**É possível prever o sucesso de um filme por meio do uso de inteligência artificial?**

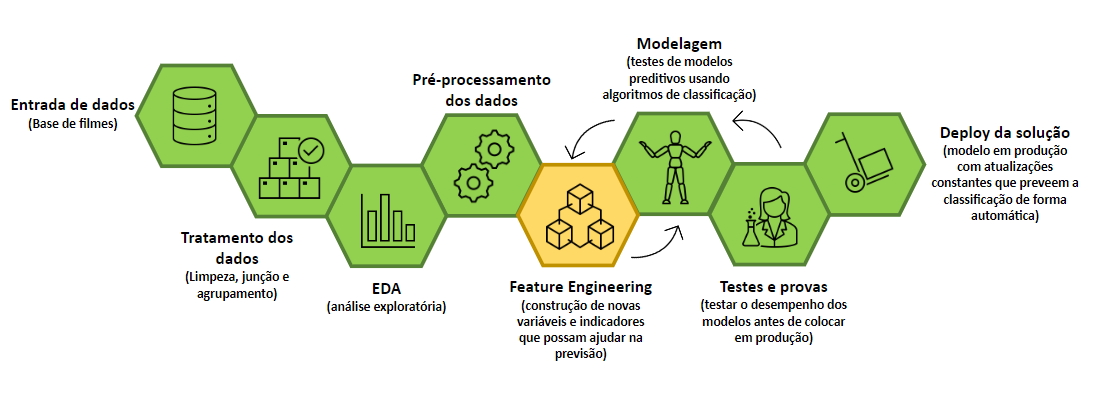
**Introdução**

O cenário do cinema contemporâneo é marcado por um ambiente altamente competitivo. Com o crescimento exponencial do *streaming* e a multiplicação de cinemas, as empresas precisam de uma visão aprofundada para satisfazer as necessidades de cada espectador e oferecer os filmes ideais para cada situação.

Nesse contexto desafiador, um projeto foi desenvolvido por um grupo de alunos da Tera com o objetivo de criar um sistema de classificação que oriente produtores e distribuidores de filmes. **Busca-se com isso determinar se um filme será bem-sucedido, com base em variáveis como gênero, roteiro e diretor. Além disso, espera-se prever a receita da semana de lançamento do filme.**

Essa questão é fundamental para a indústria cinematográfica, visto que o lançamento de um filme envolve altos investimentos e riscos financeiros significativos. Com um modelo de inteligência artificial de classificação e previsão seria possível ajudar os produtores e distribuidores a tomar decisões mais informadas sobre o lançamento e a divulgação de filmes, aumentando as chances de sucesso financeiro e de crítica.

A seguir, este artigo irá detalhar as etapas solução proposta passando pela coleta e análise de dados, modelagem da solução e o deploy, conforme a figura abaixo.

****

**Figura 1: Desenho da solução. Fonte: Autoria própria.**

**Capítulo 1: Obtendo os Dados**

Nossa primeira etapa foi reunir dados sobre o maior número de filmes possível, com informações detalhadas. Exploramos várias fontes, incluindo APIs como a da IMDB, que fornecem informações sobre filmes. No entanto, enfrentamos obstáculos, como a dificuldade e velocidade para conseguir os dados.

Fomos em busca de conjuntos de dados já existentes que pudessem fornecer uma base sólida de informações sobre filmes. No entanto, encontramos limitações em muitos conjuntos de dados do Kaggle. Muitas entradas não atendiam aos nossos critérios mínimos, como ter receita e orçamento.

Após várias tentativas de obter dados adicionais de outras fontes, decidimos concentrar nossos esforços em dois conjuntos de dados do Kaggle. Combinamos esses conjuntos e mais um genérico da IMDB, eliminando recursos específicos de cada um e ajustando nomes de colunas. Essa fusão nos deu um conjunto não muito grande, mas com todas as informações que precisaríamos, com informações sobre uns 600 filmes que abrangem o período de 2000 a 2017. A partir deste ponto, nossa ênfase mudou para melhorar a qualidade dos recursos, em vez de simplesmente aumentar a quantidade de dados. A tabela a seguir detalha as variáveis da base de dados final.

| **Variável** | **Tipo** | **Dicionário** |
| --- | --- | --- |
| Film | string | Nome do filme |
| Rotten Tomatoes critics | numérico | Nota do Rotten Tomatoes |
| Metacritic critics | numérico | Nota do Meta Critics |
| Average critics | numérico | Média da nota do Rotten e Meta |
| Rotten Tomatoes Audience | numérico | Audiência que deu nota no RT |
| Metacritic Audience | numérico | Audiência que deu nota no MC |
| Rotten Tomatoes vs Metacritic deviance | numérico | Desvio entre a nota do RT e MC |
| Average audience | numérico | Audiência média entre RT e MC |
| Audience vs Critics deviance | numérico | Desvio entre a audiência do RT e MC |
| Primary Genre | string | Gênero principal do filme |
| Genres | string | Todos os gêneros do filmes |
| Script Type | string | Tipo de roteiro |
| Opening weekend ($million) | numérico | Valor arrecadado na semana de abertura |
| Opening Weekend | numérico | Valor arrecadado na semana de abertura |
| Domestic gross ($million) | numérico | Arrecadação nos EUA |
| Domestic Gross | numérico | Arrecadação nos EUA |
| Foreign Gross ($million) | numérico | Arrecadação fora dos EUA |
| Foreign Gross | numérico | Arrecadação fora dos EUA |
| Worldwide Gross | numérico | Arrecadação mundial |
| Worldwide Gross ($million) | numérico | Arrecadação mundial |
| of Gross earned abroad | numérico | % da arrecadação fora dos EUA |
| Budget ($million) | numérico | Orçamento |
| Budget recovered | numérico | Orçamento recuperado |
| Budget recovered opening weekend | numérico | Orçamento recuperado na semana de abertura |
| Year | numérico | Ano de lançamento do filme |
| Oscar Winners | dummy | Vencedor de Oscar |
| Oscar Detail | string | Tipo do Oscar |
| Link | string | Link da sinopse do filme |

**Capítulo 2: Preparando os Dados**

Em nossa jornada rumo à previsão do sucesso dos filmes, mergulhamos fundo no processo de preparação dos dados e modelagem. Esta etapa crítica envolveu uma série de desafios e decisões inteligentes que moldaram nosso projeto.

A engenharia de recursos foi uma das partes mais desafiadoras do nosso projeto, principalmente devido à natureza complexa dos filmes e das informações disponíveis. Nosso objetivo era criar um conjunto de recursos que abrangesse anos, meses, diretores, atores, gêneros, classificações, avaliações de filmes, orçamento, produção e receita, entre outros. No entanto, as bases de dados que encontramos continham apenas alguns desses recursos. A maior parte das informações estava formatada de maneira incompatível com o que precisávamos.

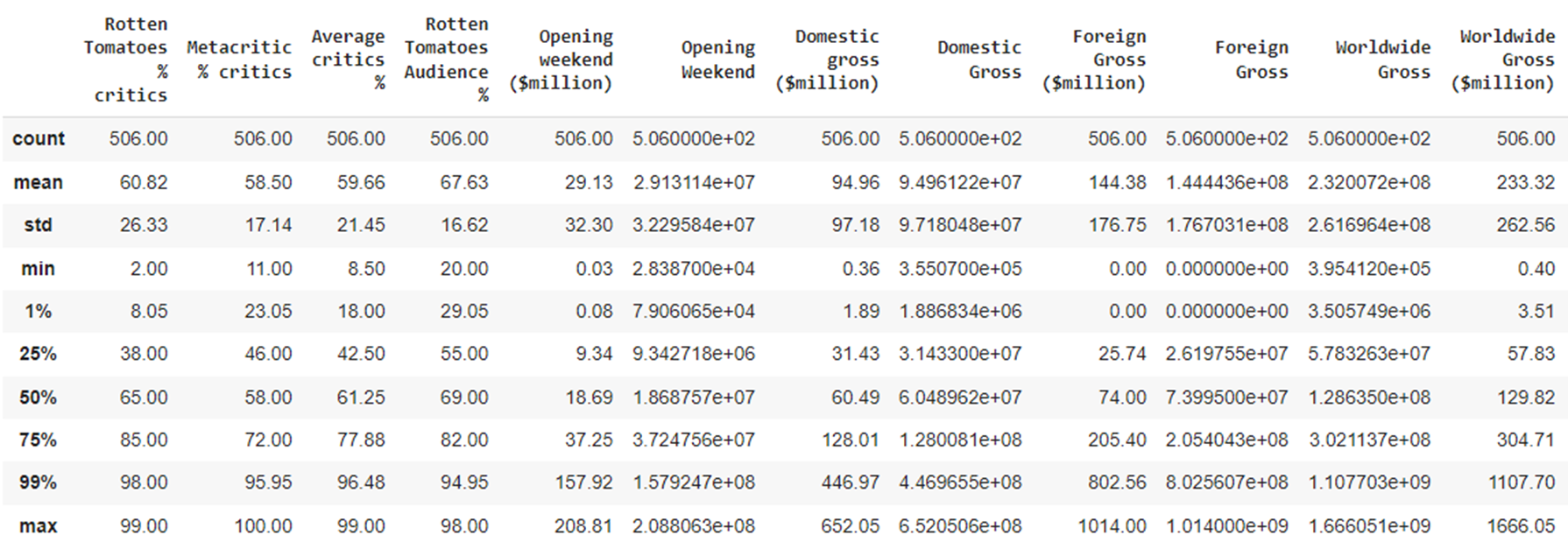
Por exemplo, as informações sobre gênero eram armazenadas como objetos de dicionário ou listas de texto com diferentes formatações. Para resolver isso, desenvolvemos um analisador capaz de extrair informações sobre os gêneros de cada filme. Optamos por representar os gêneros usando a técnica "one hot encoding", criando uma coluna para cada tipo de gênero, marcando "1" se o filme pertencesse a esse gênero e "0" caso contrário. Essa abordagem nos permitiu ter uma representação clara dos gêneros de cada filme.

