Predictions for Movie Performance

Final Report

Bingxin Wu (bw383)

Yiwen Jiang (yj76)

Contents

1 Project Description

During the past decades, the movie industry has drastically evolved. With thousands of movies produced and released each year, only a few movies stand out. Although the gross returns for movies have been increasing as a general trend, the cost of making these films has also increased, which resulted in a higher risk in movie-making. Therefore, it is important for the producers and investors to have a brief idea of how the outcomes of their movies might be beforehand.

In this project, we tried to explore ways of predicting two fundamental criteria of movie performance: box office profitability and reputation, which is measured by public ratings from the IMDB website. Because of the large MSE models for predicting profitability generally create, we later tried to simplify the profit segment to a classification problem, i.e. if the movie will have a positive profit or not.

2 Data Exploration

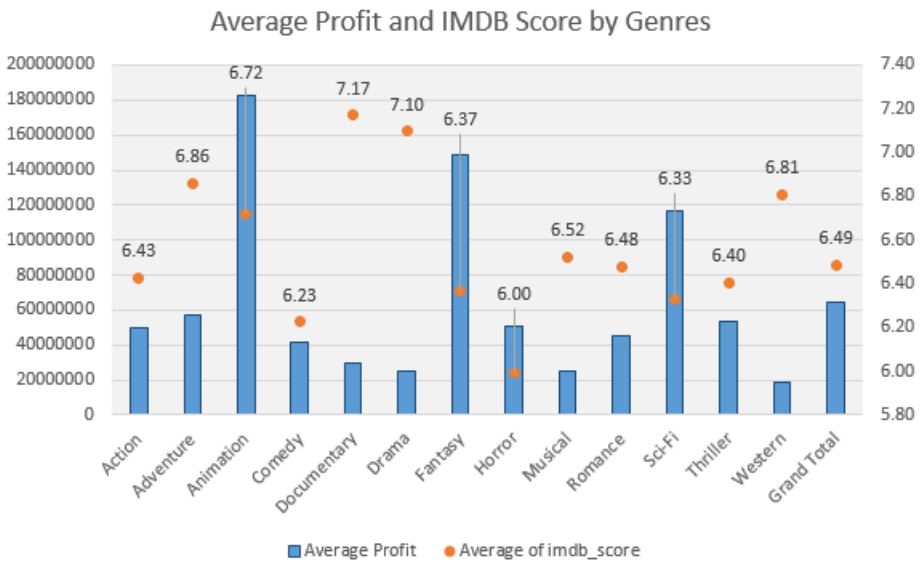
2.1 Data Cleaning

We used 3 datasets in this project: "IMDB 5000 Movie Dataset" and "The Academy Awards, 1927-2015" from Kaggle, and the budgets and gross information from The Numbers. The initial IMDB dataset has only stands for sales in the US market, and many of the budget and gross revenue listed used different currency without indicating such. We hence used the "Fuzzy Lookup" add-in from Excel with threshold of 90% to match the title name in order to correct the monetary entries using the dataset from The Numbers. In addition to the 28 variables at hand, we added two new variables: profit, which is world gross revenue minus budget, and single genre, which is the dominant characteristic of the movie out of all the genre types listed.

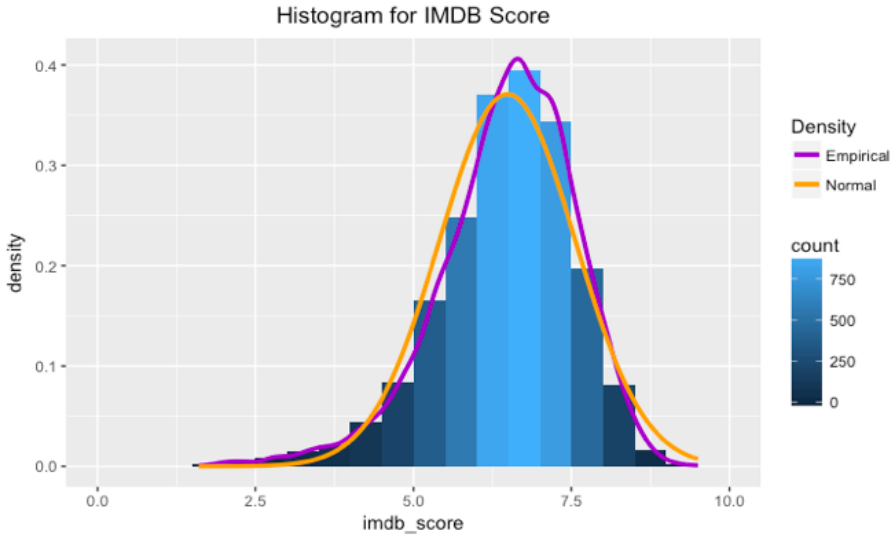
Reorganizing and matching the Oscar data, we added 4 more features to the IMDB dataset: whether the movie was nominated for best pictures, whether it won the prize, whether the director has been nominated in the past years, and the number of major actors and/or actresses who have been nominated in the past years. For the empty entries of variables involving Facebook (FB) likes, such as the number of likes on the FB page of the movie, we entered 0, since the absence of FB page is also an indication of lack of attention. For the small amount of other minor missing values, we assumed that they were missing completely at random (MCAR) and continued with the analysis.

2.2 Data Visualization

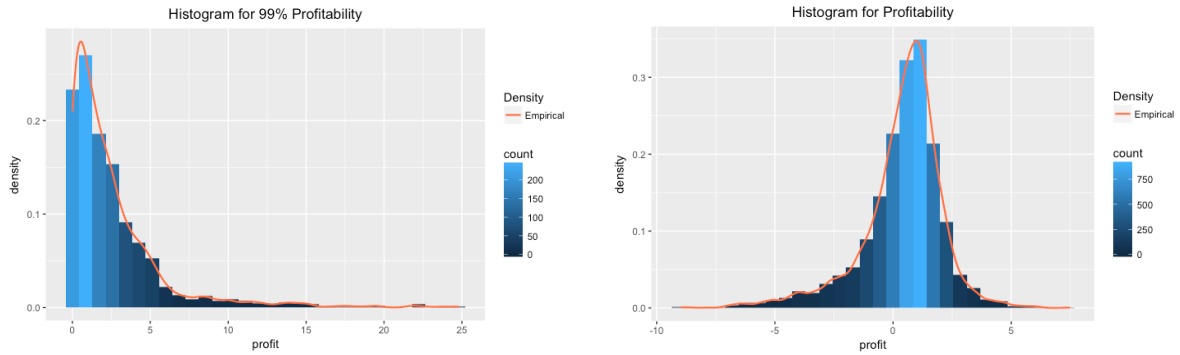
First of all, we looked at how movies of different genres perform in terms of profit (the difference between world gross and budget) and IMDB scores. From the bar plot, it can be observed that while documentary and drama have the highest average IMDB scores, they are among the least profitable genres. Animations, Fantasy and Sci-Fi are the three most profitable genres, which met our expectation, since these movies are family-friendly and are generally believed to attract more audience. We added a variable called genres_high which records 1 for movies in these three categories, and 0 otherwise.



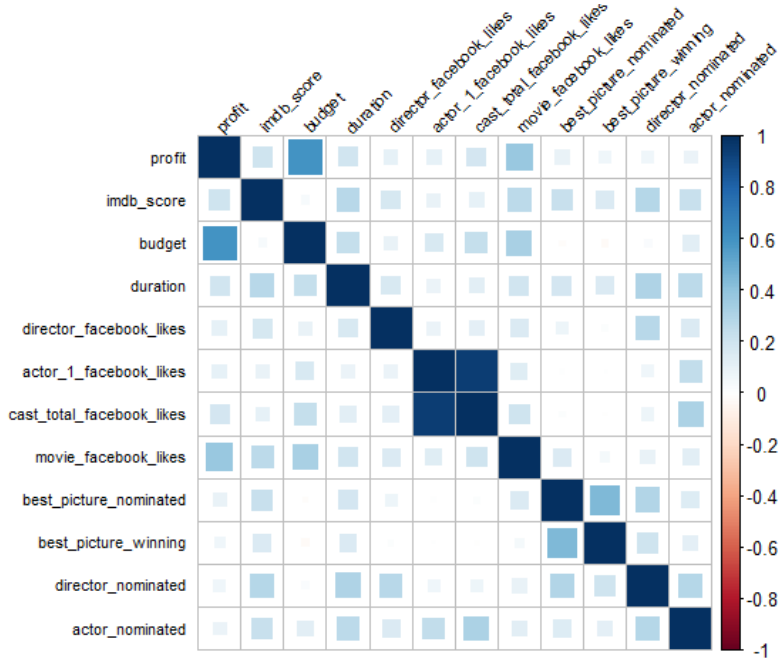
The variables of interest for this project are ratings and profitabilities. We first generated a histogram of the IMDB scores, and observed that it can be fitted to a normal distribution with mean 6.49 and standard deviation 1.0753, which follows the law of large numbers.



Similarly, we generated a histogram for profitability index, which was calculated by profit over budget. To get a more detailed look at the graph, we excluded the upper 1st percentile of the data and recreated the following left plot. The distribution is strongly skewed to the right. To make the data easier to use and reduce the effects of outliers, we wanted to apply the natural log transformation to the data. However, because there were a lot of movies at loss, these negative numbers would return NA as a result of taking logarithms. Therefore, we redefined profitability as world gross over budget, and then conducted the log transformation to obtain the normally distributed array, as shown in the following plot on the right. In addition, we created another variable called profit_or_not, which stores 1 if the movie has profit, and -1 if it does not.



Given the massive amount of features in the current dataset, we would eventually want to decrease the number of variables used for further modeling. We first created a correlation plot of chosen numeric variables to look at the pair-wise correlations, and find out which features may have significant impacts on our variables of interest. From the plot, we can see that budget and movie FB likes have relatively strong correlations with profitability, while duration, movie FB likes, whether or not the director has been nominated in previous years have relatively strong correlations with IMDB scores.



3 Model Selection

In this section, we separately tried out different methods for predicting IMDB scores and profitability. Intuitively, we chose reasonable variables to put into the model. One particular thing to notice is that we would test the effect of IMDB score on profitability, but not the other way around. The reason for this choice is that IMDB scores stabilize in the first few days after the release of the movie, and usually then influence the movie-picking choices of general audience, which eventually affects the profitability of the movie. Because the box office profit information can only be collected after the movies are off theaters, it makes no sense to use this later-collected information to predict an already-existed variable, the IMDB score.

3.1 For IMDB Scores

3.1.1 Quadratic loss with regularization

We separated the dataset into a training set (80% of data) and a test set (20%), and within the training set we used 15% of the data to perform validation.

Since the scales of the variables in the dataset varies a lot, we decided to standardize our data to ensure comparable variance among the different variables.

We first tried quadratic loss with no regularizers as a basic model. However, after fitting the model we found that the variables are not linearly independent, and thus we should add regularizers to remove some features. Hence we chose to add l_2 regularizer to the quadratic loss function. To choose the optimal lambda for l_2 regularizer, we ran used 10 fold cross validation on the validation dataset, and the lambda that minimizes MSE is 0.0112.

The objective function of this model is: $minimize \sum_{i=1}^{n}(y_i - w^T X_i)^2 + 0.0112 \sum_{i=1}^{n} w_i^2$

The resulting MSE of this model is 0.5278.

3.1.2 Huber loss with regularization

Huber loss gives smaller penalties for large outliers compared to quadratic loss. Since there are potential outliers in our dataset, we decided to try huber loss with l_2 regularization to see whether it will provide better predictions. Again, we first tried to find the optimal λ for the l_2 regularizer using 10 fold cross validation. The λ that minimizes the MSE is 5.

Hence the objective function of this model is: $minimize \frac{1}{n} \sum_{i=1}^{n} huber(y_i - w^T X_i)^2 + 5||w||^2$

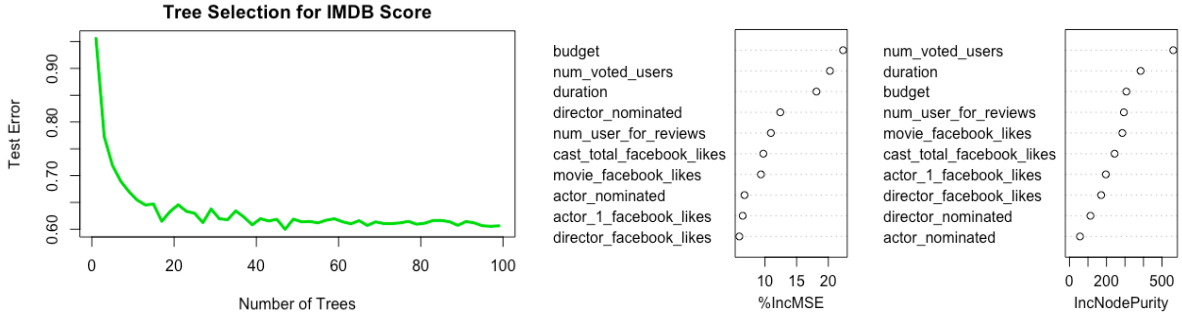
The resulting MSE of this model is 0.6563.

3.1.3 Random Forest with Regression

The random forest algorithm creates multiple groups of decision trees, which are randomly assigned a chosen number of features to split on. Because of the large amount of decision trees there are, overfitting is prevented.

To get a good estimate, we first completed a pilot run with pre-selected variables to find an appropriate number of trees to grow. We used 3 for the number of variables randomly sampled as candidates at each split ($p \approx 10/3$). Based on the output, the optimal number of trees to choose is 70, and the resulted testing MSE is 0.5598.

The graph on the right shows which variables have the greatest influence on the IMDb scores. The %IncMSE measures the increase in MSE of predictions based on the out-of-bag cross validation method, and is a robust and informative measurement. On the other hand, the IncNodePurity finds the variables in the order of how well they achieved node purities. The higher up the feature is on the list, the better it fits the criteria of having low intra-variance and high inter-variance. Therefore, we looked at the %IncMSE table, and observed that the most influential variables are budget, number of users voted on IMDb website, and the duration of the movie. Whether or not the director has been nominated in previous Oscar Awards, the number of users who wrote an review, the total FB likes of the cast numbers, and movie FB likes also have substantial effect on IMDb scores.



3.2 For Profitability

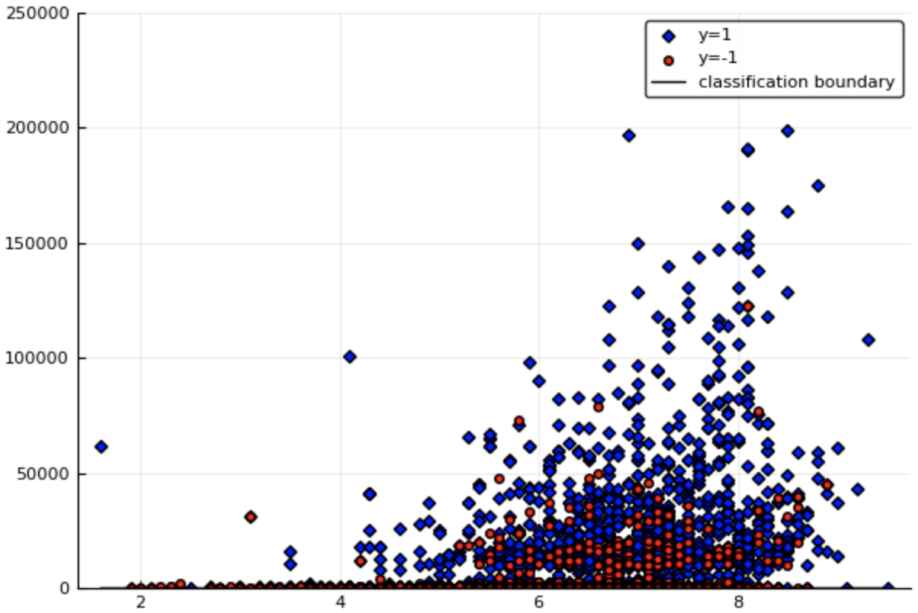
3.2.1 Quadratic Loss

We first tried to predict profitability using regression. We separated the dataset in the same way as for predicting IMDb scores, that is a training set (80% of data) and a test set (20%), and a validation set using 15% of the data. However, the result we got has huge MSE and the predicted vs. true fit is poor. We then added l1 and l2 regularizer to select appropriate variables and also to prevent overfitting. But the results were still not satisfying.

This shows that it is hard to accurately predict profitability, and thus we decided to turn it into a classification problem instead, that is to classify the movies into those with positive profitabilities (gross/budget>1) and those with negative profitabilities. And then predict whether a movie will earn profit or not.

3.2.2 Perceptron

The first classification method we tried is the perceptron, which is a simple method used to classify data into two groups. For our project, we encoded the movies with positive profitability as 1, and those with negative profitability as -1. Given the results from the feature engineering we did, we selected movie FB likes and IMDb score as the two factors to predict whether a movie will be profitable or not. However the classification boundary we got does not provide useful information for classification. Hence, a more sophisticated classification model is needed.



3.2.3 Hinge Loss

A suitable loss function for classification problems is hinge loss, which is defined as

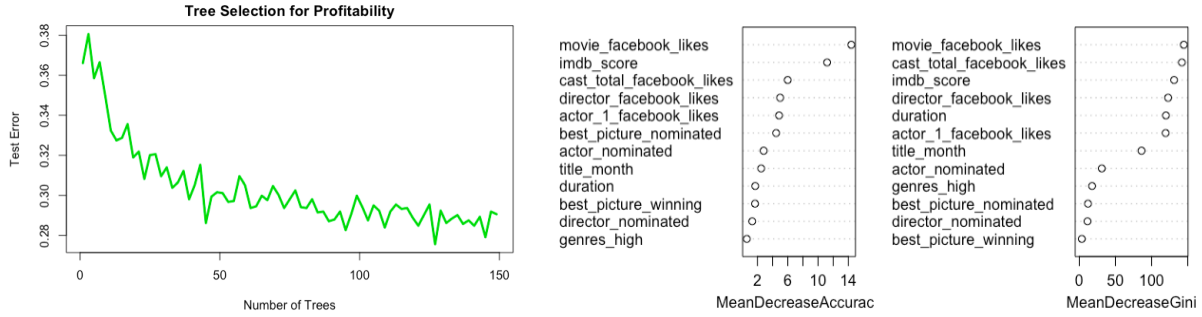
$$minimize \sum_{i=1}^{n} (X_i, y_i; w) + \lambda ||w||^2$$

There are 34 variables in our dataset (28 in the original dataset and 6 others we added ourselves), thus to prevent overfitting, we tried l_1 and l_2 regularizers to remove insignificant variables. Running the algorithm gave us misclassification rates of 36.25% and 35.57% respectively.

Therefore the overall accuracy of classification using hinge loss with regularization is around 64%.

3.2.4 Random Forest with Classification

We pre-selected features and applied the random forest algorithm with 100 trees to predict whether the movie will obtain a positive gain or not. Applying the model to the testing set, we obtained a misclassification rate of 28.75%. The list of used features are listed below in the chart, and we could see that the two most influential features are movie FB likes and IMDb scores. FB likes of the cast members and the director also moderately affect the financial performance of the movie.



4 Conclusion

For predictions of the IMDb scores and the ability to gain profit, we conducted several methods each, and summarized the results in the following table. As can be seen, for the IMDb score prediction, the quadratic loss with l_2 regularizer returned the lowest testing MSE at 0.5278, and for whether or not the movie can make a positive profit, the random forest with classification gave the lowest misclassification rate at 28.75%.

Prediction Methods and Errors			
IMDB Score		**Positive Profit or Not**	
Methods	**Mean Square Error**	**Methods**	**Misclassification Rate**
Quadratic Loss with l2	0.5278	Hinge Loss with l1	36.25%
Huber Loss with l2	0.6563	Hinge Loss with l2	35.57%
Random Forest with Regression	0.5598	Random Forest with Classification	28.75%

We are fairly confident about our models and the results obtained. We would highly recommend the score prediction model to the movie industry. With thousands of movies used in the training set and a low MSE for the test set, we believe this is a robust model that can be used for future predictions. However, the profit prediction model should be used with caution for two reasons. One is that our current best model uses random forest, which can be very hard to interpret. The other reason is due to the precision of the models.The accuracy of our classification model is only 71.25% at maximum, but because the investment on movies is usually massive, the investors might want to use other analytical tools to further reduce their risk.