Projeto Final de Ciência De Dados

Danilo Balman Garcia
Ra: 2482088
Rafael Machado Wanner
Ra: 2021013
[1]Departamento Acadêmico de Computação (DACOM)
Universidade Tecnológica Federal do Paraná (UTFPR)

Abstract

This project aimed to simulate a data scientist activity by conducting a complete data analysis, from problem definition to results communication, in order to demonstrate competencies in all stages of the analytical process: data integration, exploratory analysis, predictive modeling, and communication (report writing and presentation). To this end, we chose to conduct an in-depth analysis of the catalog of movies and series available on the main streaming platforms (Netflix, Amazon Prime, and Disney+), with the purpose of discovering patterns in the data and developing predictive models that answer central questions about content strategy and audience reception. Additionally, data from the Internet Movie Database (IMDB) were included. The cross-referencing of these four datasets allowed, based on the definition of certain attributes, to analyze critics' behavior and predict the ratings a given production would achieve.

Keywords: Data Science; Predictive Modeling; Streaming Platforms; IMDB.

Resumo

Este projeto teve como objetivo simular uma atividade como um cientista de dados conduzindo uma análise completa de dados, desde a definição do problema até a comunicação dos resultados, a fim de demonstrar competências em todas as etapas do processo analítico: integração de dados, análise exploratória, modelagem preditiva e comunicação (escrita de relatório e apresentação). Para tanto, optou-se por conduzir uma análise aprofundada do catálogo de filmes e séries disponíveis nas principais plataformas de streaming (Netflix, Amazon Prime e Disney+), com o propósito de descobrir padrões nos dados e desenvolver modelos preditivos que respondam a questões centrais sobre a estratégia de conteúdo e a recepção do público. Adicionalmente, foram incluídos dados do banco de dados da Internet Movie Database (IMDB). O cruzamento destes quatro datasets permitiu, a partir da definição de determinados atributos, analisar o comportamento dos críticos e prever que notas alcançaria determinada produção.

Palavras-Chave: Ciência de Dados; Análise Preditiva; Plataformas de Streaming; IMDB.

1 Introdução

A atividade de um cientista de dados exige a demonstração de competências em todas as etapas do processo analítico, desde a definição do problema até a comunicação efetiva dos resultados. Este projeto visa simular integralmente essa jornada, abrangendo a integração de dados, a análise exploratória e o desenvolvimento de modelos preditivos.

Para tanto, o conjunto de dados selecionado para este projeto é uma compilação de fontes distintas, as quais reúnem informações sobre filmes e programas de TV disponíveis nas plataformas

de streaming Netflix, Amazon Prime e Disney+. Coletivamente, esses conjuntos de dados totalizam aproximadamente 20.000 registros, cada um descrito por 12 atributos iniciais. Os dados iniciais selecionados foram publicados no Kaggle pelo usuário Shivam Bansal, que disponibilizou conjuntos de dados para cada uma das principais plataformas de streaming Bansal (2021a,b,c).

As variáveis abrangem diversas naturezas, incluindo dados textuais, temporais e categóricos. Este dataset foi enriquecido por meio da integração com a base de dados do IMDb, o que permitiu ampliar o potencial analítico, com a adição de mais três atributos.

Desse modo, os datasets em conjunto contêm os seguintes atributos:

Atributo	**Descrição**
show_id (textual)	Identificador único do filme/série.
type (textual)	Indica se a obra é um filme ou uma série.
title (textual)	Título da obra.
director (textual, topológico)	Diretores da obra.
cast (textual, topológico)	Atores que compõem o elenco.
country (textual, espacial)	País em que a obra foi produzida.
date_added (data, temporal)	Data em que a obra foi adicionada na plataforma.
release_year (inteiro, temporal)	Ano em que a obra foi lançada.
rating (textual)	Classificação indicativa.
duration (textual)	Duração da obra (min. para filmes, temporadas para séries).
listed_in (textual)	Gêneros em que a obra é listada.
description (textual)	Descrição breve da obra.
streaming (textual)	Plataforma de origem.
imdb_rating (inteiro)	Nota do IMDB.
run_time_minutes (temporal)	Tempo em Minutos.
number_votes (inteiro)	Número de votos no IMDB.

O código-fonte completo da implementação, incluindo todos os scripts utilizados para a análise de dados, foi disponibilizado publicamente. O projeto pode ser acompanhado através do seu repositório oficial no GitHub:

nosso repositório online

2 Definição do problema e perguntas de pesquisa

O cenário atual do entretenimento é marcado pela saturação do mercado de streaming. Plataformas gigantes como Netflix, Amazon Prime Video e Disney+ competem ferozmente, investindo somas colossais na expansão de seus catálogos. Nesse contexto altamente competitivo, a otimização do conteúdo e a retenção de assinantes aparecem como desfaios críticos e permanentes.

Uma das principais dificuldades é justamente obter uma previsão confiável da recepção do público antes que uma produção seja produzida, lançada ou adquirida. A quantidade de projetos, filmes e séries disponíveis, aliada aos critérios de avaliação interna da plataforma e as métricas externas (como as do IMDB), certamente cria uma zona de incerteza para a tomada de decisões estratégicas.

Assim, o problema que este estudo se propõe a resolver é o desenvolvimento de um mecanismo preditivo que utilize a Ciência de Dados para cruzar e analisar o vasto conjunto de atributos dos projetos e produções. O objetivo é superar a dificuldade em traduzir a complexidade de dados em algo prático: prever a classificação que um determinado título irá alcançar. Essa capacidade preditiva é crucial para minimizar o risco de investimento e alinhar a estratégia de conteúdo com o desejo do público.

Assim, este projeto conduziu uma análise aprofundada do catálogo de filmes e séries disponíveis nas principais plataformas de streaming (Netflix, Amazon Prime e Disney+) com o propósito de descobrir padrões nos dados e desenvolver modelos preditivos que respondam a questões sobre a estratégia de seleção de conteúdo baseado na predição da recepção do público. Para guiar esta investigação, foi formulada a seguinte questão de pesquisa: É possível prever a classificação que o assinante dará a determinada produção? O projeto pretende textar as seguintes hipóteses:

- *H1: Filmes de terror costumam ter uma duração menor*
- *H2: Filmes no Disney Plus chegam mais rápido ao streaming*
- *H3: Atores dedicados ao gênero de terror participam de menos produções*

3 Metodologia e limitações

Na Etapa 2, desenvolvemos inicialmente um *script* para realizar o *download* dos arquivos do *dataset* principal e dos dados auxiliares (IMDb). O primeiro passo do processamento consistiu na concatenação das bases de dados de streaming (Netflix, Disney+ e Prime Video) em um único *dataframe*.

Foi necessária, nessa etapa, a definição de cada plataforma de streaming como um atributo, bem como a aplicação a transformação *wide*, que subdivide os atributos e atribui-lhes o valor 1 para positivo e zero para negativo. Com os dados unificados, aplicamos uma rotina de normalização nos gêneros. Este processo incluiu:

- A separação de gêneros compostos (ex: ”Animals & Nature”);
- A padronização de termos (convertendo ”TV Dramas”para ”Drama”);
- A remoção de subgêneros, sinônimos e formatos que não representam temas (como ”TV Shows”).

O link do *Github* acima inclui o código utilizado para o mapeamente.

Após tratarmos os gêneros, buscamos complementar o *dataset* com informações do IMDb, adicionando métricas como a nota média e a duração em minutos. Para superar inconsistências nos nomes das obras, adotamos uma estratégia de *fuzzy matching* na etapa de merge, seguindo uma ordem de prioridade hierárquica::

Prioridade 1: Correspondência de Nome (similaridade $\geq 80\%$)E mesmo Diretor;

Prioridade 2: Caso não houvesse correspondência na etapa anterior, verificava-se Nome (similaridade $\geq 80\%$)E mesmo Ano de Lançamento.

Essa estratégia resultou em uma taxa de sucesso de 72%, consolidando aproximadamente 14 mil títulos no dataset final. Por fim, foi realizada uma limpeza nos dados de duração e votação para o tratamento de valores nulos.

A etapa 3 consistiu em converter o conjunto de dados para o formato tidy data, para que cada variável tivesse sua própria coluna e cada célula contivesse um único valor, pois poucos modelos de *machine learnig* conseguem tratar atributos multivalarados. Assim, para gênero e país de origem são optou-se também por aplicar a transformação *wide*.

Para os atributos de ator e diretor, considerando-se que as possibilidades de variáveis são potencialmente gigantescas, optou-se por processar os dados calculando-se o percentual de participações de cada um pelo gênero das produções, com atributos como, por exemplo:

"Dir_Ratio_Genre_Adventure" ou "Cast_Ratio_Genre_Action".

Para o atributo "Nota", considerando que o modelo de *machine learnig* não conseguiria prever de modo aceitável para as pretensões deste trabalho se os valores fossem subdividos de 0,1 a 10,0, optou-se por agrupar os valores nas seguintes cinco categorias distintivas:

- De 0,1 a 2 - Classificação 0: (Muito Ruim)',
- De 2,1 a 4 - Classificação 1: (Ruim)',
- De 4,1 a 6 - Classificação 2: (Médio)',
- De 6,1 a 8 - Classificação 3: (Bom)',
- De 8,1 a 10 - Classificação 4: (Excelente).

Como o modelo utilizado também não é capaz de processar textos diretamente, datas que contenham barras entre os números são impeditivas. Assim, foi necessário transformar as datas em *Timestamp* (segundos) no arquivo. Pela mesma razão (incapacidade do Modelo de processar textos), os atributos de classificação por faixa etária receberam categorias distintivas de 1 (livre) a 6 (Maiores de dezoito).

Na Etapa 4 aplicamos dois algoritmos da biblioteca *scikit-learn*: Regressão Linear e SVM (Support Vector Machine).

Para o treinamento dos modelos, selecionamos as seguintes variáveis (*features*), abrangendo dados temporais, técnicos, de elenco e de distribuição:

```
Python
colunas_features = ['release_year', 'run_time_minutes', 'number_votes'] +
[c for c in df.columns if c.startswith('Genre_')] +
[c for c in df.columns if c.startswith('Type_')] +
[c for c in df.columns if c.startswith('Cast_')] +
[c for c in df.columns if c.startswith('Dir_')] +
[c for c in df.columns if c.startswith('Streaming_')]
```

Utilizamos uma divisão dos dados de 80% para treinamento e 20% para teste.

Ao analisarmos as matrizes de confusão resultantes, identificamos um problema de desbalanceamento de classes no *dataset*. A grande maioria dos títulos concentra-se na faixa de notas entre 4.1 e 8.0. Por conta disso, os algoritmos conseguem predizer com maior precisão os filmes situados nessa faixa média, mas apresentam dificuldade de generalização para as notas muito baixas ou muito altas.

Uma possível solução para este esse problema seria a geração de dados sintéticos para simular filmes de notas baixas, pois essa técnica, dependendo do *dataset*, não costuma introduzir ruídos significativos. Além disso, discutiu-se com o professor/orientador que a aplicação de validação cruzada, com a repetição dos testes, poderia aprimorar a robustez da predição. No entanto, devido ao alto custo computacional, de aproximadamente 30 minutos por execução, optou-se por manter a validação simples.

Abaixo, apresentamos os resultados finais de classificação:

Text
Relatório Detalhado:

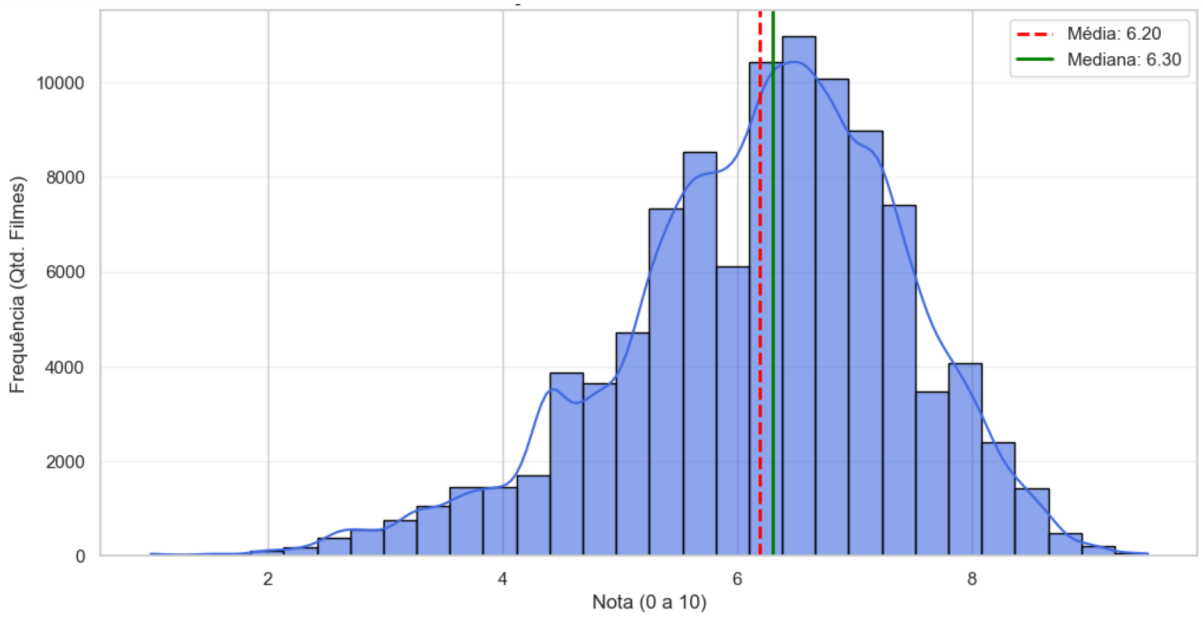
	precision	recall	f1-score	support
0-2 (Muito Ruim)	0.00	0.00	0.00	12
2.1-4 (Ruim)	0.62	0.03	0.06	532
4.1-6 (Médio)	0.68	0.64	0.66	3608
6.1-8 (Bom)	0.73	0.87	0.80	5587
8.1-10 (Excelente)	0.94	0.18	0.31	455
accuracy			0.71	10194
macro avg	0.59	0.35	0.36	10194
weighted avg	0.72	0.7	0.69	10194

4 Resultados das análises

4.1 Análise Univariada

A análise inicial buscou entender o comportamento geral do catálogo. Observando o histograma gerado (Figura 1), nota-se que a distribuição das notas do IMDb se assemelha a uma "montanha" centrada na faixa média (entre 4 e 8). A curva de densidade confirma que avaliações extremas, sejam elas de obras muito boas ou muito ruins, são raras estatisticamente.