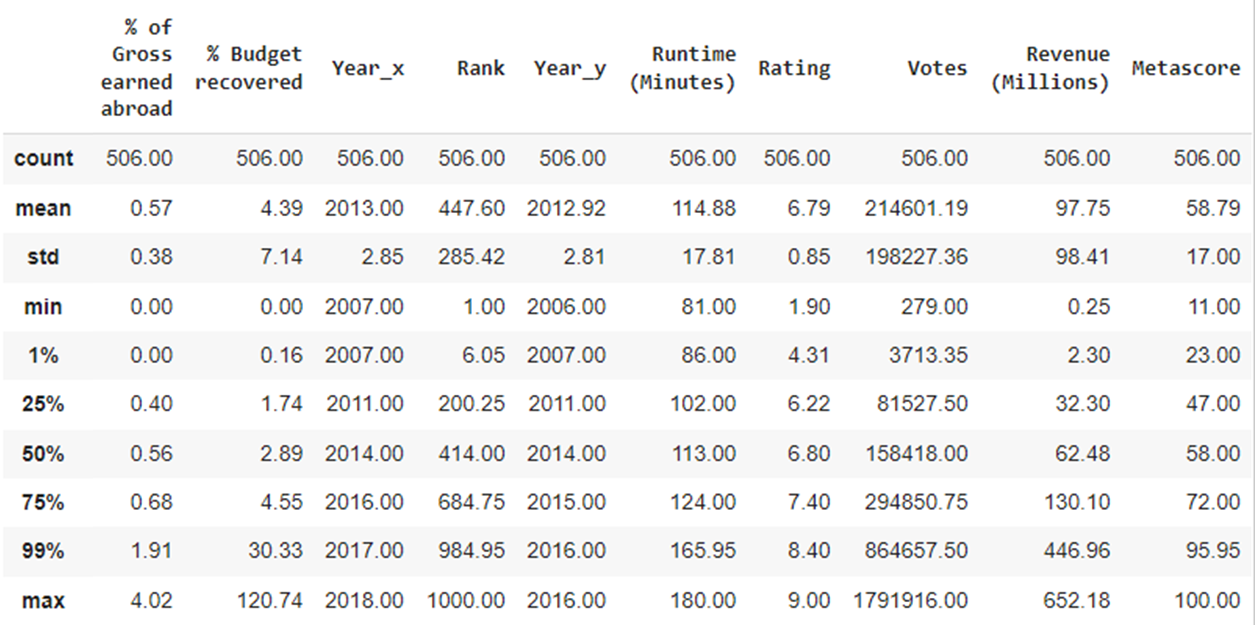
Além disso, os recursos categóricos que continham nomes de diretores foram transformados em recursos quantitativos para serem utilizados em nossos modelos. Além de criarmos uma coluna com a quantidade de prêmios do Oscar por diretor.

No geral, nossa estratégia de engenharia de recursos visava aprimorar os dados existentes e enriquecer nosso conjunto de dados com informações adicionais. Com cerca de 180 dimensões no total, nosso conjunto de dados se tornou mais robusto e completo em comparação com os estudos anteriores.

**Capítulo 3: Análise Exploratória**

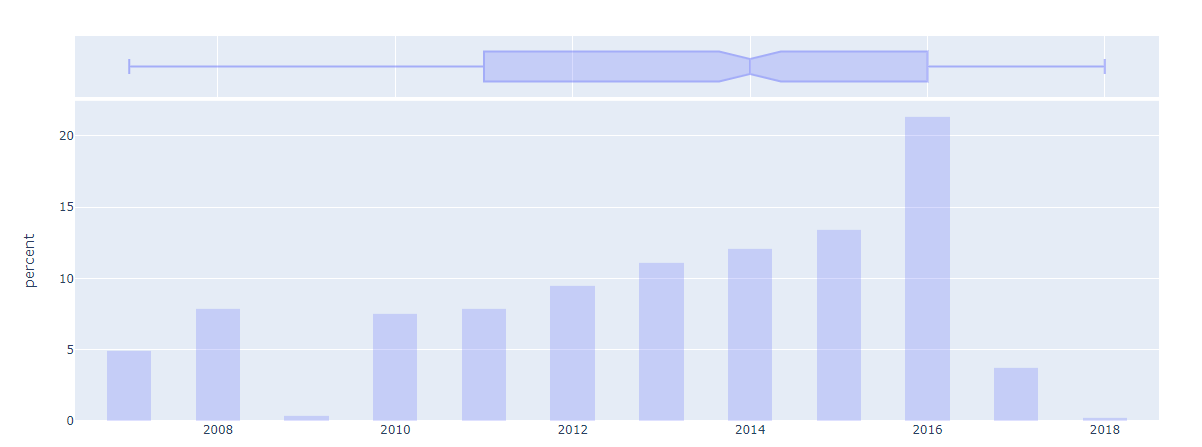
De posse da base de dados preparada, o passo seguinte foi avaliar o comportamento das variáveis que a compõem. Apresentaremos a seguir algumas conclusões resultantes dessa análise exploratória sendo esta não-exaustiva. O notebook contendo a EDA completa pode ser encontrado no Git Hub desse projeto (link ao final do artigo).

****

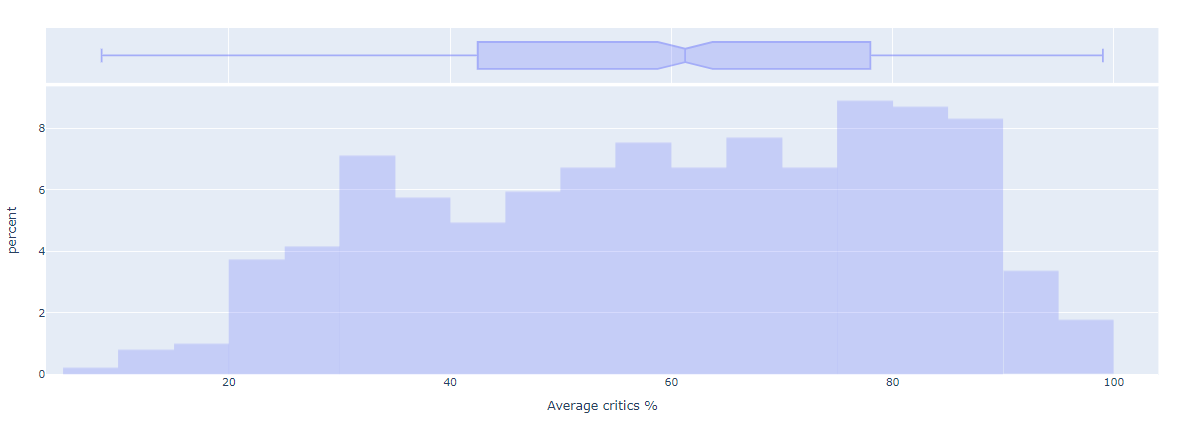
****

**Algumas conclusões EDA univariada - quantitativa (os demais gráficos são encontrados no github):**

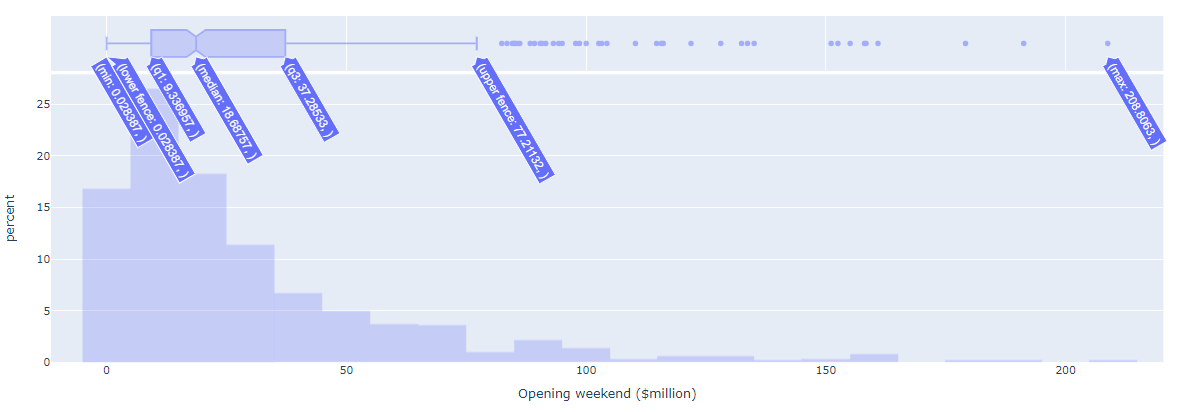
1. 2016 concentra mais filmes do que os outros anos



1. No geral, a mediana das críticas gira em torno de 61



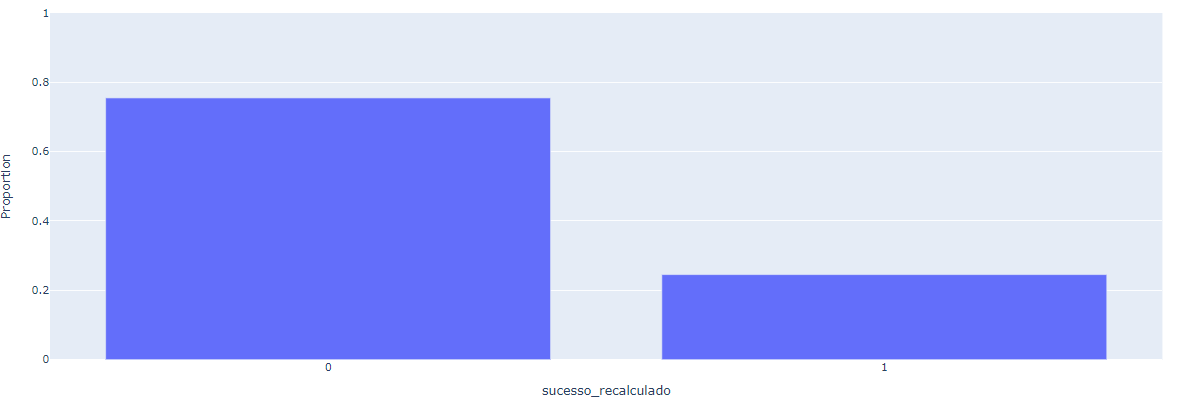
1. No geral, a mediana da arrecadação da semana de abertura é 18 milhões, a arrecadação doméstica gira em torno de 60 milhões, a arrecadação estrangeira em torno de 73 milhões e a arrecadação mundial em torno de 130 milhões;



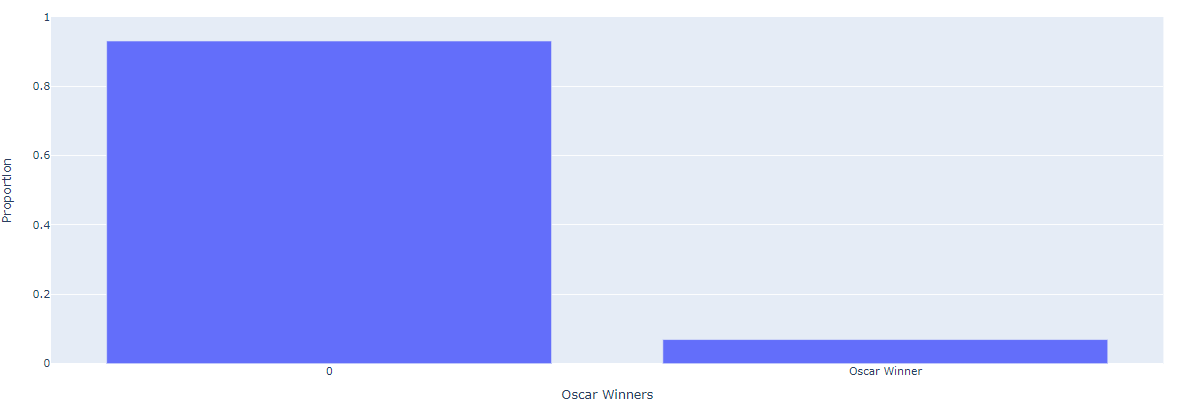
1. Quase 60% da arrecadação é fora dos EUA;
2. O budget mediano gira em torno de 47 milhões;
3. A mediana do budget recuperado gira em torno de 3% (0.4% na semana de abertura);
4. O tempo de duração mediano é 113 minutos;
5. A mediana do metascore é 58
6. boa parte dos filmes concentra uma receita entre 25 e 75 milhões de dólares
7. O rating mediano é de 6.8

**Algumas conclusões sobre a EDA univariada - qualitativa**

1. 25% dos filmes foram sucesso

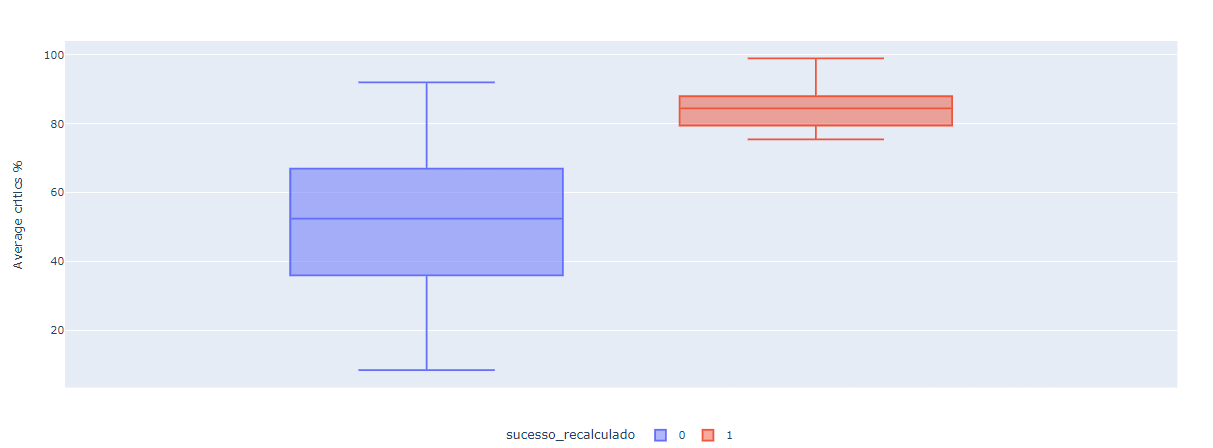


1. A maioria dos filmes é de ação, comédia e drama;
2. A maioria é adaptação ou roteiro original;
3. Apenas 7% ganharam algum Oscar

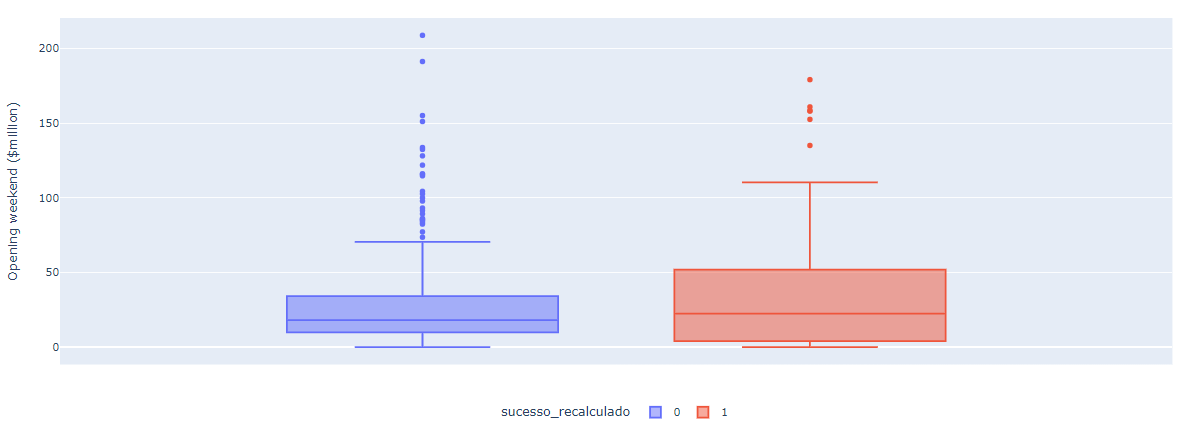


**Algumas conclusões sobre a EDA bivariada - quantitativa**

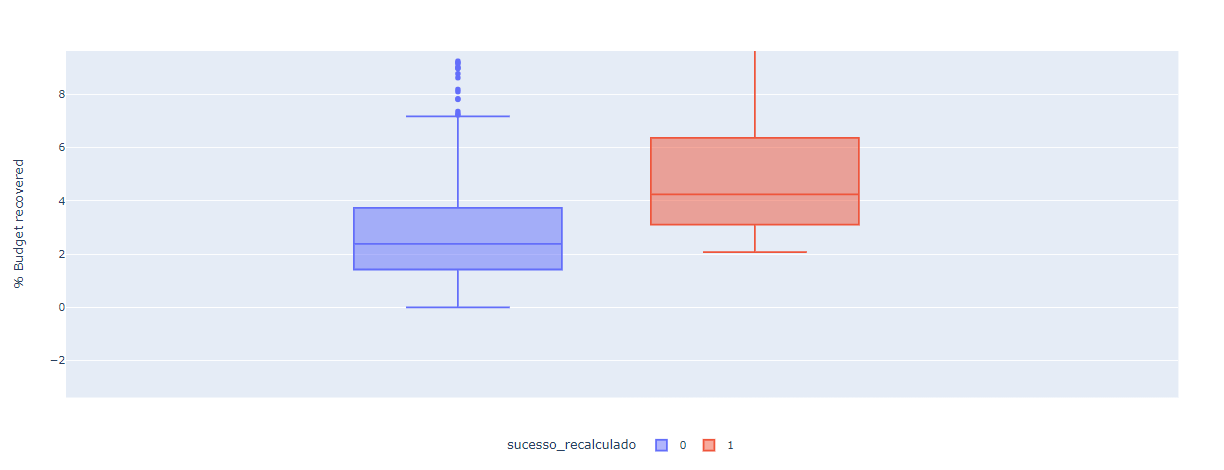
1. Os filmes que são um sucesso tem uma crítica mediana de 84 pontos, enquanto os filmes que não são sucesso têm a mediana em 52.



1. A arrecadação mediana na primeira semana dos filmes que são um sucesso (22 milhões) não difere muito dos filmes que não são sucesso (18 milhões);



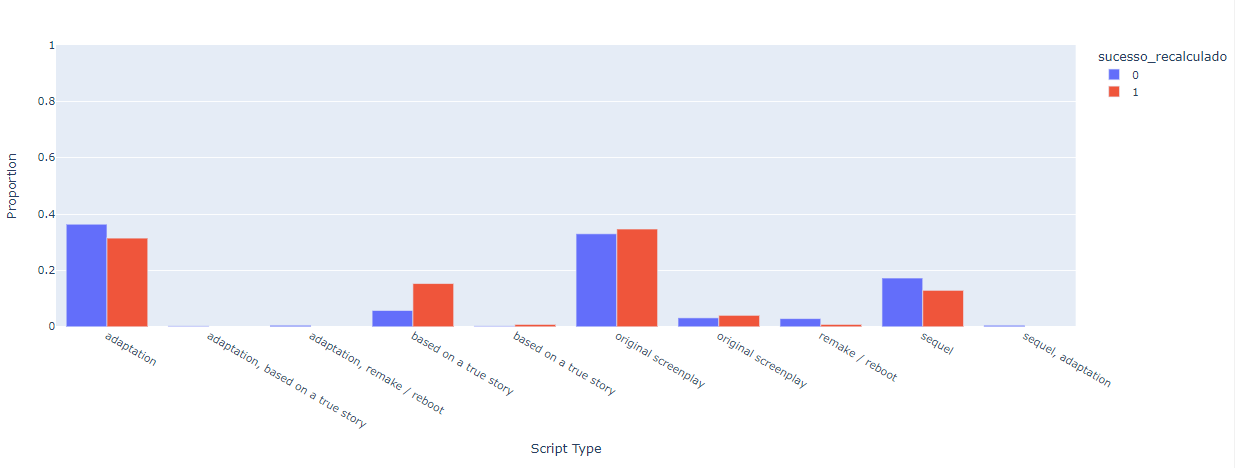
1. A arrecadação nos EUA dos filmes que são um sucesso (112 milhões) difere o dobro dos filmes que não são sucesso (54 milhões) - o mesmo acontece considerando o mundo todo;
2. A arrecadação nos outros países dos filmes que são um sucesso (93 milhões) difere um pouco dos filmes que não são sucesso (64 milhões);
3. O budget mediano dos filmes que são sucesso é menor do que o dos filmes que não são (37 x 50 milhões);
4. O budget mediano recuperado na primeira semana dos filmes de sucesso é quase o dobre do que os filmes que não fazem sucesso (4,2% dos sucessos x 2,3% dos não-sucesso);



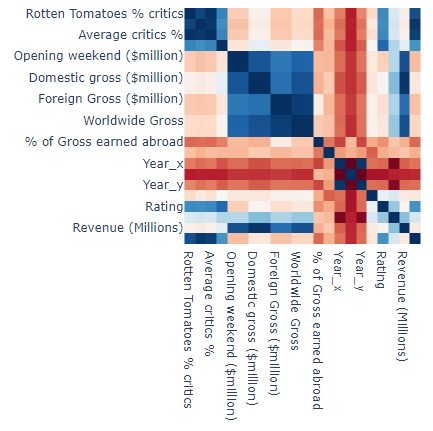
1. O tempo de duração não difere muito (115 min dos sucessos x 112 min dos não-sucesso);
2. Parece que filmes mais recentes fizeram menos sucesso que filmes menos recentes.

**Algumas conclusões sobre a EDA bivariada - qualitativa**

1. Dos filmes que são um sucesso, grande parte é adaptação, roteiro original ou sequência; de ação, comédia ou drama; mas esses também são os filmes mais frequentes, logo, era esperado



1. Praticamente a metade dos filmes de animação, de crime e de família são sucesso - isso também acontece para os filmes baseados em fatos reais; enquanto que a maioria dos filmes feitos de ação, comédia e drama (mais frequentes) não são sucesso;
2. Ou seja, se você for fazer um filme, a chance de que ele seja um sucesso parece que vai ser maior se você fizer um filme de animação, de crime ou de família - e baseado em fatos reais.

****

**Algumas conclusões da matriz de correlação:**

1. A correlação entre Budget do filme e a média das notas do Review é baixa (-0.014). Influência quase nula.
2. A correlação entre Budget do Filme e Budget Recovered é negativa (-0.17). Esperava-se que quanto maior o Budget, melhor seria o ROI

**Algumas conclusões encontradas por meio de testes de hipóteses (ANOVA e Tukey Test):**

**Hipótese Nula (H0): "o tipo de script não impacta na arrecadação mundial."**

**Hipótese Alternativa (H1): "o tipo de script impacta na arrecadação mundial."**

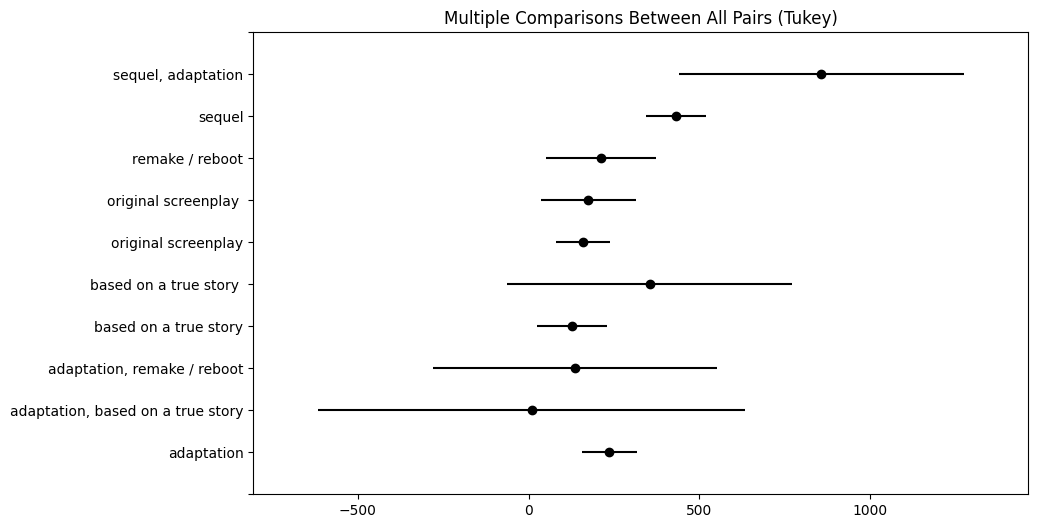
**Como p < 0,05 → Rejeitamos H0**

**o tipo de script parece influenciar na arrecadação mundial.**

1. Existem alguns tipos de script que arrecadam mais do que outros, por exemplo

* Sequência arrecada mais que adaptação
* Sequência arrecada mais que filmes baseados em histórias reais
* Sequência arrecada mais que roteiro original

Conclusão: se você estiver em dúvida de que tipo de roteiro escrever pra arrecadar muito dinheiro mundialmente, escreva uma sequência

****

**Capítulo 4: Criando os Modelos**

Nosso objetivo era criar modelos de regressão para prever a receita de filmes e modelos de classificação para determinar se um filme seria um sucesso ou não. No entanto, essa tarefa se mostrou desafiadora devido ao grande número de colunas em nossos dados.

Começamos nossa jornada com modelos básicos, como regressão linear e regressão logística. No entanto, esses modelos não atingiram nosso objetivo, e as métricas de desempenho, como o MAE (Mean Absolute Error), estavam bem abaixo do desejado, na verdade as regressões lineares foram uma falha total, apesar da regressão logística ter ficado boa decidimos não levar esses modelos para nossa versão final e continuar com os modelos mais “barra-pesada”.

Para otimizar nossos resultados, exploramos outros modelos como alternativas como Random Forest, SVM e Gradient Boosting e realizamos ajustes de hiperparâmetros e cross-validation.

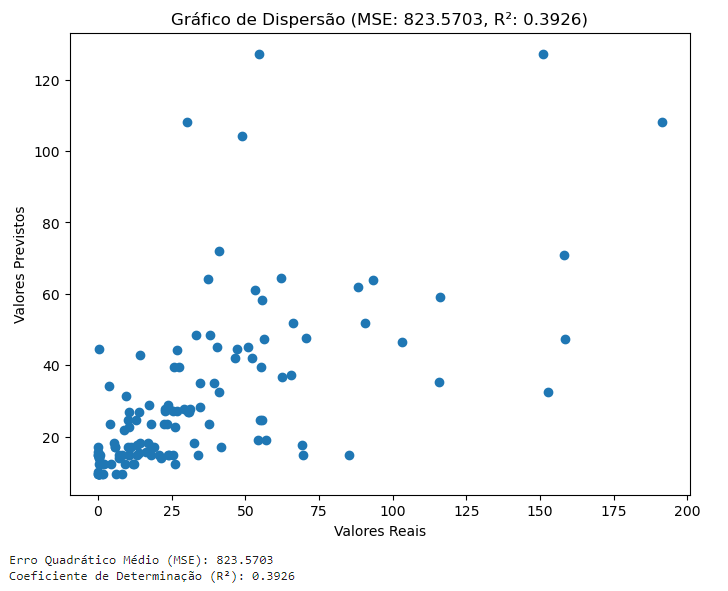
Em nossa busca contínua por modelos de alta precisão, empregamos uma técnica crucial conhecida como "Grid Search" ou busca de grade. Essa abordagem é fundamental para encontrar os melhores hiperparâmetros que otimizam o desempenho de nossos modelos.

O Grid Search envolve a exploração sistemática de várias combinações de hiperparâmetros de um modelo, como taxas de aprendizado, profundidade máxima da árvore, regularização e outros. Essa exploração é realizada por meio de iterações em uma "grade" de valores possíveis para cada hiperparâmetro.

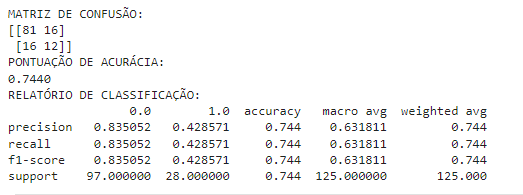
O MSE foi escolhido como nossa métrica devido à sua capacidade de medir a precisão das previsões em termos de erro quadrático médio, o que se encaixava bem em nosso problema, considerando a variância nas receitas de filmes.

Além disso, no contexto de nossos modelos de classificação, empregamos a precisão como uma métrica crítica. Isso porque como é pior classificamos um filme como sucesso e na realidade ele não ser (do que classificar como fracasso e ele ser), devemos tomar cuidado com os falsos-positivos. Essa métrica nos permitiu avaliar a eficácia de nossos modelos em identificar corretamente os filmes de sucesso, levando em consideração tanto os verdadeiros positivos quanto a minimização de falsos positivos e falsos negativos. Dessa forma, abordamos a natureza desafiadora de nosso problema de classificação, em que o equilíbrio entre a precisão e a capacidade de detecção é fundamental para obter resultados confiáveis.

Nosso modelo campeão para a regressão foi o Gradient Boosting Regressor com as seguintes métricas



E para nosso modelo de classificação, o ganhador foi o SVM (Support Vector Machine) com as seguintes métricas:



**Capítulo 5: O Lado Prático da Inteligência Artificial - MLOps**

A inteligência artificial não se limita apenas à criação de modelos sofisticados, mas também à capacidade de implantá-los com eficácia, tornando-os prontamente acessíveis e úteis para a indústria cinematográfica. Para alcançar esse objetivo, utilizamos duas ferramentas poderosas: Streamlit e scikit-learn.

O Streamlit é uma plataforma que desempenhou um papel fundamental em nossa jornada. Essa ferramenta de código aberto nos permitiu criar uma interface interativa e amigável para nossos modelos de previsão e classificação de filmes. Com o Streamlit, desenvolvemos um aplicativo web que simplifica o processo de entrada de dados e apresentação de resultados. Isso não apenas facilita o uso de nossos modelos, mas também permite que profissionais da indústria cinematográfica e entusiastas explorem as previsões e avaliações de filmes de forma intuitiva. O Streamlit tornou a ponte entre a inteligência artificial e o mundo real do cinema, possibilitando que nossos modelos fossem aplicados com facilidade.

Outra peça fundamental de nossa infraestrutura foi a biblioteca scikit-learn. Essa biblioteca é uma caixa de ferramentas versátil para aprendizado de máquina, que inclui funcionalidades essenciais para normalização e encoding de dados em tempo real. Com o scikit-learn, conseguimos implementar etapas críticas, como normalização de dados numéricos e codificação de variáveis categóricas, de maneira eficiente e consistente em nosso aplicativo. Isso garantiu que nossos modelos fossem capazes de lidar com diferentes tipos de dados e garantiram a confiabilidade das previsões.

Assim, o Streamlit e o scikit-learn não apenas tornaram nossos modelos acessíveis a um público mais amplo, mas também demonstraram a importância da integração da inteligência artificial com o mundo real. Com essas ferramentas, nossa abordagem se tornou prática e orientada para resultados, capacitando a indústria cinematográfica com insights valiosos para a tomada de decisões informadas.

Link do nosso app: <https://projetofilmetera.streamlit.app/>

Link do GitHub com todos os códigos: <https://github.com/Ciro125/projeto_filme_tera>





**Conclusão**

Desenvolvemos um modelo de aprendizado de máquina capaz de prever o sucesso crítico e a bilheteria de um filme antes de seu lançamento. Este projeto é uma ferramenta poderosa para a indústria cinematográfica, permitindo decisões mais sólidas e alinhadas com um mercado em constante evolução.

Nossa exploração demonstra que a combinação de cinema e inteligência artificial é uma receita de sucesso. O futuro da sétima arte é promissor, com a capacidade de tomar decisões informadas, otimizar recursos e estratégias, e oferecer ao público o que realmente desejam.

Como próximos passos para levar a solução desenvolvida um passo adiante, são recomendadas:

* Adicionar mais gráficos e funcionalidades no aplicativo;
* Consertar o modelo de regressão linear e adicioná-lo ao aplicativo;
* Adicionar features que ficaram de fora no modelo com o objetivo de melhorar as métricas;
* Treinar os modelos com bases maiores;
* Automatizar a atualização diária da base de filmes e do aplicativo

**Bibliografias e Referências:**

* **Tera**. (2023). Tera: Programa para Desenvolvedores.<https://somostera.com/>
* **Streamlit**. (2023). Streamlit: O Aplicativo mais Rápido para Data Apps.<https://streamlit.io/>
* **Scikit-learn.** (2023). scikit-learn: Machine Learning in Python.<https://scikit-learn.org/stable/>
* **Otimizando Campanhas de Marketing: Selecionando Clientes com Machine Learning.** (2019).
* <https://medium.com/@gouveadiones/otimizando-campanhas-de-marketing-selecionando-clientes-com-machine-learning-9f00871a3a2b>
* **The Revengers: Age of Movie Prediction Models.** (2018).<https://medium.com/ee379k-team-3-movie-revenue-prediction/hi-its-me-gilad-6d7f7734cf12>

**Links Importantes:**

Link do nosso app: <https://projetofilmetera.streamlit.app/>

Link do GitHub com todos os códigos: <https://github.com/Ciro125/projeto_filme_tera>

Base de dados de filmes: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset