# What Does Matter in the Total Revenue of a Movie?
## Group 15 Summary Paper

Data: https://www.kaggle.com/juzershakir/tmdb-movies-dataset

Objectives:

The above dataset was used in order to determine which variables are best able to predict the net revenue of a movie. The global pandemic has hit the movie industry hard, mainly due to such a large decrease in ticket and concession sales at theatres. We are investigating to see if there are certain trends that production companies could follow in order to attempt to maximize their revenue. The main factors that will be focused on are movie popularity, budget, run time, vote count (on imdb website), release year, home page (yes or no), tag-line (yes or no), number of actors, and number of genres.

Analysis:

There are multiple methods that were used for our analysis, with the first being linear regression. A correlation matrix was created in order to compare the net revenue to the other descriptive variables. Only the vote-count variable showed some strong correlation with the movie's revenue. Apache spark ml regressor was used through databricks in order to run the linear regression. As a result of low correlation between revenue and the other variables, low R-squared values were achieved, as well as a RMSE of 150492315.282752 in the training data and 1.5266e+08 from the test data. For these reasons, it was decided that linear regression was not a suitable model in predicting adjusted revenue.

The next model used was a random forest model, specifically a random forest classifier. This model was used in order to determine if a movie having a “home page” will tend to also have a higher adjusted revenue. It was determined that there was a strong enough correlation in the model to run the home page analysis in order to maximize adjusted revenue.

A second random forest model was also used, but this time it was a random forest regressor. For this model some of the non-numeric data such as strings or Booleans were transformed in a way that they could be used in the model. For example, all of the actors from a given movie were combined into a “Number of actors” column. The challenge with this was that the current raw data did not make it easy to create this column. Actors having the same name as well as different delineators being used made it more difficult than expected. The same was done for the number of genres in a film, and the last column is a Boolean of whether or not the film has a tagline. The random forest regressor was run multiple times, each time with adjusted revenue and a different

combination of variables. The most successful run of the model was the one including both the number of actors column and the number of genres column. The number of actors/genres in a movie both had a positive correlation with adjusted revenue, while the correlation with movies which had a tag-line was weaker than expected.

The final type of model used in order to analyze the data was a XG Boost Regression model. This was run for the purpose of home page analysis. The challenges faced with this method are that there is not much documentation on feature selection and importance using pyspark. Popularity, homepage presence, runtime, vote count, vote average, release year and adjusted budget were the features used. A positive correlation was found between the adjusted revenues and our predicted revenues. While it predicted well, it still had a relatively low R-squared value of below 0.5.

Conclusions:

Four different models were used in our assignment; linear regression, random forest classifier, random forest regressor, and XG Boost regressor. Based on our analysis, we can conclude that the random forest models were the most successful in predicting adjusted revenue. The R-squared value of nearly 0.9 in the random forest regressor is much better than the low ones found from linear regression and XG Boost. Overall, we determined that higher budgets and vote counts seem to have a high importance when it comes to generating more revenue. Additionally, having a website seems to correlate positively with adjusted revenue. Having a greater number of actors in the movie also influences revenue, as this likely goes in hand with the budget of the movie. Lastly, a higher number of genres in a movie also predicts that it will in turn earn more revenue. This is likely due to the film reaching additional audiences that prefer a certain genre.