Figura 1: Distribuição Univariada das Notas IMDB

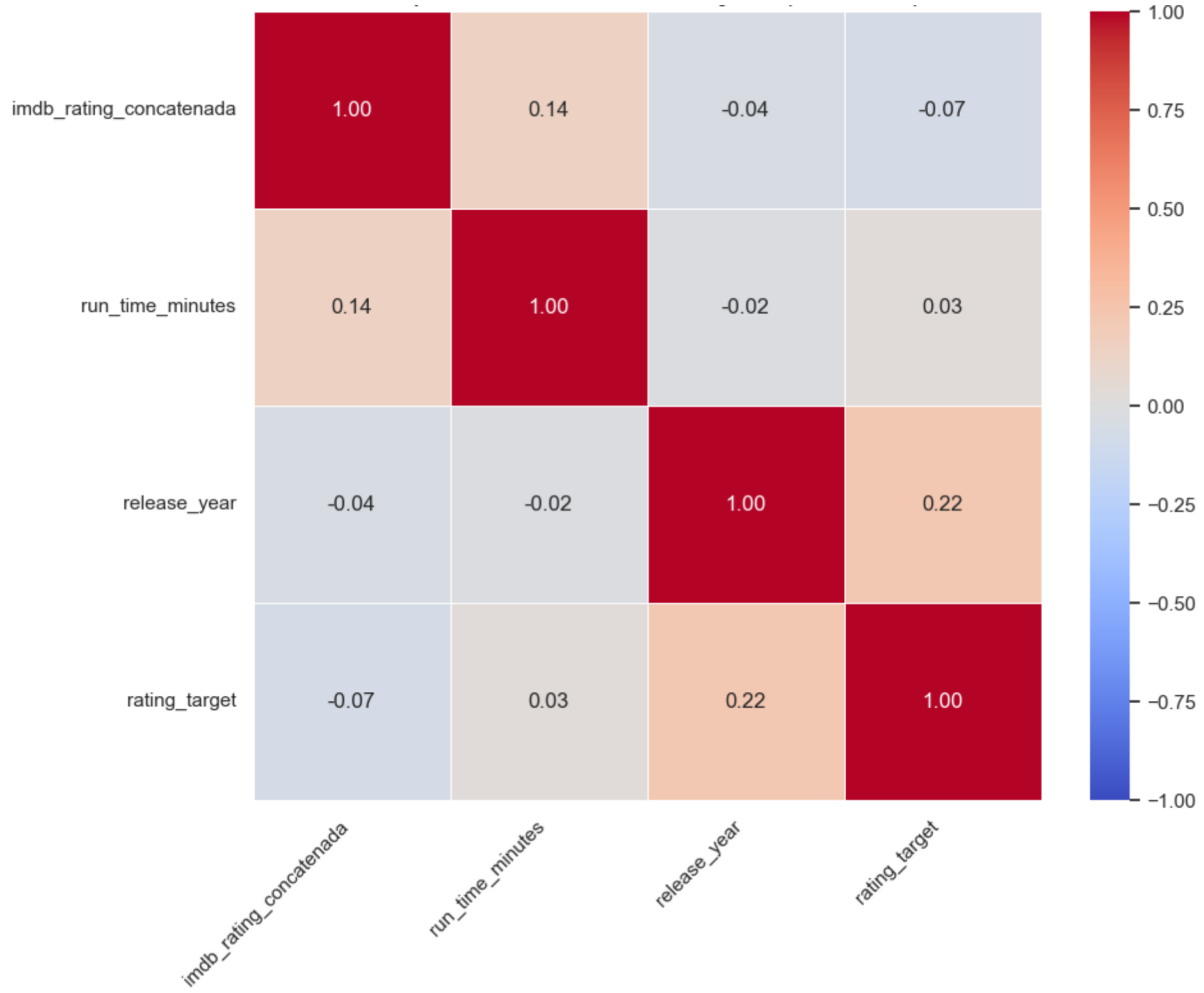


Elaboração própria

4.2 Análise Bivariada

Ao cruzarmos as variáveis para entender o que define o sucesso de uma obra, utilizamos um mapa de calor (*heatmap*), onde cores quentes indicariam fortes conexões (Figura 2). Visualmente, a predominância de tons neutros entre "Duração" e "Nota" revela que filmes mais longos não são, necessariamente, percebidos como melhores pelo público. Contudo, ao segmentar os dados por faixa etária, notou-se um padrão: produções classificadas para adultos tendem a possuir uma média de avaliação muito aproximada às produções de classificação livre, sugerindo níveis de exigência semelhantes entre esses públicos.

Figura 2: Mapa de calor das correlações

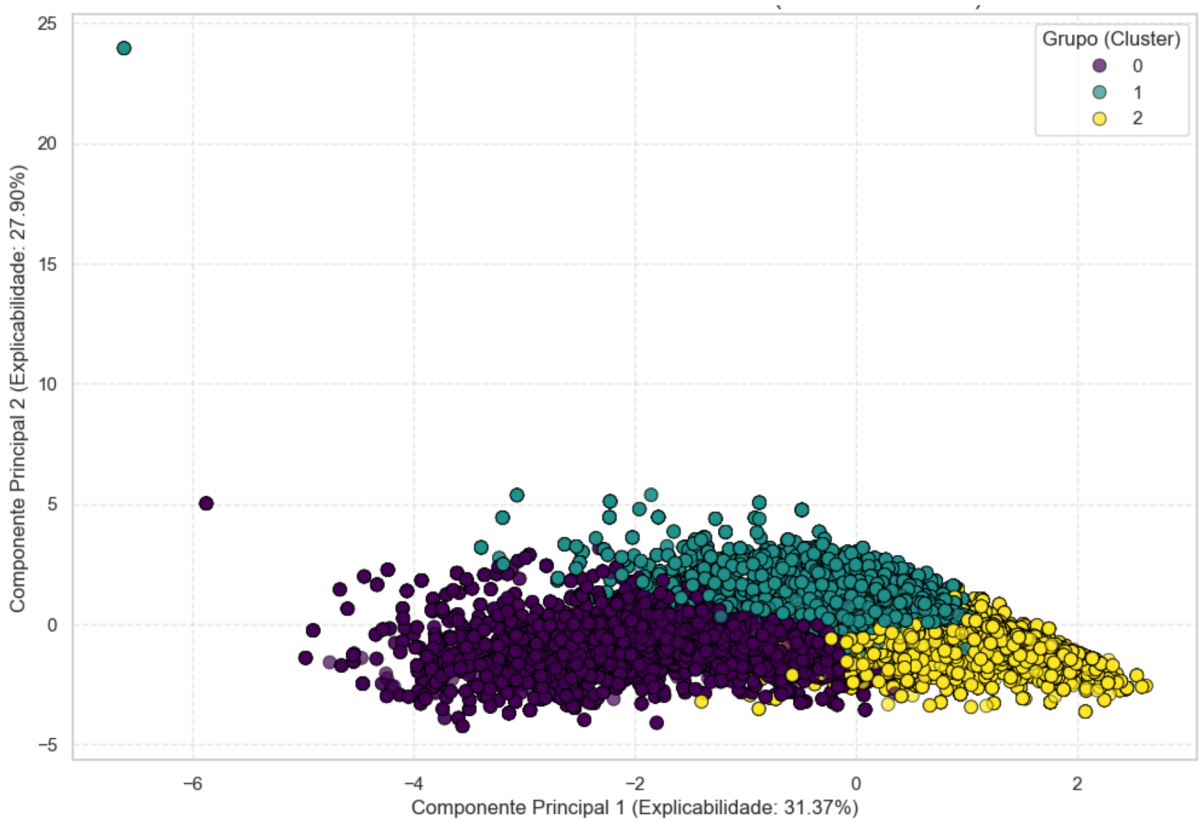


Elaboração própria

4.3 Análise Multivariada

Para detectar nichos que escapam à observação humana, aplicou-se o algoritmo *K-Means*, que segmentou o catálogo em 3 perfis automáticos. A projeção gráfica (Figura 3) mostra esses grupos como "nuvens" de cores distintas. A separação visual clara entre os grupos indica que o catálogo é heterogêneo, composto por bolhas de conteúdo com características técnicas próprias (como a distinção entre filmes clássicos antigos e lançamentos rápidos de *streaming*).

Figura 3: Análise Multivariada - Cluster de Filmes



Elaboração própria

4.4 Testes de Hipóteses

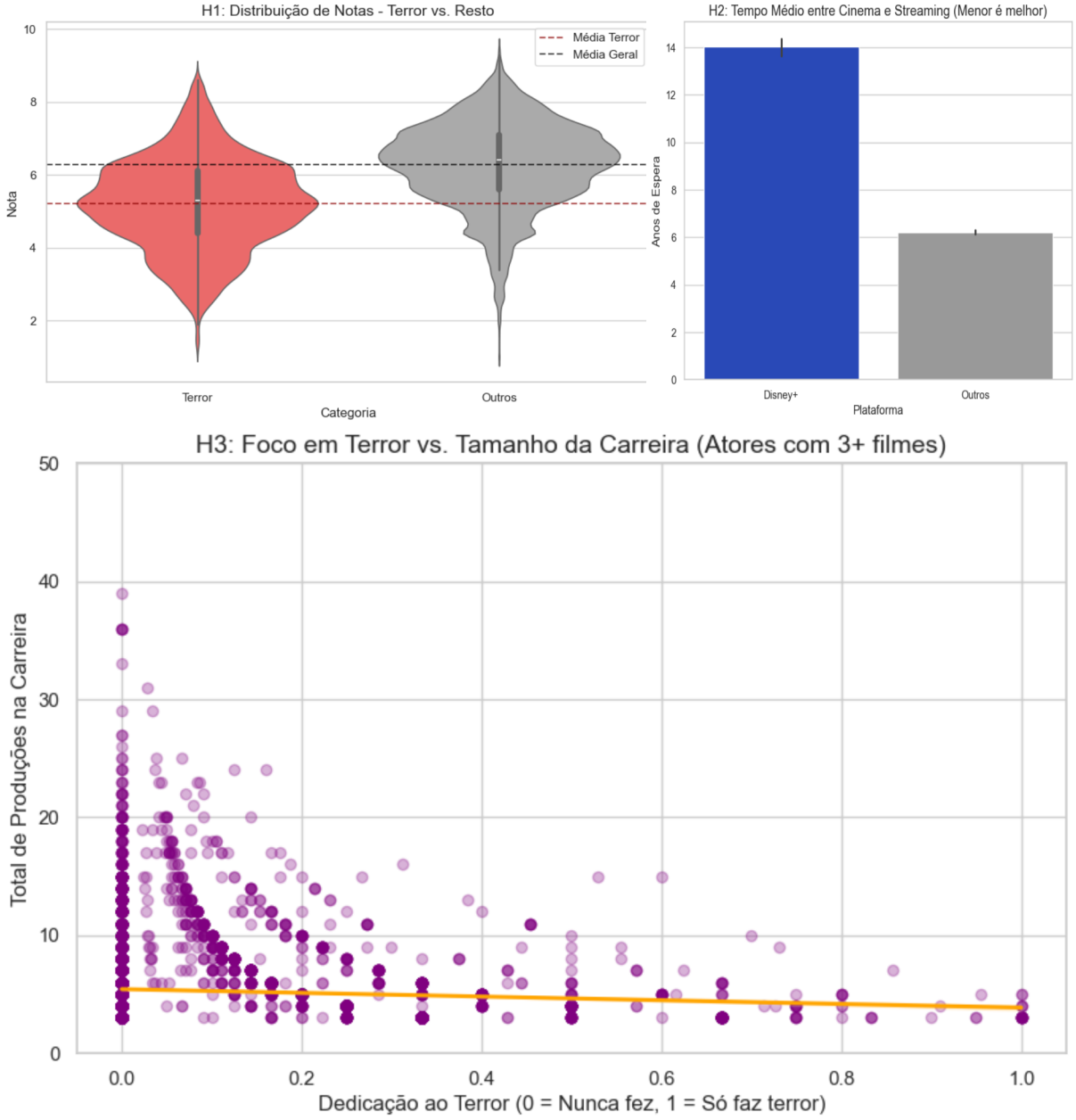
Por fim, submetemos as suposições estratégicas a testes estatísticos, ilustrados na Figura 4:

O estigma do Terror (H1): O gráfico de violino demonstrou uma concentração de notas mais baixas para o gênero de Terror em comparação aos demais, confirmando estatisticamente que este gênero enfrenta uma resistência crítica histórica.

A agilidade do Disney+ (H2): A comparação visual das barras de atraso (delay) comprovou que o Disney+ possui a menor janela de tempo entre o lançamento da obra e sua disponibilização na plataforma, validando sua estratégia de exclusividade e rapidez frente aos concorrentes.

Carreira no Nicho (H3): A linha de tendência na análise dos atores mostrou uma inclinação nregativa. Isso indica que, ao contrário do senso comum, a especialização em filmes de terror prejudica o volume total de trabalho do artista, provando ser um nicho sustentável para a carreira.

Figura 4: Testes de Hipóteses



Elaboração própria

5 Discussão dos resultados

A análise crítica dos dados revela que o desempenho do nosso modelo preditivo — com uma taxa de acerto global de 71% — é um espelho bastante fiel da própria indústria cinematográfica: um mercado saturado de obras medianas, onde grandes sucessos ou fracassos retumbantes são exceções nas estatísticas. Observou-se que o modelo utilizado, ao aprender com esse histórico tende a classificar a maioria das obras na zona de conforto das notas médias (entre 4 e 8).

Embora isso garanta consistência, cria uma dificuldade para as produções ou projetos muito bons ou muito ruins, dificultando a identificação antecipada de possíveis sucessos ou desastres de crítica, justamente porque a base de dados oferece poucos exemplos desses casos para aprendizado. Além disso, a validação das hipóteses apontou uma tendência cultural importante, com o gênero de terror recebendo notas geralmente inferiores, sugerindo que a avaliação do público carrega preconceitos que podem ir além da qualidade técnica da obra.

6 Recomendações práticas

Diante desse cenário, propomos diretrizes estratégicas para transformar esses *insights* em vantagem competitiva. Para a gestão de conteúdo, os dados indicam que o investimento em narrativas maduras (classificação adulta) traz retornos de crítica superiores, desmistificando o receio de segmentar o público. A estratégia de lançamento rápido do Disney+ também se provou uma estratégia eficaz, sugerindo que reduzir a janela temporal entre cinema e *streaming* potencializa o engajamento. No campo técnico, a recomendação primordial para a equipe de dados é focar no balanceamento artificial do *dataset*. É necessário testar mais o modelo e estudar mais profundamente os casos raros de excelência e fracasso, para que ele evolua de um classificador mediano para uma ferramenta capaz de apontar sucessos com precisão.

7 Trabalhos futuros

Para possíveis próximas etapas, o foco ideal seria tornar o modelo mais "inteligente" na interpretação do conteúdo. Nas configurações utilizadas, o sistema analisa apenas categorias fixas,como Ação ou Drama, mas se fosse possível, por exemplo, incluir análise de texto e processar o conteudo do resumo das obras, ele passaria a entender o contexto e o tom da história, detalhes que podem ser decisivos para o público, e ampliaria os percentuais de acerto.

Além disso, seria válido migrar para algoritmos mais robustos, capazes de detectar conexões complexas sozinhos. Diferente do método atual, essas novas ferramentas conseguiriam descobrir automaticamente o peso real da fama de um diretor ou a química de um elenco, sem que fosse necessário criar regras manuais para isso. Por fim, um bom aprimoramento seria equilibrar os dados para corrigir a aparente timidez do modelo, fazendo o sistema estudar mais a fundo os casos de sucesso e de fracasso absolutos, para que aprender a identificar obras diferenciadas com a mesma segurança que é capaz de classificar um filme mediano.

Referências

Bansal, Shivam. 2021a. Amazon Prime Movies and TV Shows. Acessado em: 14 de outubro de 2025. https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows.

Bansal, Shivam. 2021b. Disney+ Movies and TV Shows. Acessado em: 14 de outubro de 2025. https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows.

Bansal, Shivam. 2021c. Netflix Movies and TV Shows. Acessado em: 14 de outubro de 2025. https://www.kaggle.com/datasets/shivamb/netflix-shows.