**Nanodegree Engenheiro de Machine Learning**

**Projeto final**

Vinicius Benter Gomes  
14 de Junho de 2019

**I. Definição**

**Visão geral do projeto**

Na era onde o lucro se torna o fator mais importante de um produto, nas grandes obras cinematográficas cujos orçamentos são milionários, o lucro ou prejuízo de um filme na maioria das vezes determinam se a franquia vai ter uma continuação ou não. Quais são os fatores mais importantes para o sucesso de um filme? Se o filme é de ação, se o ator X é o protagonista ou se o diretor Y está dirigindo o filme. Pensando nisso como podemos utilizar a ciência de dados para prever um possível sucesso ou fracasso financeiro de um filme antes do mesmo ser produzido.

Neste problema de regressão será utilizado um conjunto de dados com 45 Mil Filmes Registrados sendo retiradas do site TMBD até Julho de 2017 com as features numéricas de orçamento, nota dos usuários e as features categóricas como gênero do filme, as keywords de identificação dos filmes, elenco, diretor.

**Descrição do problema**

É possível prever o sucesso de um filme que será lançado se deve apenas do orçamento gasto, o gênero ou o elenco? As críticas de por parte dos seus espectadores são um fator importante para este sucesso?Até que ponto podemos supor um sucesso de um filme com base nas suas características?

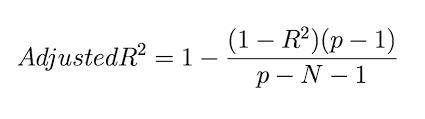
Inicialmente será levantada a hipótese que os fatores mais importantes para a receita do filme são os gêneros do qual ele pertence, orçamento utilizado para realização, o elenco escolhido e o diretor, sendo criado um modelo de benchmark utilizando regressão linear com alguns desses fatores.

Com o benchmark criado será realizado o tratamento de novos dados para posteriormente adicioná-los ao modelo esses novos fatores identificados como importantes para a receita do filme e comparado se o modelo com mais fatores possui um assertividade melhor do que o modelo de benchmark. Ao final do projeto será possível identificar quais são os fatores mais importantes para o sucesso de um filme.

Um resultado muito interessante que ao final do projeto teremos um modelo que consiga realizar uma previsão de receita do filme de um filme não lançado ou podemos simular um filme que já tenha sido lançado com um elenco diferente e comparamos se com este novo elenco a receita é melhor do que o elenco original.

**Métricas**

O problema a ser resolvido é enquadra como um problema de regressão sendo assim uma das métricas mais indicadas para este caso é o coeficiente de determinação, conhecido como R² possuindo como resultado entre 0 e 1, indicando o quanto o modelo consegue explicar os valores a serem observados, neste caso quanto mais próximo de 1 mais as características do filme podem ser utilizadas para prever a receita do filme. Entretanto neste projeto iremos utilizar a métrica Adjusted R² pois está métrica baseada no R² penaliza o modelo caso seja adicionada características menos importantes e gratifica caso for adicionado características mais relevantes.A formula a ser utilizada será:



Fórmula Adjusted R2

Sendo R2 o resultado obtido pelo modelo de regressão, p o numero de características do modelo e N o numero de amostras utilizadas no modelo, no nosso caso o numero de filmes estudados.

**II. Análise**

**Exploração dos dados**

Os dados utilizados foram disponibilizados em dois arquivos csv sendo um arquivo possuindo as características do filme sendo denominado como metadados e o segundo arquivo o casting do filme sendo denominados como créditos contendo as informações do elenco e diretor, estes arquivos serão descritos detalhadamente do próximo parágrafo.

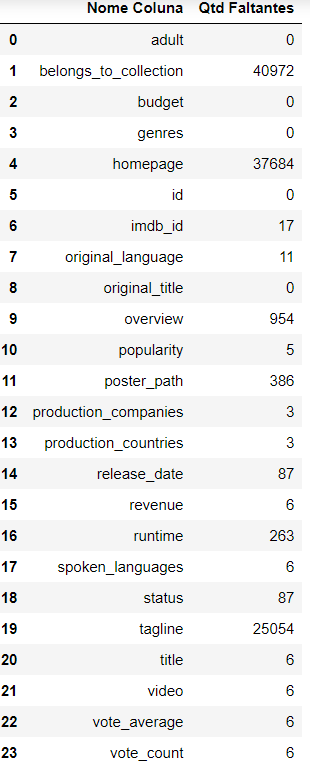
No arquivo de metadados estão descritas as características categóricas do filme como o título original, a data de lançamento, os gêneros, a língua original, a franquia do qual ele pertence e as características numéricas como a receita arrecadada e o orçamento do filme.



Dados Iniciais MetaDados

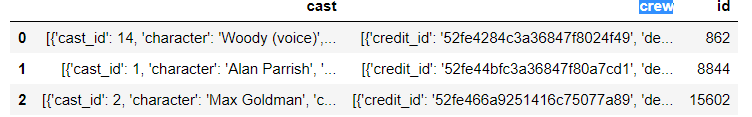
Na característica de gêneros, as informações estão concentras em apenas um campo como “{'id': 16, 'name': 'Animation'}, {'id': 12, 'name': 'Adventure'}”, devendo ser realizado o tratamento dessas informações de forma que o campo apenas possua apenas o gênero como “Animation, Adventure”.

O arquivo de metadados possui poucos que segundo a hipótese de importância de campos com valores nulos.



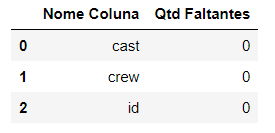
Análise Dados Faltantes MetaDados

O arquivo de créditos possui apenas duas características categóricas o elenco e a equipe que trabalhou no filme, como no arquivo de metadados o campo de cast possui todas as informações concentradas com a informação do ator e que papel ele atuou. No campo de equipe podemos tratar as informações para obter diretor do filme considerando o job “director” para isto.



Análise Dados Iniciais Créditos

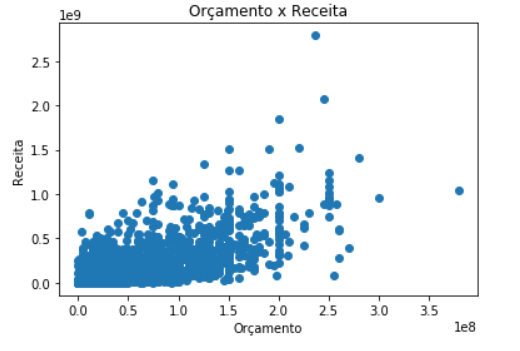
O arquivo de créditos não possui muitas informações estando compatível com o arquivo de metadados como a imagem abaixo.



Análise Dados Faltantes Créditos

**Visualização exploratória**

Para auxiliar no desenvolvimento projeto foi gerada uma visualização de que compara a relação entre o orçamento de um filme com a receita arrecadada de forma que seja possível validar que esses dois atributos estão correlacionados

****

Orçamento X Receita

Conforme o grafico acima é possivel deduzir que existe uma corelacao entre os atributos já que para os filmes com maior orcamento existe uma receita maior associdada.

**Algoritmos e técnicas**

Neste projeto será utilizado o algoritmo de regressão linear, pois neste projeto é um problema característico de regressão já que estamos tentando prever um valor com base nas características do filme.

Para ser possível a utilização do algoritmo de regressão todas as características categóricas serão transformadas em características numéricas através da utilização da técnica de LabelEncoder, que, por exemplo, irá traduzir o gênero ‘Animation’ para o valor 1.

**Benchmark**

Será utilizado como modelo de benchmark um modelo de regressão linear que contenha apenas as características do primeiro gênero do filme e o seu orçamento, sendo comparado posteriormente se a adição de novas características melhora o resultado obtido.

**III. Metodologia**

**Pré-processamento de dados**

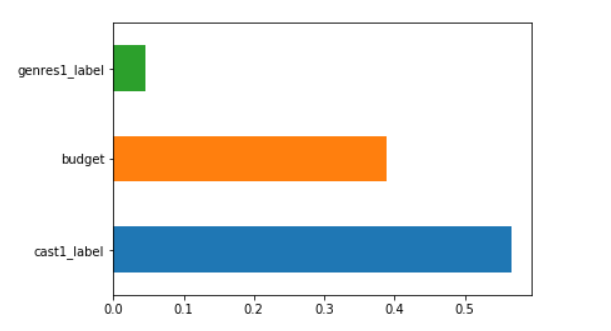
No arquivo de metadados foram retirados todos os filmes cujo orçamento seja zero ou nulo e cuja sua receita seja zera ou nula, adicionalmente será realizado o tratamento do campo de gêneros e posterior a segregação dos três primeiros gêneros em campos distintos para possibilitar o modelo considerar se todos os gêneros são importantes ou apenas primeiro.

No arquivo de créditos foi realizado o tratamento para segregar os três primeiros atores do elenco e o diretor do filme em campos distintos para aumentar as possibilidades da analise. Posteriormente foram unificadas as informações de ambos os arquivos em apenas um.

Com o arquivo unificado foi realizado o transformação de todos os campos categóricos relevantes para campos numéricos no qual seja possível a utilização do modelo de regressão linear, sendo utilizada a técnica de LabelEncoder.

**Implementação**

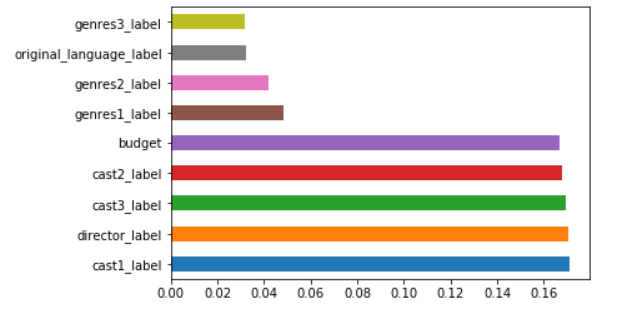
Com o arquivo devidamente tratado foi utilizado o um algoritmo de ExtraTreesClassifier no qual possibilita a identificação de quais são as características mais importantes para o conjunto de dados. Inicialmente foi realizada a verificação da hipótese prévia que as características do primeiro gênero do filme e o orçamento.



Características Importantes Benchmark

Sendo essa visualização confirma em parte a nossa hipótese inicial, entretanto para a criação do nosso modelo de benchmark foram utilizados os mesmos atributos da hipótese inicial. Para criação do modelo foi utilizada a técnica de segregação da base em 75% para treino do algoritmo e 25% para testar a eficiência do modelo, sendo utilizada regressão linear.

Após a implementação do modelo de benchmark foi gerada a visualização com os demais atributos para analisarmos os atributos que serão fornecidos ao novo modelo, sendo utilizadas no novo modelo apenas as características com o resultado superior a 0.05.



Características Importantes Modelo

Utilizado as características selecionadas acima foram fornecidas ao novo modelo de regressão linear sendo apurado o resultado, utilizando a mesma técnica de particionamento de antes 75% da base para treino e 25% para teste.

**Refinamento**

Para refinarmos os resultados obtidos foram realizados testes com outros tipos de algoritmos comumente utilizados em problemas de regressão, sendo utilizados os algoritmos de LinearRegression, RandomForestRegressor, DecisionTreeRegressor e BayesianRidge.

Para cada modelo foi realizado o refinamento através de GridSearch sendo identificado os parâmetros possíveis do modelo através da documentação disponível, para parâmetros booleanos serão utilizados os atributos True/False e para atributos numéricos será utilizado números com intervalos como 1,10,30 por exemplo.Será fornecido ao modelo final os melhores parâmetros sugeridos pelo algoritmo através da função best\_estimator.

**IV. Resultados**

**Modelo de avaliação e validação**

O modelo final escolhido foi o de Bayesian Ridge que obteve um resultado R2 0.510822, enquanto o segundo melhor modelo testado de LinearRegression obteve um score de 0.5093.

Com o modelo de regressão escolhido como o melhor foi utilizada a técnica de GridSearch com os parâmetros de fit\_intercept, normalize sendo estes booleanos e testados como True e False e o parâmetro n\_iter com os parâmetros 100,300,500,700.Os parâmetros sugeridos pelo algoritmo através da função Best\_estimador foram fit\_intercept = False e normalize = True e n\_iter = 10 , através de melhoria dos seus parâmetros obtivemos um resultado R2 de 0.510823 e Adjusted R2 0.5086.

Durante a etapa de refinamento foi utilizado o parâmetro de Cross Validation na técnica de GridSearch para testar a eficiência do modelo para dados não vistos desta forma o resultado obtido acima indica que o modelo consegue generalizar bem os dados não vistos, tornando o resultado do modelo confiável.

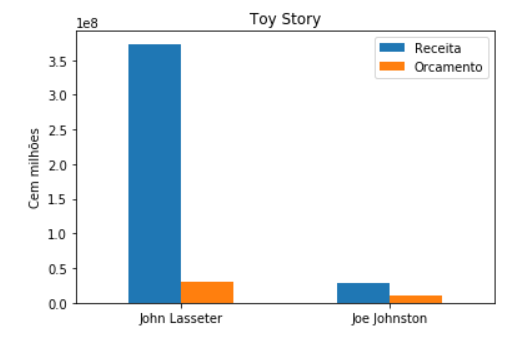
**Justificativa**

Os resultados obtidos após treinar o modelo com os atributos mais importantes estão muito próximo ao modelo de benchmark, sendo 0.5102 para o benchmark e 0.5086 para o modelo final, isto nos leva a entender que mesmo utilizando as novas características do filme o fator mais relevante passa o sucesso permanece o orçamento do filme, esse conceito é confirmado ao analisar os últimos sucessos de bilheteria que apesar do seu elenco, o marketing que compõe o orçamento utilizado para promover o filme possui um fator extremamente importante. Desta forma podemos inferir que quanto mais gasto para realização de um filme maior vai ser a sua receita.

**V. Conclusão**

**Forma livre de visualização**

Como um dos resultados obtidos foi a utilização do modelo para prever como seria o orçamento de um filme no caso do filme Toy Story com outro diretor e um orçamento menor 1/3 menor. Considerando essa características o modelo previu que a receita seria 1/10 menor do que o filme original, reafirmando a hipótese que tínhamos suposto antes, que o orçamento influencia muito o a receita do filme.



Modelo Novo Diretor e Menor Orçamento

**Reflexão**

Este projeto nasceu de uma paixão pessoal que é gostar muito do universo cinematográfico, foi sempre um pensamento curioso entender como certos filmes conseguem produzir grandes receitas para os seus estúdios e outros não.

Durante o desenvolvimento do projeto tive muita dificuldade em realizar os tratamentos dos dados e geração das visualizações, devido ao falta da experiência da linguagem, mas ao final do projeto tive certeza que para os próximos problemas deste tipo, estarei mais preparados para solucioná-los.

O mais interessante ao final do projeto foi descobrir que os gêneros do filme não são tão importantes para a receita, sendo crucial o orçamento e o elenco. Refletindo isto com a realidade dos últimos filmes lançados pode se perceber que está hipótese está correta pois muitas vezes assistimos os filmes quando determinado ator está nele ou que ele possui um bom numero de efeitos especiais de qualidade que estão diretamente ligados ao orçamento do filme.

**Melhorias**

Como algumas possibilidades de melhorar a precisão do modelo pode ser sugerido à utilização do pelo do filme, isto é, utilizar qual será a historia que o filme está contando. Poderia ser utilizado a técnica de TF-IDF para encontrar as palavras mais importantes do plot do filme e alimentar o modelo com está nova variável.

Outra melhoria sugerida é a de considerar se o filme possui um antecessor e considerar o sucesso ou fração do antecessor alimentando o modelo com esta nova variável.

A solução apresentada neste projeto é apenas o inicio de um estudo de quais fatores tornam um filme um sucesso de bilheteria, podendo ser expandido este estudo agregando as informações existentes em outras plataformas como a popularidade da franquia, atores ou diretor pelo seu numero de seguidores no Twitter por exemplo.