Predicting Movies' Revenue

Final Report

Ali Naji

Table of Contents

Project Idea

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.

Data Cleaning

The very first step is to import necessary libraries to wrangle the data. For example, we need to import pandas and numpy with their standard aliases. Others will be imported later as needed We then load the dataset as pandas dataframe to handle it. The next step is to inspect the dataset especially the first few rows, the column labels and their data types. This is important for the upcoming steps to have a good idea about the dataset. It helps identify, and ultimately drop, useless columns such as ID, codes, etc.

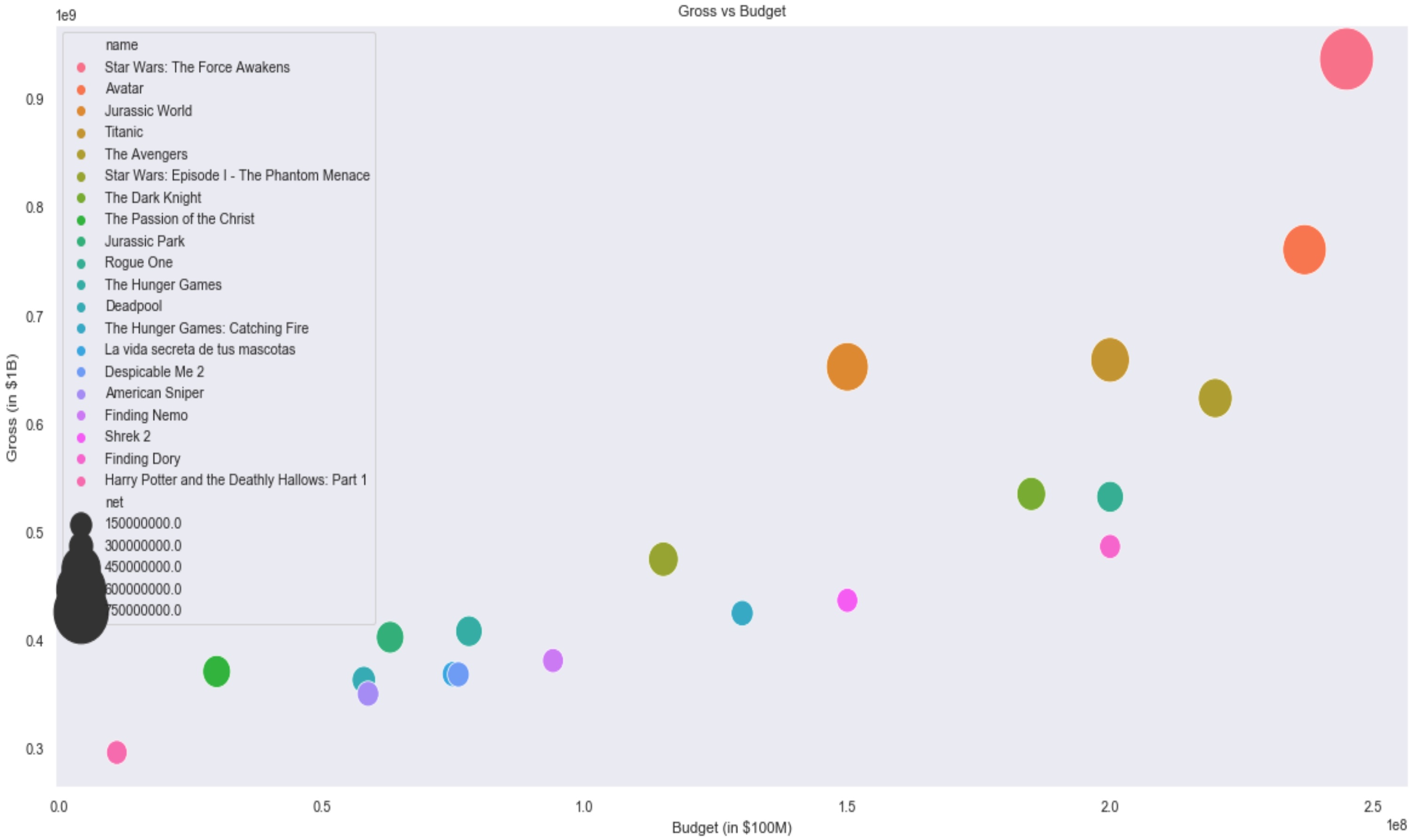
Next, we need to inspect for nulls and missing values in the dataset. In our case, there were no nulls although there were some zeros in the budget column. Going back to the dataset description, it was mentioned that the dataset had 0 budget for films for which the budget was unknown. In that case, we had to replace the zero values with some better representative values. Median was chosen as the substitute due to its independency of outliers.

In many numeric columns, there may be significant number of outliers that may need to be handled in order not to disrupt plotting and learning the data later. Those could be handled early on by creating histograms of numerical data columns and removing outliers below or beyond specified percentile of their respective columns

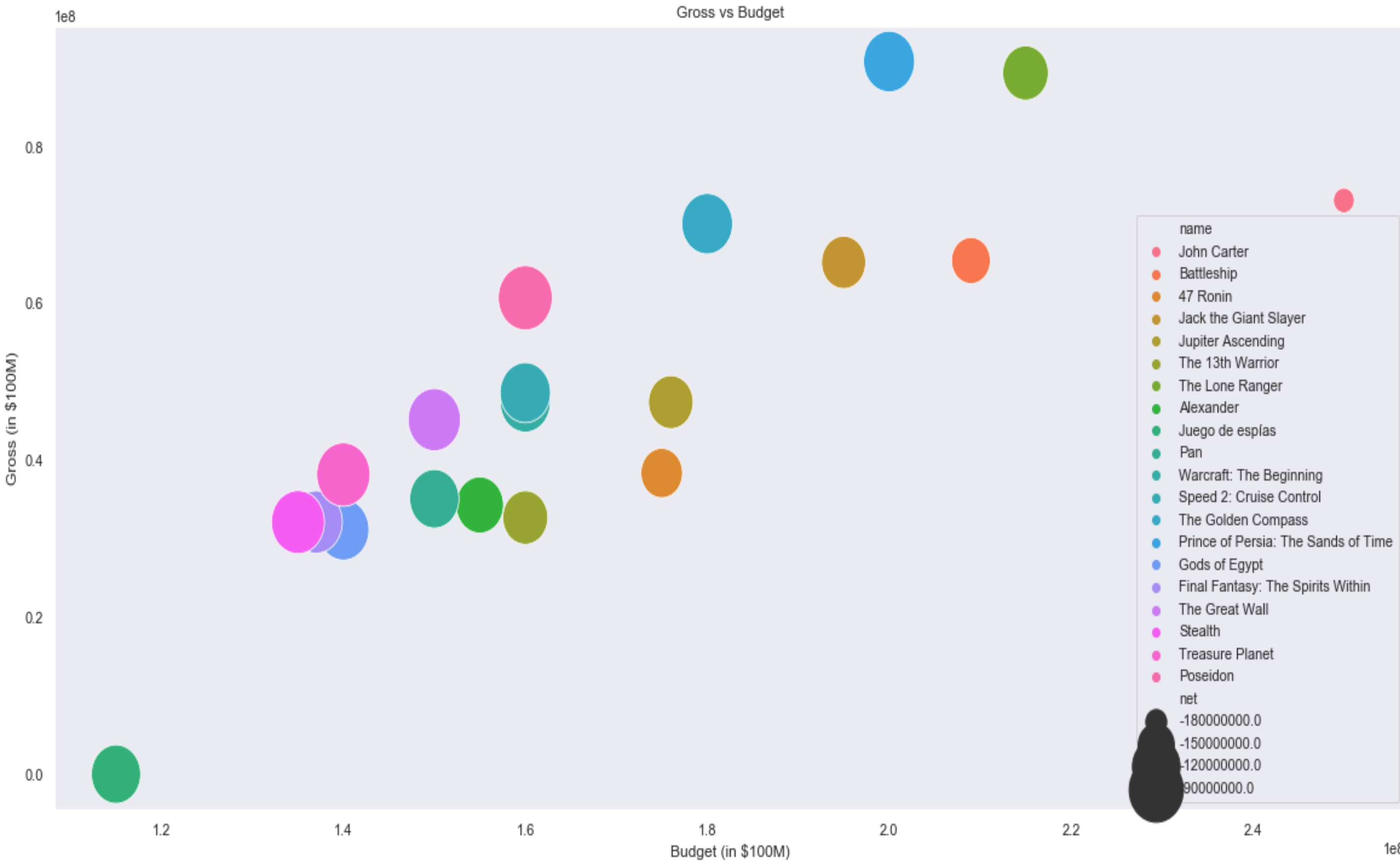
Data Story

What Movies Make Great Profits?

Movie fans have increased dramatically in the past few decades dramatically increasing demand for movies. But what kind of movies do those fans like most? And what kind of movies are most successful? The following graph shows the world's most profitable movies from 1986 to 2016 in bubbles, the bigger the bubble the more profits the movie made. The x-axis is the budget in $100M while the y-axis represents the gross in $1B. The famous action & sci-fi movie, *Star Wars: The Force Awakens* (2015), topped the list of greatest profits (net revenue, gross - budget) with net revenue of about $691M. *Avatar* had the second greatest success with net revenue of about $523M. Other films had net revenues between $285M to $500M. The common theme between most of these movies, 85% of them, is that they were released in the 21st century and in the US. 55% of them had PG-13 rating and 50% of them had action genre.



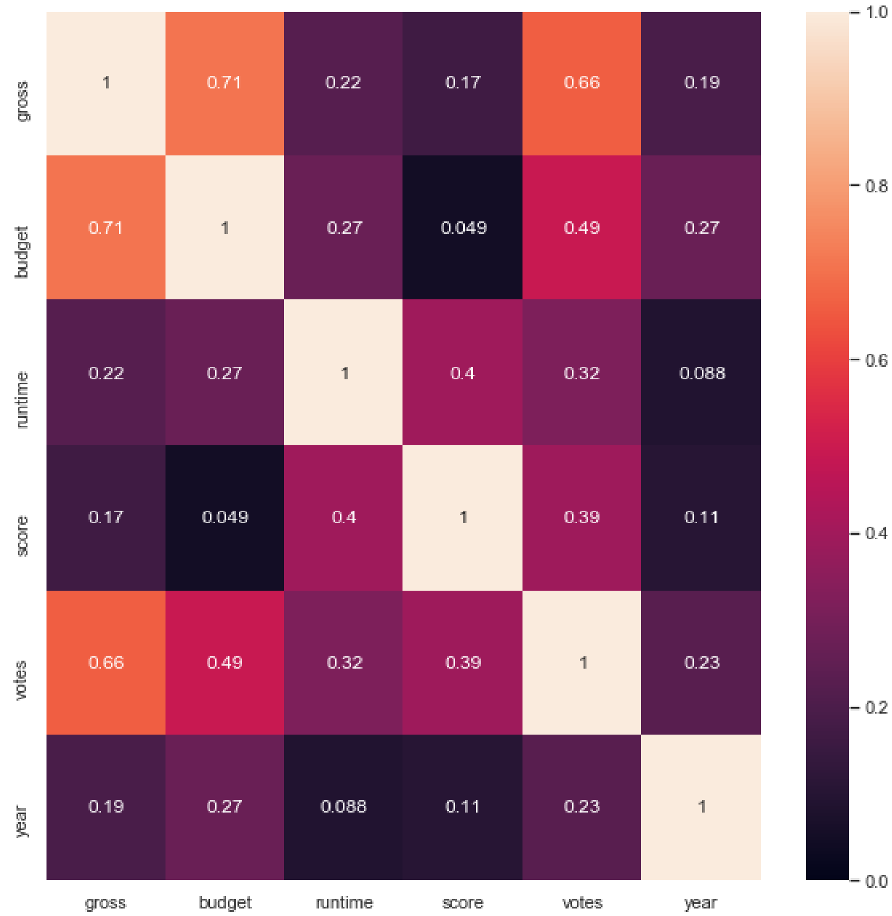
Does that mean that modern movies with PG-13 rating and action genre will most likely be successful? Not really. Looking at the worst performers during that same period in the below similar graph, we can see that action movie *John Carter* (2012) had net revenue of about $-177M (that's $177M lost) which also had PG-13 rating and was released in the US. That applies to many of the other worst performers list.



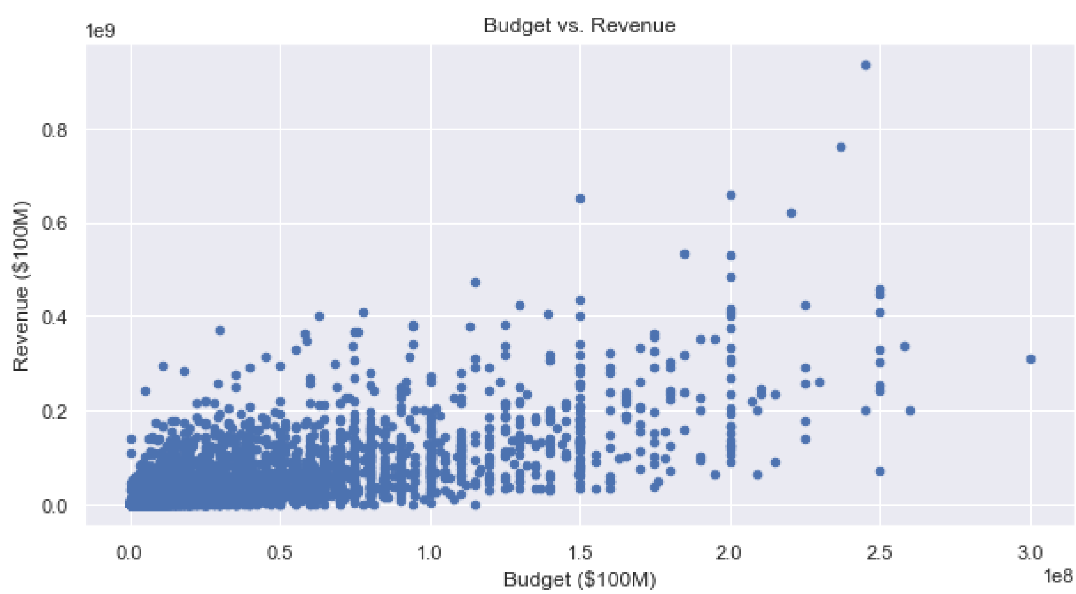
So is it possible to predict movie success based on its info? It turns out that the answer to this is complex and additional information about the movie context, its starring actor, director, and writer are needed to sufficiently answer this. In the next few sections, we will explore the data and find any correlations between different features. Later, more sophisticated methods will be used to reveal secrets behind successful movies.

Data Exploration Findings

After visualizing relations between several string columns and target variable, revenue of movie, many of them had weak or mild correlations with the target. However, finding correlations between numerical columns using a correlation matrix have revealed interesting results as shown in the map below



This shows an interesting result, that there is a strong correlation between budget and revenue (0.71 pearson correlation). When creating a scatter plot between budget vs. revenue we get the following plot:



This shows a roughly linear trend. In order to verify that this is always the case between the two, we need to do a hypothesis test where fraction of pearson correlations in permuted replicates are at least as extreme as observed one to the total number of replicates, which is p_value. If that p_value < 0.05, we reject the null hypothesis as there is statistically significant correlation

H0: No significant correlation between Budget and Revenue

H1: There is significant correlation between them

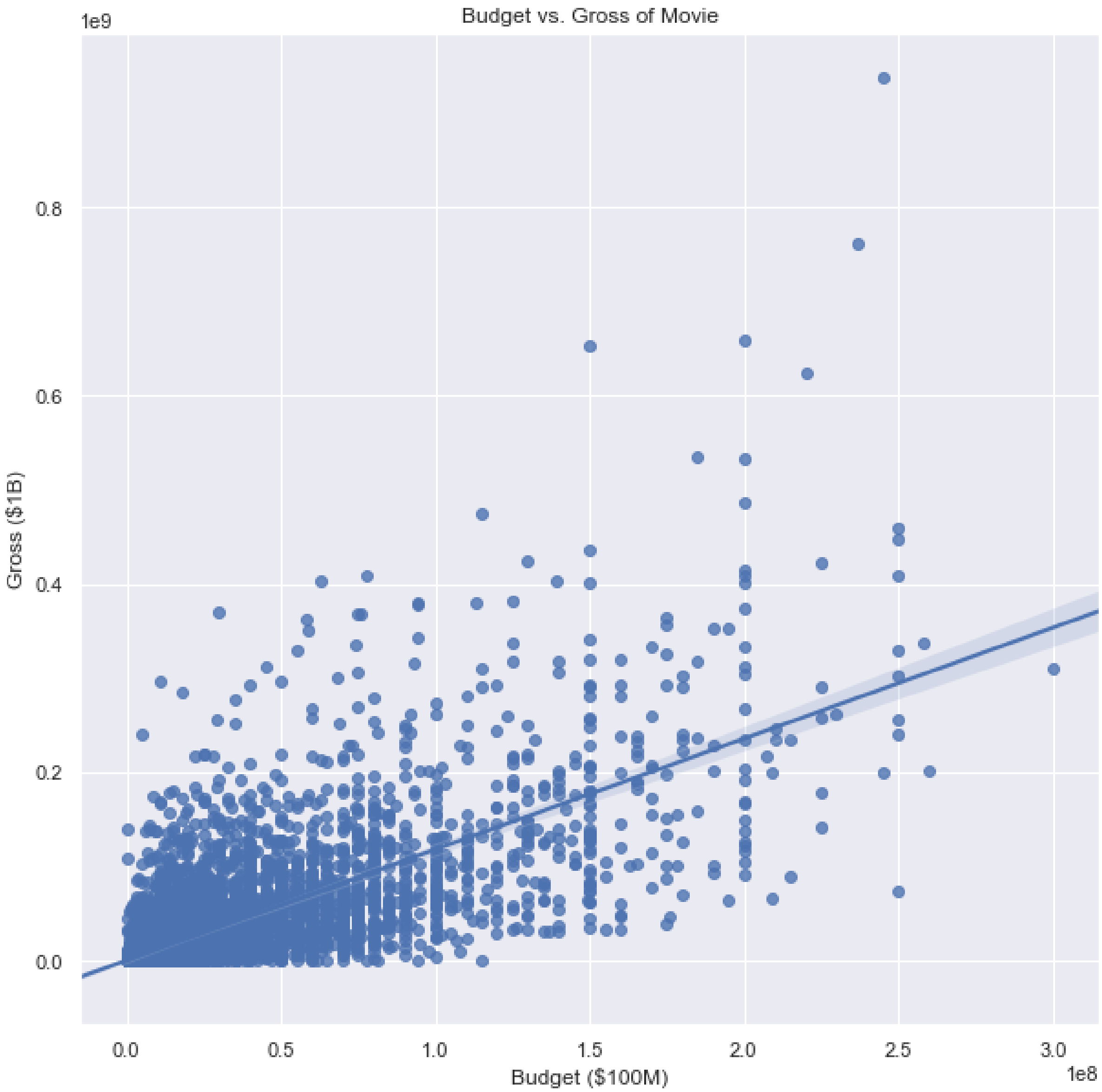
After conducting the test, we got a p_value of 0.0 which allows us to reject the null hypothesis and conclude that there is a statistically significant correlation between budget and revenue. This is a primary result that will help us create models to estimate revenue based on budget.
Next steps include encoding string columns such that they will be useful to ML models down the line. After creating ML models, we will test their accuracies and determine best model to fine-tune it further.

In-Depth Analysis

In this section, we try to create a model that is able to predict revenue of the movie with high precision. First, we try out different models, with their standard out-of-the-box parameters, to check which performs best. The best performer will then be fine tuned to achieve best results. First, we need to encode the categorical data to input them in any model created. We will numerically encode these categories (0 to N-number of categories). The string columns we have are the company, country, director, genre, name, rating, star, and writer of the movie.

After encoding categorical columns, our dataframe is all numerical type and can be used in ML algorithms. We split the data to train and test sets. We will choose 30% for test set and specify random state for reproducibility.

Based on our hypothesis test from previous section, the first model we will create is a linear regression model to estimate gross based on budget only. The model will be using L2 regularization (Ridge regression) as it is less prone to outliers. We will instantiate, train, and test the model and then evaluate its performance using root-mean-squared-error (RMSE). A regression line between the two would look like the following figure:



After creating the ridge model, predicting test set, and evaluating its RMSE, the resultant RMSE is about 34E6. That is, for every gross prediction, there is, on average, about 34M dollars in tolerance about the true revenue. Although this is a powerful result, we will investigate other models that will account for all other variables in the decision to see if it is possible to narrow down this tolerance We will create a function that will instantiate, train, and test different algorithms and return their results in a dataframe. We will use 5 different regressors:

- Random Forest regressor: Ensemble method with lots of randomization to account for diverse features of the dataset
- Adaptive Boosting regressor: Ensemble method learning sequentially paying special attention to the mistakes of its predecessor predictor

- Gradient Boosting regressor: Ensemble method with sequential learning allowing it to adapt for great complexities
- K-Nearest Neighbor regression: predicting outcome from weighted average of neighbors' values
- Ridge regressor: Linear regression model with L2 regularization

After running the function with the training and testing data and printing the results matrix (shown in the table below), we see that Gradient Boosting, with test accuracy of about 76%, followed by random forest and adaboost regressors perform best with similar accuracies on test data compared to Ridge and KNN models. It is worth mentioning, however, that Gradient Boosting had significantly lower train accuracy compared to random forest and adaBoost. Random forest and adaBoost regressors, for example, had about 13% and 16% higher train accuracies, respectively. This indicates that those models are overfitting the training set and need to be simplified to account for variance in the data.

Model	Train_Score	Test_Score
Gradient Boosting	84.3%	76.53%
Random Forest	96.16%	75.46%
AdaBoost	99.93%	74.44%
Ridge	63.85%	65.74%
KNN	77.01%	55.56%

Next, we proceed with Gradient Boosting regressor. We will fine-tune this model by creating a grid of parameters and possible values using 'GridSearchCV'. The parameters we are going to tune are learning rate, number of estimators, minimum number of samples/leaf, and maximum depth of base tree. After doing the grid search, we have found the optimal parameters for the model to perform best. Plugging those in the model and testing the accuracy, we get an accuracy of about 78%

After tuning the model and testing its accuracy, it's more appealing to quantify error in a way that makes sense. Since our task is regression, we will use root mean squared error (RMSE) as our evaluation metric as it's directly interpretable to measurement units, which is dollars in gross in this case. We evaluated RMSE to about 27E6. That is, for every movie's gross prediction, there is, on average, $27 million tolerance about the true gross/revenue. This is huge improvement since our budget-gross ridge regression model. Although this amount seems high as movie companies gross estimation may be off by that amount, nevertheless, this amount makes only small portion of most of nowadays movies' budgets and estimating movie gross with that tolerance may still be of great benefit to evaluate a movie even before its make.

Conclusion

The main task was to predict movies’ revenues based on their information like genre, release date, etc. with relatively high accuracy to give movie companies/makers an idea about how successful a certain movie may be. After exploring the ‘Movie Industry’ dataset, we have found that some parameters correlate strongly with gross of the movie, namely, the budget. The latter was tested for correlation with gross and it was found that it has significant correlation. A regression model was created and its performance was evaluated. Other regression models that utilize all features of the dataset were used to improve performance. Those have shown better performance than budget-gross regression model. The best model had an accuracy of about 78% with RMSE of abou $27M

Next Steps

These models and their results, although may be helpful to lots of movie-making companies, need a lot more data and context to accurately predict revenue. For example:

- More information is needed about the trend of movies in every year and how people's’ taste about movies change over time.
- Historical information about the movie like past remakes, classic versions, and/or movies with similar scripts, etc.. can help decide if a movie will be successful
- Audience of the movie information, such as how large of an audience the movie targets, can greatly help predicting turnout towards a certain movie. That does not necessarily mean age-rating but other indicators like movies about pets, kids, space, disabled people or general public movies like superheroes movies.

Having those predictors in would have nailed down the prediction tolerance greatly. Movie makers need to study those indicators and account for them in order to produce models with superior accuracy.