            Family       0.62      0.29      0.40       572
                           Fantasy       0.62      0.16      0.25       449
                           Foreign       0.12      0.01      0.01       328
                           GoHands       0.00      0.00      0.00         0
                           History       0.36      0.11      0.17       273
                            Horror       0.73      0.58      0.65       972
Mardock Scramble Production Committee    0.00      0.00      0.00         0
...
                         macro avg       0.39      0.28      0.30     18064
                      weighted avg       0.56      0.57      0.54     18064
                       samples avg       0.56      0.64      0.56     18064
```

A interpretabilidade por termos confirma o bom senso semântico: palavras como “magic” e “wizard” puxam **Fantasy**; “murder” e “detective” associam-se a **Crime**; “tour”, “band” e “concert” aparecem em **Music/Documentary**. Para uso prático, o pipeline atual é adequado como **pré-tagueamento**: sugere gêneros prováveis e prioriza a revisão humana onde a incerteza é maior (probabilidades dispersas, baixa confiança).

```
Sinopse: A group of teenagers spend the night in a haunted house where strange events begin to unfold.
Gêneros previstos (prob.):
  - Horror: 0.914
  - Thriller: 0.536

Sinopse: A heartfelt story about a family overcoming challenges during a road trip across the country.
Gêneros previstos (prob.):
  - Drama: 0.758
  - Comedy: 0.494

Sinopse: A young wizard begins his journey at a school of magic
Gêneros previstos (prob.):
  - Family: 0.503
  - Fantasy: 0.474
  - Drama: 0.451
```

```
Top termos — Drama
['drama', 'true story', 'affair', 'prostitute', 'relationship', 'lives', 'love', 'mother', 'friendship', 's

Top termos — Comedy
['comedy', 'hilarious', 'comic', 'comedic', 'bumbling', 'funny', 'comedian', 'spoof', 'wedding', 'satire',

Top termos — Horror
['horror', 'vampire', 'zombie', 'blood', 'zombies', 'evil', 'terrifying', 'vampires', 'supernatural', 'kill

Top termos — Action
['assassin', 'cop', 'action', 'martial', 'fight', 'cia', 'warrior', 'mercenary', 'agent', 'yakuza', 'ruthle

Top termos — Documentary
['documentary', 'interviews', 'filmmaker', 'footage', 'look', 'history', 'portrait', 'film', 'journey', 'dc
```

Do ponto de vista técnico, as limitações decorrem do desequilíbrio natural do catálogo e do fato de a sinopse ser curta e ruidosa. As próximas iterações devem focar em três frentes: ampliar o vocabulário (ou migrar para **embeddings** semânticos), calibrar thresholds por classe e explorar

estratégias de **balanceamento** mais finas. O artefato final está salvo em models/h9_multilabel_pipeline.joblib, pronto para integração por API ou lote.

6. Determinantes de faturamento – H10

Esta seção investiga quais fatores se associam à **receita bruta** dos filmes, oferecendo evidências quantitativas para decisões de investimento e posicionamento. A análise utiliza a versão otimizada do Kaggle, com normalização de tipos financeiros e derivação de variáveis (por exemplo, log_budget, log_revenue, roi, year, release_season, overview_len). Para reduzir viés e multicolinearidade indesejada, controlamos por variáveis de engajamento (**popularidade**, **vote_count**) e por efeitos de idioma e gênero.

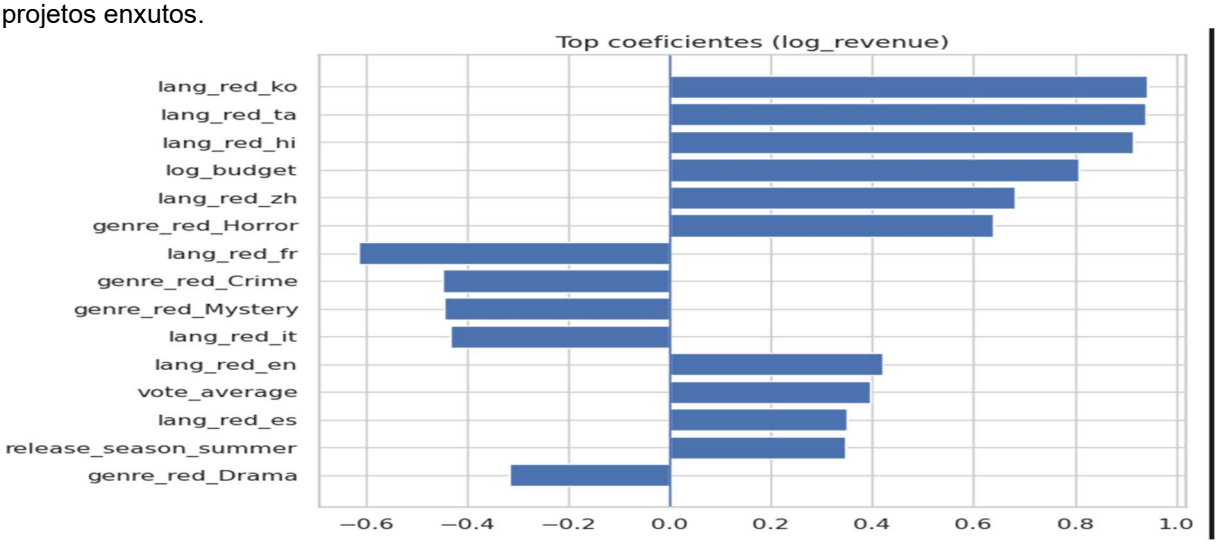
Partimos de um modelo **OLS** com erros-padrão robustos (HC3), onde a variável dependente é log_revenue. A especificação segue a forma:

- *log_revenue = β0 + β1·log_budget + β2·popularity + β3·vote_count + β4·is_english + β5..βk·(controles de gênero/estação/ano) + ε.*

Os resultados indicam que **log_budget é o principal determinante**: o coeficiente estimado é de aproximadamente **0,80 (p < 0,001)**, o que implica, na escala adotada, um aumento de **~72%** na receita esperada para um incremento unitário em log_budget.

	feature	pearson_logR	spearman_logR	pearson_rev	spearman_rev
0	log_budget	0.702083	0.710472	0.398413	0.710472
1	budget	0.507680	0.710472	0.730286	0.710472
4	vote_count	0.452639	0.750512	0.770559	0.750512
7	num_ratings	0.312354	0.583290	0.379984	0.583290
2	popularity	0.271600	0.584587	0.440527	0.584587
3	runtime	0.184250	0.222686	0.189634	0.222686
11	is_english	0.177134	0.208601	0.134349	0.208601
5	vote_average	0.162291	0.127567	0.167775	0.127567
9	n_genres	0.143017	0.176312	0.165220	0.176312
13	release_year	0.105268	0.116752	0.159713	0.116752
12	year	0.105268	0.116752	0.159713	0.116752
6	mean_rating	0.074134	0.054831	0.114825	0.054831
8	n_countries	0.042972	0.064089	0.025112	0.064089
14	release_month	0.042697	0.049890	0.031125	0.049890
10	overview_len	-0.021820	0.044036	0.018840	0.044036

Popularidade e **número de votos** também aparecem positivos e significativos, reforçando o papel do engajamento como **indicador antecedente** de performance. O **idioma inglês** apresenta coeficiente positivo (p < 0,01), consistente com maior alcance internacional. Em controles por gênero, observa-se efeito **positivo para Horror quando condicionado pelo orçamento**, sugerindo eficiência relativa em projetos enxutos.



Modelos de árvore serviram como verificação de robustez. Em **Random Forest** e **HistGradientBoosting**, a **importância de log_budget** se destaca (aprox. **56%** da importância total em uma das parametrizações), com **vote_count** e **num_ratings** complementando o sinal (cerca de **20%** no conjunto). A leitura por **Permutation Importance** reforça essa hierarquia: investimento e engajamento explicam grande parte da variação de receita, enquanto idioma, gêneros e janela modulam a magnitude.

Testes estatísticos descritivos completam o quadro. Uma **ANOVA** simples sugere **Adventure** associado a níveis absolutos de receita mais altos, ao passo que **Documentary, Drama, Comedy e Horror** ficam abaixo em termos absolutos. Um **t-teste** indicou que filmes em inglês, na amostra observada, faturam em média **≈6,6 vezes** mais do que não-ingleses — associação que deve ser interpretada com cautela, pois mistura efeitos de distribuição, marketing e mercado-alvo. Esses resultados não são causais; funcionam como guias para priorização.

	sum_sq	df	F	PR(>F)
C(genre_primary)	47954.66	19.00	68.53	0.00
Residual	1581804.55	42948.00	NaN	NaN

Categorias significativas vs. baseline (p<0.05):

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.4623	0.091	49.236	0.000	4.285	4.640
C(genre_primary)[T.Adventure]	1.0504	0.181	5.815	0.000	0.696	1.405
C(genre_primary)[T.Animation]	-1.5474	0.202	-7.644	0.000	-1.944	-1.151
C(genre_primary)[T.Comedy]	-1.6120	0.111	-14.480	0.000	-1.830	-1.394
C(genre_primary)[T.Crime]	-1.3538	0.173	-7.805	0.000	-1.694	-1.014
C(genre_primary)[T.Documentary]	-3.7781	0.138	-27.398	0.000	-4.048	-3.508
C(genre_primary)[T.Drama]	-1.9898	0.106	-18.721	0.000	-2.198	-1.781
C(genre_primary)[T.Family]	-1.8671	0.280	-6.664	0.000	-2.416	-1.318
C(genre_primary)[T.Fantasy]	-0.0922	0.246	-0.374	0.708	-0.575	0.391
C(genre_primary)[T.Foreign]	-3.7857	0.566	-6.689	0.000	-4.895	-2.676
C(genre_primary)[T.History]	-2.2696	0.374	-6.061	0.000	-3.004	-1.536
C(genre_primary)[T.Horror]	-2.0072	0.149	-13.445	0.000	-2.300	-1.715
C(genre_primary)[T.Music]	-2.8643	0.290	-9.892	0.000	-3.432	-2.297
C(genre_primary)[T.Mystery]	-1.8688	0.273	-6.838	0.000	-2.404	-1.333
C(genre_primary)[T.Romance]	-2.1899	0.198	-11.070	0.000	-2.578	-1.802
C(genre_primary)[T.Science Fiction]	-1.4006	0.255	-5.488	0.000	-1.901	-0.900
C(genre_primary)[T.TV Movie]	-4.4171	0.321	-13.754	0.000	-5.047	-3.788
C(genre_primary)[T.Thriller]	-2.0090	0.174	-11.530	0.000	-2.351	-1.667
C(genre_primary)[T.War]	-2.3973	0.325	-7.385	0.000	-3.034	-1.761
C(genre_primary)[T.Western]	-2.9214	0.300	-9.745	0.000	-3.509	-2.334

A síntese técnica é que **orçamento** e **engajamento** explicam boa parte da **receita bruta**, enquanto **gênero** e **janela** ajustam expectativa e risco. Em termos de portfólio, **Animation/Family/Adventure** sustentam bilheteria e consistência, mas requerem capital elevado e retornos proporcionais moderados; **Horror/Crime/Mystery** tendem a maior **ROI** por custo menor, com variância mais alta de resultados. Para a PProductions, isso recomenda equilibrar **"tratores de bilheteria"** com **apostas eficientes** de médio/baixo orçamento.

7. Previsão da nota do IMDb – H11

Esta seção desenvolve e avalia um modelo de regressão para estimar a nota média do IMDb (*vote_average*, escala 0–10). O objetivo é obter um preditor estável, de baixo custo e plugável no fluxo de análise, capaz de absorver informações numéricas, categóricas e textuais.

Preparação e desenho do experimento:

A variável-alvo é vote_average. O conjunto de preditores combina: (i) numéricos transformados e estabilizados — log_budget, log_revenue, roi, runtime, popularity, vote_count, mean_rating, num_ratings, overview_len, year; (ii) categóricos com redução de cardinalidade por top-K — original_language, release_season, genre_primary, director, star1..star4; (iii) texto — overview, representado por **TF-IDF** (1–2-gram) com vocabulário limitado.

O pré-processamento é unificado num ColumnTransformer com *imputer* mediana para numéricos e OneHotEncoder tolerante a categorias desconhecidas. O particionamento segue **corte temporal por year** (≈70% treino, 15% validação, 15% teste), reduzindo *look-ahead bias*.

Modelos comparados:

Avaliamos uma família linear (Regressão Linear, Ridge, ElasticNet) e duas famílias de árvores (Random Forest, HistGradientBoosting). A motivação teórica é clara: modelos lineares são transparentes, mas tendem a sub-ajustar relações não lineares e interações; modelos de árvores, por sua vez, capturam não linearidades e combinam bem dados heterogêneos com mínimo *feature engineering*.

Resultados:

Os modelos de árvores superaram sistematicamente os lineares. Em validação, o **HistGradientBoosting** apresentou **RMSE ≈ 0,95**, **MAE ≈ 0,65** e **R² ≈ 0,68**; em teste, **RMSE ≈ 1,13**, **MAE ≈ 0,76** e **R² ≈ 0,66**. A **Random Forest** ficou muito próxima (RMSE teste ≈ 1,14; R² ≈ 0,65).

Já os lineares permaneceram na casa de **RMSE ≈ 1,56** em validação, com **R²** bem inferior, e pioraram no *hold-out*. Em termos operacionais, o erro típico do melhor modelo gira em torno de **0,6–0,8 ponto de nota**, o que é adequado para *ranking* e *screening* inicial.

	model	rmse_val	mae_val	r2_val	rmse_test	mae_test	r2_test
0	hgb	0.95	0.65	0.68	1.13	0.76	0.66
1	rf	0.97	0.67	0.66	1.14	0.77	0.65
2	enet	1.56	1.07	0.14	2.06	1.33	-0.14
3	ridge	1.56	1.07	0.14	2.03	1.32	-0.10
4	lin	1.56	1.07	0.13	2.03	1.32	-0.10

A **importância de atributos** (Permutation Importance) confirma que sinais de **engajamento** e **escala** explicam boa parte da variância: vote_count, popularity, mean_rating e num_ratings aparecem entre os principais, seguidos por overview (TF-IDF) e blocos categóricos reduzidos (genre_primary, original_language, director, star1..star4). A presença de year e release_season atua como ajuste de período e janela.

Caso de uso — The Shawshank Redemption.

Aplicamos o pipeline salvo ao exemplo do enunciado, montando a linha de entrada a partir dos campos Lighthouse e projetando-a nas colunas esperadas pelo pré-processamento. O modelo retornou **≈ 6,3** para *Shawshank*, enquanto o valor do dataset é **8,5**. A diferença é compatível com dois efeitos conhecidos: **regressão à média** (o modelo tende a "puxar" extremos para a região central, reduzindo variância) e ausência de alguns **preditores causais** ou *proxies* finos (prestígio de premiações, contagem de telas, campanha de marketing, franquia, boca-a-boca longitudinal).

```python
# Real vs. Previsto - Erro absoluto
IMDB_ID = "tt0111161"
if "imdb_id" in df.columns:
    m = df["imdb_id"].astype(str).eq(IMDB_ID)
else:
    m = df["original_title_norm"].astype(str).str.contains("shawshank", case=False, na=False)

cols = [c for c in ["vote_average","vote_count"] if c in df.columns]
cand = df.loc[m, cols].dropna(subset=["vote_average"])
if len(cand):
    if "vote_count" in cand.columns:
        cand = cand.sort_values("vote_count", ascending=False)
    real = float(cand["vote_average"].iloc[0])
    print(f"Real: {real:.2f} | Previsto: {pred:.2f} | Erro abs.: {abs(real - pred):.2f}")
else:
    print("Nota real não encontrada no dataset.")
```

✓ 0.0s Python

```
Real: 8.50 | Previsto: 6.34 | Erro abs.: 2.16
```

Em cenários práticos, esse enviesamento nos extremos pode ser mitigado com *ensembles* especializados, *winsorization* do alvo ou enriquecimento de variáveis.

O artefato final está salvo em models/h11_imdb_rating_model.pkl como pipeline completo, apto a receber novos registros com o mesmo esquema de colunas e produzir previsões reprodutíveis.

8. Limitações e riscos

Os resultados são **associativos**; não estabelecem causalidade. O **viés de fonte** (Kaggle) implica cobertura incompleta do mercado e políticas de cadastro heterogêneas. Campos financeiros contêm **zeros ou valores faltantes** que podem representar ausência de informação, não montantes reais; apesar das transformações log e do *coerce*, isso limita a precisão de sinais como ROI. Variáveis categóricas de **alta cardinalidade** (diretores, elenco) foram reduzidas por top-K, o que melhora estabilidade, mas perde granularidade; efeitos de indivíduos específicos não são inferência causal. O **texto de sinopse** é curto e ruidoso, e seu poder preditivo depende do vocabulário adotado. Por fim, há **deriva temporal**: preferências mudam, a ascensão do streaming altera distribuição e janelas; métricas treinadas em períodos longos podem requerer recalibração periódica.

9. Reprodutibilidade

O projeto segue um fluxo versionado nos notebooks 00_setup_e_context, 01/02/03 e K01/K02/K03, culminando no 00_final_movie_analytics.ipynb, que consolida cargas, engenharia e inferências de forma executável ponta a ponta.

Os **artefatos** estão em models/ (h9_multilabel_pipeline.joblib, h11_imdb_rating_model.pkl), enquanto tabelas de apoio e figuras são exportadas para reports/ (por exemplo, reports/hypotheses_summary.csv e as imagens de importância/medianas).

O ambiente é definido por requirements.txt. *Seeds* e random_state foram fixados para reprodutibilidade em validação e teste.

A organização de caminhos (PATHS) garante que o notebook funcione tanto na raiz quanto dentro de notebooks/, e o *pipeline* encapsula todo o pré-processamento (imputação, One-Hot, TF-IDF), eliminando risco de *drift* entre treino e produção.

10. Conclusões e próximos passos

Os achados convergem para um quadro coerente. **Orçamento** e **engajamento** explicam grande parte da **receita bruta**; **gênero** e **janela** modulam retorno e risco. Em portfólio, **Animation/Family/Adventure** sustentam bilheteria com consistência, ao custo de investimentos altos; **Horror/Crime/Mystery** tendem a melhor **ROI** por exigirem menos capital, embora com maior variância.

O modelo **H9** demonstra que a **sinopse** contém sinal semântico suficiente para acelerar curadoria com pré-tagueamento; e o modelo **H11** entrega um preditor de nota com erro típico abaixo de um ponto, suficiente para priorizar projetos e calibrar expectativas de recepção crítica.

Em uma próxima iteração, o plano técnico foca em três frentes.

Na **regressão de nota (H11)**: ampliar representação textual (TF-IDF mais profundo ou **embeddings**); criar *features* temporais mais ricas (por exemplo, **decade**, *franchise flag*), refinar top-K de director e stars, e testar **blending** entre modelos de texto-puro e *tabular*, com meta de **reduzir o RMSE em ≥20%** e **elevar o R^2 em +0,03–0,05**.

Na **multilabel de gêneros (H9)**, calibrar *thresholds* por classe, reforçar balanceamento e avaliar embeddings (**BERT**), buscando **+3–5 p.p. em F1-micro** e **+2–3 p.p. em F1-macro**.

Em **faturamento (H10)**, estender variáveis de negócio (indicadores de marketing, telas de estreia, premiações, *franchise/seqüência*), o que tende a melhorar a leitura estrutural do mercado e o poder explicativo dos modelos.

Esses passos incrementais preservam a simplicidade do *stack*, mantêm custos computacionais controlados e avançam precisamente nas alavancas que limitam a performance atual. Com os artefatos salvos e o notebook final reprodutível, a PProductions dispõe de uma base técnica sólida para evoluir análises, experimentar políticas de portfólio e tomar decisões com mais contexto e menos incerteza.