Luc Rieffel, Michelle Zhang, Zak Usac
MGSC 310 Final Project Report

Business Case

Our goal is to build accurate, interpretable, and applicable statistical models in order to predict a film's revenue and profit to assist producers and production companies in identifying what predictors to focus on in order to generate the most profit for a film.

Dataset Overview and Feature Transformation

The dataset we are using comes from a platform called TMDB and their dataset which we found on Kaggle contained over 3,000 rows and 8 predictor variables. After reading in the dataset to RStudio, we printed the initial summary statistics which are shown below.

```
 Movie_Name         Certification      Release_Date          Genres            Language            Budget             Revenue             Runtime     
 Length:3966        Length:3966        Length:3966        Length:3966        Length:3966        Length:3966        Length:3966        Min.   : 61.0  
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 91.0  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :102.0  
                                                                                                                                      Mean   :105.7  
                                                                                                                                      3rd Qu.:116.0  
                                                                                                                                      Max.   :248.0  
                                                                                                                                      NA's   :264    
```

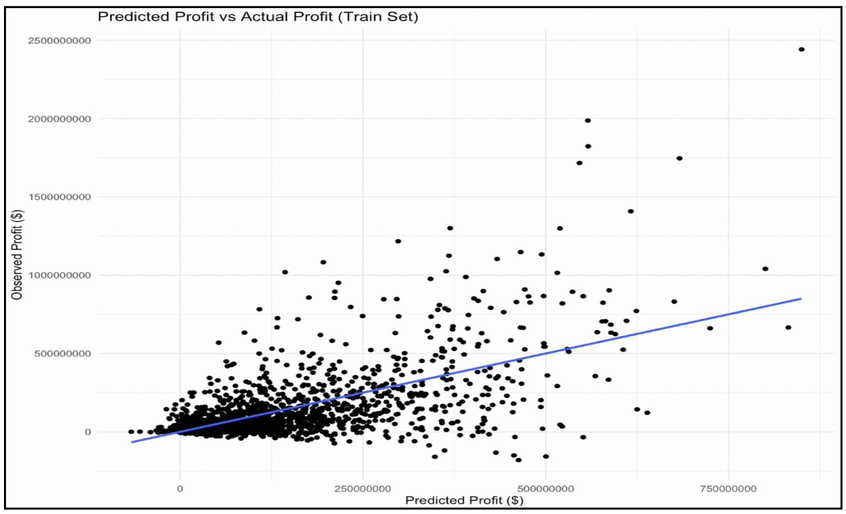
Nearly none of the columns were read in as the correct data type and it is apparent that we needed to clean the dataset and apply feature transformation techniques to make the columns usable in our models. We dropped *Movie_Name* and *Release_Date*, factored *Certification* and renamed it *Rating*, factored *Language* and applied fct_lmp, and finally typecasted *Budget* and *Revenue* to remove the commas and dollar signs. We also added a column called *profit* which was a calculation of *Revenue - Budget*. Furthermore, we applied log transformation to profit, budget, and revenue because their distributions were heavily skewed. The resulting modified dataset is shown below.

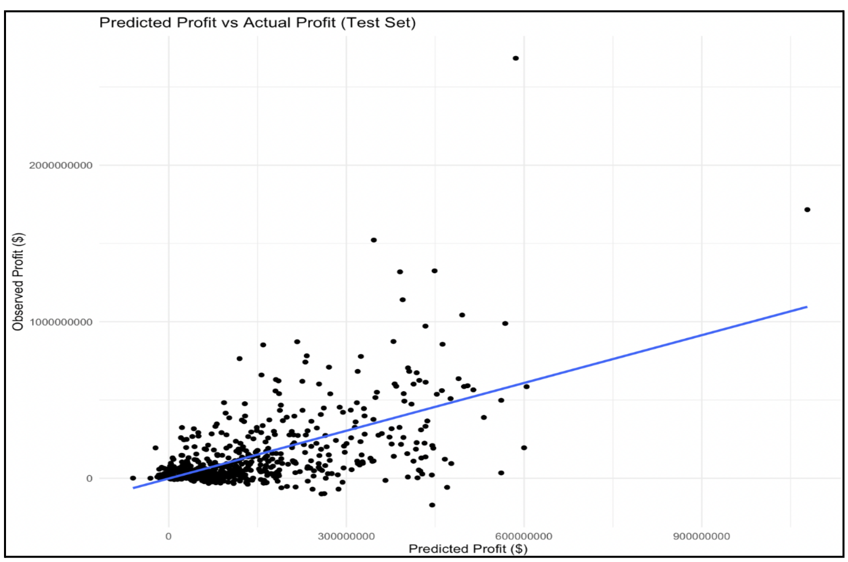
```
    Language          Budget          Revenue          Runtime          rating          profit     
 English :1687   Min.   : 9.616   Min.   :10.35   Min.   : 64.0   PG    :348   Min.   : 9.099  
 Released:   0   1st Qu.:16.677   1st Qu.:18.10   1st Qu.: 97.0   R     :471   1st Qu.:17.523  
 Spanish;:   0   Median :17.504   Median :18.85   Median :109.0   PG-13:470   Median :18.440  
 Status  :   0   Mean   :17.355   Mean   :18.76   Mean   :112.6   G     : 82   Mean   :18.259  
 Other   :   8   3rd Qu.:18.315   3rd Qu.:19.57   3rd Qu.:124.0   Other:324   3rd Qu.:19.255  
                 Max.   :19.947   Max.   :21.79   Max.   :248.0                Max.   :21.710  
```

Model 1

Luc Rieffel, Michelle Zhang, Zak Usac
MGSC 310 Final Project Report

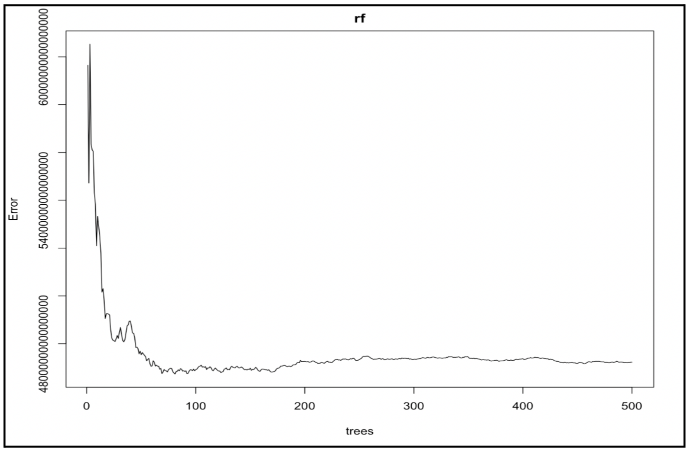
Given the nature of our business case, we will be attempting a regression task to predict *profit* based on all other variables shown in the image above. In our first model, we will be using linear regression because it is the easiest model to interpret and analyze quantitatively. After running the model, we obtained an Adjusted R Squared value of 0.3475. The significant predictors with 95% confidence turned out to be *Language(English), Budget, Runtime, rating(PG),* and *rating(G)*. For model evaluation, we obtained Train RMSE of 191,148,775 and Test RMSE of 200,914,204. This signifies very slight overfitting. The figures below show scatterplots of the Actual Profit vs. Predicted Profit.

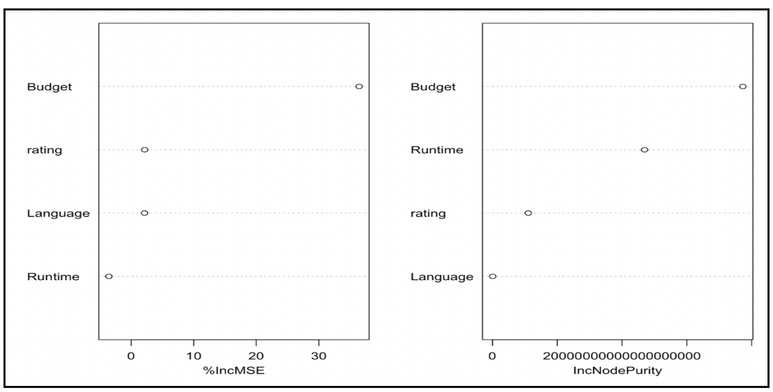


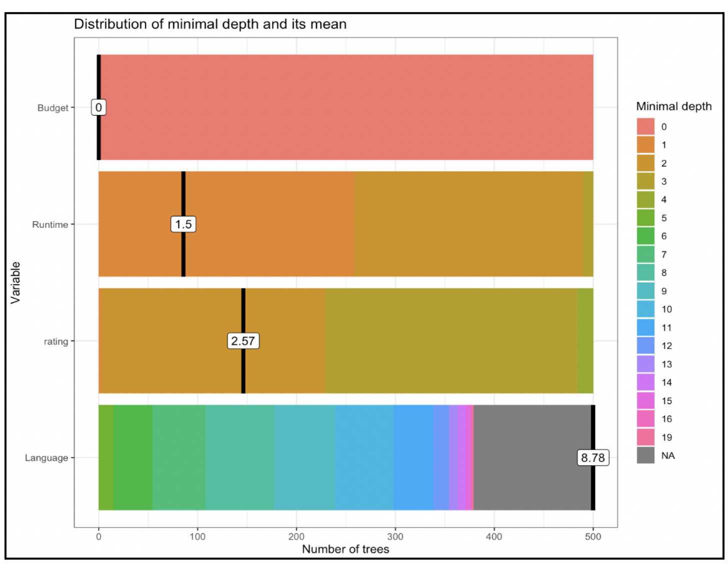


<u>Model 2</u>

For the second model, we decided to implement a Random Forest Model to experiment with a higher-complexity model on our data. The same predicted and predictor variables were used. We ran the model with *ntree* set to 500 and *mtry* set to 5 because we have 5 predictor variables. The figures below show the error plot, importance plot, and minimal depth plot distribution.







From the plots above, we can conclude that the most important predictor variable is *Budget*. This makes sense because *profit* was derived from the budget. The next two predictors are all pretty close in importance as seen in the importance plot and minimal depth distribution plot. *Language* seems to be the least influential in predicting *profit* according to our model. In terms of model evaluation, the Train RMSE came out to be 109,202,440 and the Test RMSE came out to be 199,242,934. This suggests that the model is overfit which is an issue as this model may not be good at predicting unseen data.

Conclusion

Out of these two models, we believe the better model is the linear regression model. The RMSE for both training and testing sets is low and does not have an overfitting issue like the random forest model does. The Adjusted R Squared value is on the lower side which suggests that not a lot of the data can be explained by the model. This may also imply that the relationship between *profit* and the predictors is not linear and requires a more complex model to represent. One big disadvantage of the random forest model is the overfitting issue. However, as we kept experimenting with different values of *mtry*, we realized that *mtry = 1* produced the least amount of MSE and resulted in a Train RMSE of 180,203,525 and a Test RMSE of 202,552,732. Therefore, with the random forest model, the overfitting issue can be resolved and the error values are both lower than those for the linear regression model leading us to conclude that it is the better model.

Motivation

As avid movie watchers and huge fans of the film industry, we wanted to do a statistical analysis on a movie's dataset because we wanted to see if there were any underlying patterns or trends within the data that could pose an explanation as to why some movies are more successful than others. By analyzing the dataset, we hope to be able to identify what key factors contribute and lead to a high profit for that is a primary indicator of a movie's success. Ultimately, we felt that by choosing a dataset surrounding movies, as business and data science majors, we felt that this project would allow us to combine our career interests along with our passion for films in order to gain a deeper understanding of the movie industry.