Project Proposal

Movie industries lose a lot of money every year on movies that do not create profit. According to a podcast, about 80% or more of Hollywood movies fail to turn a profit. One example is 'Gone in 60 Seconds' starring Nicholas Cage and Angelina Jolie where the movie grossed to $240 million but when marketing costs and movie budet were minused, it turned out the production company had lost $212 millions. This creates a problem for movie production companies, and having a sense of profitability is essential in the company's long term survival.

A production company can greatly benefit from their past experiences in movies' successes and failures. It turns out that gauging movie's profitability is too complex for humans to understand as there are many features that a movie can have that play greatly in the prediction of success. Nevertheless, those companies can greatly benefit from modern machine learning models that can handle complex features and give out great prediction results. Therefore, employing historical movies' data and using them in ML algorithms to predict success of the movie could save companies millions of dollars on rather unsuccessful movies.

For this task, we need the maximum amount of information about every movie in the relatively not too distant past. The kind of information include a mixture of movie genre, release year, starring actor, director and other info. But most importantly, we need the revenue and budget of the movie in order to determine the net revenue of the movie (gross - budget) and hence, determine if the movie is profitable at all. Other information like critic reviews, although could be quite useful in determining success of the movie, will not be helpful/applicable for future scenarios where there are not critic reviews yet. For these reasons, the data we're going to use for this task is Daniel Grijalva's Kaggle 'Movie Industry Three decades of movies' dataset. The dataset contains 15 columns (14 features and 1 target) and 6820 rows. It contains many indicators to do the task with numerical, date, and categorical string data types. It is relatively clean and contains relatively low number of 'Nans' (unknown values) which makes it easier to wrangle. The column features are all unique, and from a human perspective, are good indicators of movie's performance. The dimensionality of the dataset makes a great deal of sense as well. Many string columns will need to encoded to make them useful in the regression task.

For this problem, we need a regression model that would estimate revenue of the movie. We are going wrangle and explore the data to get it ready for processing. Then, we're going to split the movies to train and test sets, train ML models on the train set, and test them on the test set. As for the ML models that will be used for this task, a mixture of algorithms will be used and their performances/accuracies will be compared against each other and the best performer will be chosen

and fine tuned for optimal performance. We will also manipulate the data as needed to achieve best results.