Springboard - Data Science Program
Capstone Project Proposal
By Nicholas Dean
April, 2021

- **Problem Statement**

  Can I accurately predict box office revenue of a film based on that film's various attributes and historical box office revenue information?

- **Context**

  Making movies is a huge investment of time and financial resources for a production studio, often with a large payout that is well worth it. However historically many of these decisions are made based on domain expertise and a gut feeling. Should a movie have a $5m budget or $10m budget? If box office revenue is only going to be $15m regardless, $5m would be better clearly, but what if providing a $10m budget resulted in earning $30m at the box office?

  Such a large investment is even more fraught today given that many movie theaters have been closed for over a year. This has resulted in many movies skipping the box office entirely in favor of streaming platforms. In order to make such a decision it's important to accurately predict the revenue from either option.

- **Criteria for success**

  This problem will be successfully solved when I have built and evaluated the performance of regression models that predict the revenue of movies as a function of features associated with them.

  Phase one of this process will be to organize and clean the dataset with movie attribute and revenue information so that I have a dataset to train models on. I'll also be recombining the training and testing data since it was provided in separate files. Success in this phase will involve identifying and removing unnecessary data and accounting for outliers in the data.

  Phase two will involve exploratory data analysis and identifying potential drivers of revenue for movies. This will involve examining features in relation to one another and may involve principle component analysis. Success here will involve identifying the attributes most closely correlated with revenue for a movie.

In Phase three I'll pre-process the data and begin to train several regression models on the dataset, partitioning the data and assessing the value of the mean as a revenue predictor.

In Phase four I'll train a linear regression and random forest model to fit the dataset and assess their ability to predict revenue.

- **Scope of solution space**

  This project will be limited to specifically predicting box office revenue for movies *holding box office revenue constant irrespective of COVID*. A follow-on project will examine how to predict revenue for a movie after theaters have begun to open up again and how that might compare to streaming revenue.

  This project will conclude with a model to predict revenue for a movie based on it's attributes as scored by appropriate metrics (mean average error (MAE), MAE^2, R-squared, etc). I will also make recommendations around what dependent variables could be used as predictors for streaming performance and revenue in a future project.

- **Constraints**

  The largest constraint to this project is that there has been no box office data generated in the past year, and that there remains no way to assess if box office behavior by consumers will continue on the same trend soon. The uncertainty around the effect of COVID-19 on the movie theater industry and if this past year will have been a 'pause' or 'reset' on recent box office trends is not something that can be solved for in this project.

  However it does drive home the importance of a strong baseline understanding of what the revenue of a movie *would* be in a normal world in order to have a basis for comparison and understanding of box office behavior as new trends begin to emerge.

  There will also be constraints on my ability to lay groundwork for future projects with regards to the value provided to a streaming service by films given that the business model for a movies 'streaming premier' is variable and that these companies don't make viewership data publicly available.

- **Stakeholders:**

  Movie producers and directors, studio executives, theater companies.

- **Data Sources:**

  Kaggle. (May 2019). TMDB Box Office Prediction, V1. Retrieved 04/16/2021 from https://www.kaggle.com/c/tmdb-box-office-prediction/data.

A historical dataset containing information on over 7,000 different movies and 23 different types of attribute data (language spoken, genre, IMDB rating, etc). The vast majority of this data is either strings or categorical, however I do have numeric information such as revenue and budget. Data was sourced and organized for a Kaggle Playground competition, Kaggle's original data source was the TMDB API.