# Nanodegree Engenheiro de Machine Learning

## Proposta de projeto final

### Histórico do assunto

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.

Existem trabalhos relacionados ao tema como por exemplo o "Prediction of Movies Box Office Performance Using Social Media"[1] que foi utilizado como um dos parametros a popularidade dos atores para identificar o possivel sucesso do filme,que será umas das features do estudo a ser realizado.Um outro exemplo que ratifica o interresse para estre projeto é que um estudo de cinema está desenvolvendo um algoritmo similar para prever quais serão os proximos filmes de sucesso[2].

A inspiração deste projeto é pessoal, pois sou um aficionado por filmes e presto muito atenção qual é o diretor, atores que farão parte do elenco do filme e se a franquia está dando lucros para as produtoras, pois assim teremos mais filmes da franquia.

### Descrição do problema

Será que é possível prever o sucesso de um filme se deve apenas do orçamento gasto, o gênero ou o elenco? As criticas de por parte dos seus espectadores são um fator importante para este sucesso? Este problema que podemos classificar como um problema de regressão pois possuimos as features como orçamento, rating do filme, genero, diretor, elenco e querendo concluir qual será a receita do filme com base nela.Ao final do projeto poderemos utilizar o modelo para prever o sucesso de filme que ainda não foi lançado.

### Conjuntos de dados e entradas

Para o este projeto será utilizada uma base de dados disponível do site Kaggle com 45 Mil Filmes Registrados[3] sendo retiradas do site TMBD até Julho de 2017 com as features numericas de orçamento, rating do usuários e as features categoricas como gênero do filme,as keywords de identificação dos filmes,elenco,diretor.Para a analise de Ratings do filme possuimos 26 milhões de notas dos seus 270 mil usuários aos 45 mil filmes deste dataset.Este dataset apesar de possuir muitos dados,a quantidade de informações faltantes é pequena e os valores de rating por exemplo variam até 0 a 5 sendo bem consistentes.

### Descrição da solução

Para entender os fatores de sucesso de um filme, será realizada uma analise de quais são as variáveis mais importantes ao sucesso do filme e ao final do projeto teremos as respostas se o elenco influencia mais o sucesso, o orçamento é mais importante ou o gênero é crucial para isto,sendo ultilizado algoritmos de regressão linear ou outro algoritmo similar que possua uma precisão melhor para este dataset.Ao final do projeto poderemos prever com base nos filmes já lançados se um um novo filme que será produzido terá o resultado esperado ou não.

## Modelo de referência (benchmark)

Para um modelo de referência iremos utilizar um modelo de regressão linear com apenas algumas features como gênero e orçamento para ajudar a entender se a adição de mais features melhora a precisão do modelo ou não.

## Métricas de avaliação

Para analise da precisão do modelo se tratando de um problema de regressão será utilizado a métrica de Adjusted $R^2$ pois este tipo de metrica penaliza se adicionarmos features menos importantes e gratifica caso for adicionado features mais relevantes ao modelo estando alinhando ao linha de estudo a ser seguida.

## Design do projeto

Inicialmente será realizada uma analise exploratória dos dados para entender quais são tipos de dados disponíveis e um tratamento de dados para obter novos dados, por exemplo, o elenco do filme está todo relacionado em campo da base de dados, devendo ser extraído o nome dos primeiros atores do elenco deste campo. Após o tratamento destes dados será realizado um estudo de quais são os atributos mais correlacionados com a receita do filme.

Identificado os atributos mais relevantes ao modelo será utilizado a abordagem dos algoritmos supervisionados de regressão linear ou outro algoritmo que possua uma melhor precisão sempre utilizando a metrica de Adjusted $R^2$.Com o Algoritmo escolhido será utilizado o método de GridSerach para identificar os melhores parametros e assim melhorar a precisão do modelo. Com o modelo otimizado iremos comparar se a precisão dele é superior ao modelo de Benchmark.

## Referências

1 - https://www.researchgate.net/publication/262249589_Prediction_of_movies_box_office_performance_using_s (https://www.researchgate.net/publication/262249589_Prediction_of_movies_box_office_performance_using_s
2 - https://futurism.com/the-byte/fox-machine-learning-movies-ai (https://futurism.com/the-byte/fox-machine-learning-movies-ai) 3 - https://www.kaggle.com/rounakbanik/the-movies-dataset (https://www.kaggle.com/rounakbanik/the-movies-dataset)

In [ ]: