Predictions for Movie Performance

Midterm Report

Bingxin Wu (bw383)

Yiwen Jiang (yj76)

1 Data Description

1.1 Data Cleaning

We first looked at the data statistics to have a general idea of the dataset. The original data we downloaded: "IMDB 5000 Movie Dataset" have three major problems when being used to examine the profitability of movies. Firstly, many entires were missing, secondly the variable *gross* only stands for the sales in the US market, and thirdly, many movies used different currency for budget and gross revenue information without indicating such.

To solve these problems, we obtained a new dataset of movie budgets and gross income from "the-numbers.com". Because the movie names in these two datasets are slightly different, we used the "Fuzzy Lookup" add-in from Excel with threshold of 90% to match the title name and generated a new dataset. In addition to the data collected, 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. For the empty entries, we assumed that they were missing completely at random (MCAR) and continued with the analysis.

1.2 Data Visualization

We are most interested in the ratings and profitabilities of the movies, and therefore we first generated the following two plots. We first plotted a histogram of the IMDB score, and observed that the scores can be fitted to a normal distribution with mean = 6.49, and standard deviation = 1.0753, which follows that law of large numbers.

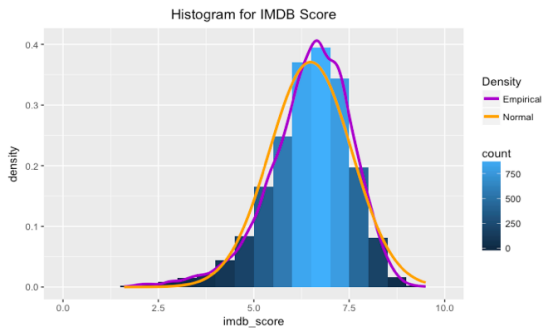

Figure 1.2.1: IMDB score Histogram

Similarly, we generated a histogram for profits. To get a more detailed look at the graph, we excluded the upper 5th percentile of the data, and recreated the plot.

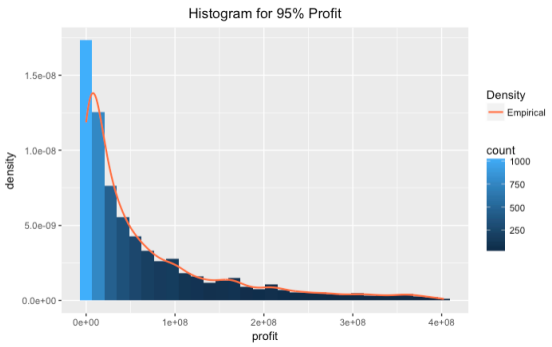

Figure 1.2.2: 95% Profit Histogram

The distribution is strongly skewed to the right, which means that in the future analysis, we should look at the median = 13514848 instead of the mean = 64760472 to obtain a more precise interpretation.

Furthermore, we looked at how movies of different genres perform in terms of profit and IMDB scores.

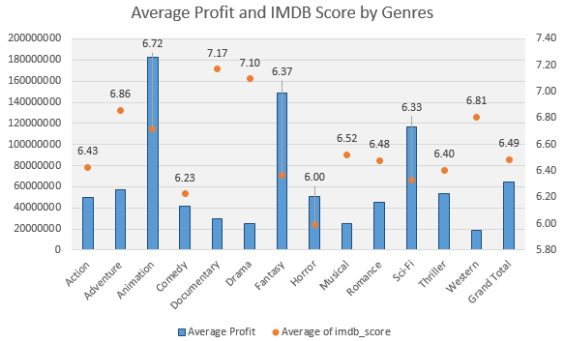

Figure 1.2.3: Barplot showing average profit and IMBD score for different genres

From the barplot, 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 being the three most profitable genres fitted our expectation, as these movies are familiy-friendly and are generally believed to attract more audience.

The boxplots below give a better understanding of the score distribution within each genre.

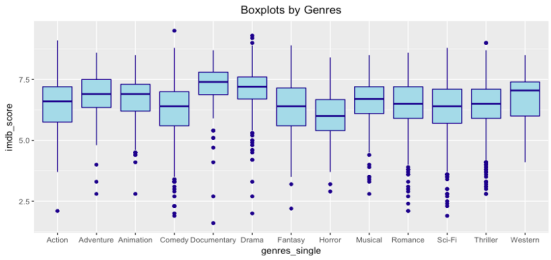

Figure 1.2.4: Boxplots showing the imdb score for different genres

2 Preliminary Analyses

2.1 Feature Analysis

Because of the massive amount of predictors in the dataset, we would eventually want to decrease the number of variables being used for further modeling. This will return a better model with smaller AIC and less chance of overfitting. From the correlation plot, we can see that there are strong correlations between budget, gross and profit. Also, IMDB scores tend to be correlated with cast facebook likes and duration of the movies.

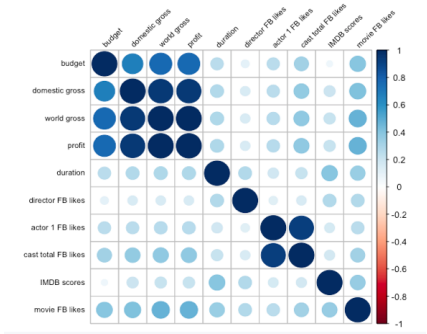

Figure 2.1.1: Correlation plot of selected factors

Two separate methods were used to complete the process of reducing predictor numbers. Knowing the correlation matrix, we attempted to use PCA to select the principal variables, the ability of explaining variances of which were listed in the next table.

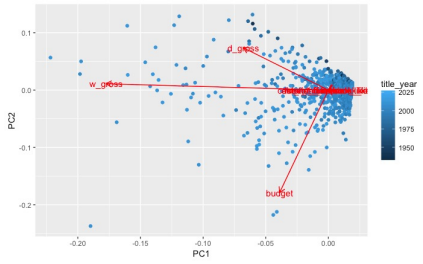

Figure 2.1.2: PCA

```
Importance of components:
                          PC1       PC2       PC3   PC4   PC5   PC6   PC7   PC8
Standard deviation     1.655e+08 2.808e+07 1.762e+07 22537 18700 3840 3082 18.26
Proportion of Variance 9.614e-01 2.767e-02 1.089e-02     0     0     0     0  0.00
Cumulative Proportion  9.614e-01 9.891e-01 1.000e+00     1     1     1     1  1.00
```

Figure 2.1.3: Variable Importance

Next, we used random forest. To get a good estimate, we first completed a pilot run to find an appropriate number of trees to grow. Given the output, the optimal number of trees to choose is about 200.

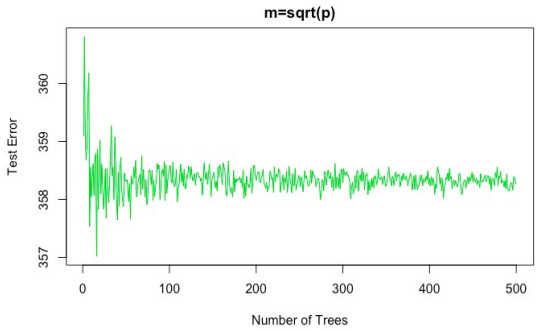

Figure 2.1.4: Pilot run MSE

Under this context, the selected variables and their importance are listed in the following charts.

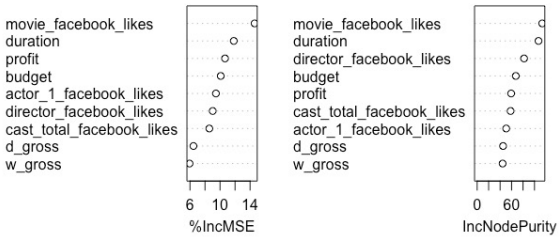

Figure 2.1.5: Variable Importance

```
          Mean of squared residuals: 0.7935679
                    % Var explained: 36.64
[1] 0.606471
                           %IncMSE IncNodePurity
budget                   10.096591      67.39901
d_gross                   6.460972      45.18650
w_gross                   5.942565      44.39448
profit                   10.666142      59.29092
duration                 11.831430     107.00088
director_facebook_likes   9.026663      81.82396
actor_1_facebook_likes    9.456715      50.65810
cast_total_facebook_likes 8.577402      58.30872
movie_facebook_likes     14.605699     113.48140
```

Figure 2.3.2: Variable Importance

2.2 Linear Regression

The first analysis we did was a linear regression showing the relationship between budget and profit of the movies.

The fitted line is: y=-5.934e+06+3.082x, with both coefficients being significant. The R^2 of the regression is 54.72%, which is a satisfying result for linear regression.

The scattered points give a growing trend in budgets as times passes.

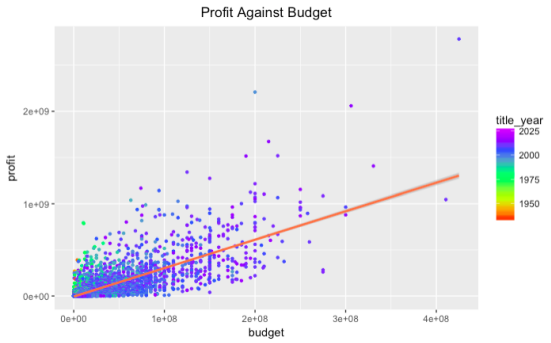

Figure 2.2.1: Linear regression of budget vs profit

Our correlation plot suggested that duration is a significant indicator for IMDB score, and hence we did another linear regression of score against duration. Our explanation for the significant slope of the fitted line is that a more complicated and well-illustrated storyline will require longer time to present. Movies are generally designed as 90 minutes to 180 minutes long.

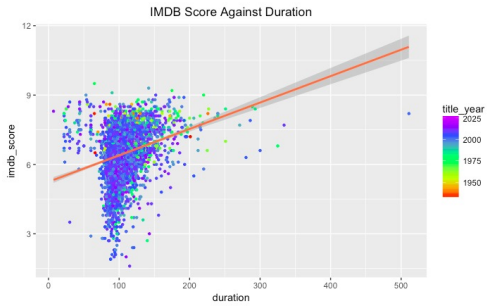

Figure 2.2.1: Linear regression of budget vs profit

2.3 Bagging

We also built a model with bagging to look at the IMDB scores. The original data were separated into two groups: training set and testing set. After using the training data to fit a bagged-tree model, we applied the model to the testing set and obtained their predicted IMDB scores based on other factors. As can be seen in the plot below, when plotting the actual value against the forecasts, the fitted line goes through the origin with slope equals 1. This means that the fit is good and to be more accurate, the MSE is only 0.5968.

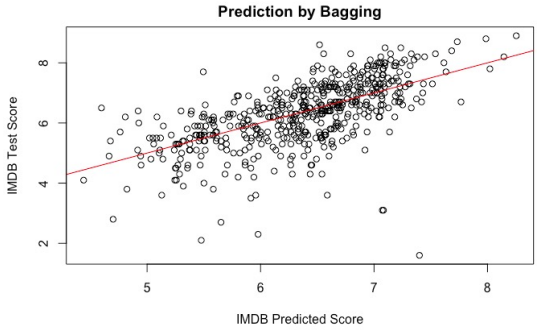

Figure 2.3.3: Variable Importance

3 Next Steps

For future analyses, we will first continue to focus on the variable selection process, and find good combinations of predictors for both the rating and the profit predictions. We will run logistics regressions of score against duration to better explore the underlying relationship. Also, we will try out more methods such as Lasso for both analysis, in order to achieve a relatively good balance between bias and variance for future models and thus avoid over/underfitting.

The models we have for now are relatively simple as we are currently in the data-exploration phase. We will try to develop more complicated and precise models for the predictions.

Another thing we will work on is to relate the Oscar Awards results to our data. We will include the new Awards dataset into the analysis, and look at questions such as whether the directors and actors that featured in the award-winning movies tend to generate more box office revenue and receive higher ratings for their other movies, and also will the award has an impact on the gross and ratings of the movies.