Hollywood Blockbuster Case Study Report

Priyal Sanjay Maheshwari
maheshwari.pri@northeastern.edu

Summary

Predicting a movie's success is a challenging task due to a high degree of variation in its success rate. In this case study, the goal is to develop a predictive model that will enable us to outperform the classification accuracy achieved in previous academic models. The movies are classified ranging from 1 to 9[1] categories where 1 is assumed to be a flop movie whereas 9 is assumed to be a blockbuster. I have explored the data set by performing exploratory data analysis on various features, applied NLP techniques on textual data, and implemented supervised machine learning model Decision Tree and Ensemble method Random Forest for classification. Random Forest is selected as the final model as it outperforms other models in terms of cross-validation accuracy.

Data Set

The Data Set is divided into two parts, the Training data, and the Scoring data. The training data consist of 1196 data points, 14 features, and our Target Variable Category, whereas the Scoring data consists of 91 data points and 14 features and we aim to predict the categories for this dataset. The data sets were observed to be tidy and did not consist of missing values. Training data had column 'total' which represented Total Gross Earnings (in Millions of $) of a movie that was not present in scoring data. After the initial data analysis [2], it was observed that 'Total' was highly correlated with category and each distribution of total earning could be binned to represent each category. Hence, we consider dropping the feature. Similarly, features like 'name', 'id', and 'display_name' which differ for every row will also be dropped. Scoring data had the feature 'Production_budget' which was not present on the training set and can be ignored.

Methodology

NLP: While dealing with some features like 'board_rating_reason' I came across some textual data. I used an implementation of NLPs Spacy library to clean the data and remove stop words. After observing the most frequent words and most frequent bigrams, I tried to categorize sentences using Topic Modelling with 4 topics. However, it was observed that the same terms were shared by two or more topics. Thus, using topic modeling, not much can be inferred about the data. Further, I went on to perform sentimental analysis on the reason, by calculating the polarity and subjectivity of the sentences. It was observed for most of the textual data that polarity lied between -0.5 and 0.5[3] and can be classified as the neutral sentiment. Similarly, sentences were also classified as positive and negative sentiments. We can in a way say that data consists of more facts. An additional categorical feature 'Sentiment_Score' was created based on the type of sentiment.

Data Modelling: As most of our features were categorical, I applied one-hot encoding on the datasets to generate an encoded matrix to be used as an input for the classification models. Both train and scoring data were considered while one-hot encoding, as the unique categories of each feature differed in both datasets. Decision Tree was selected as the base model and with fine-tuning of depth as 7 and sample leafs as 50, a 5-fold cross-validation accuracy of 25% was achieved. To improve the performance further, ensemble methods Random forest was used to achieve a 5-fold cross-validation accuracy of 28% [4]. Random forest was tuned to use depth as 7, leaf nodes as 50 and 1000 estimators. Random forest uses the Bagging technique to create decision trees and the average of all the outputs is considered to be the final classification. It decreases the variance in the data using Bootstrapping to produce multi-sets of original data, taking into consideration all possible outcomes and can handle our data imbalance in an effective way by tracing each possible path.

Conclusion

Finally, Random Forest was used to predict Categories of Scoring Data Set and cross-validated accuracy of 28% was achieved. Further results [5] can be seen in the Appendix.

Appendix

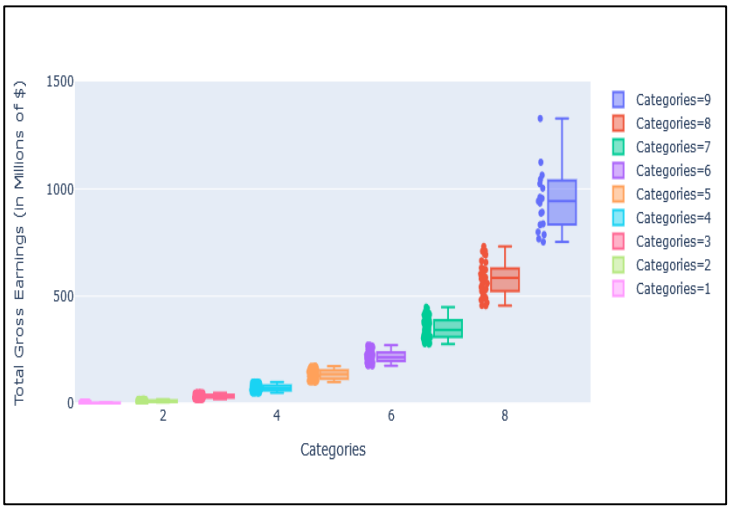

Figure 1: Distribution of Categories across Total Gross Earnings

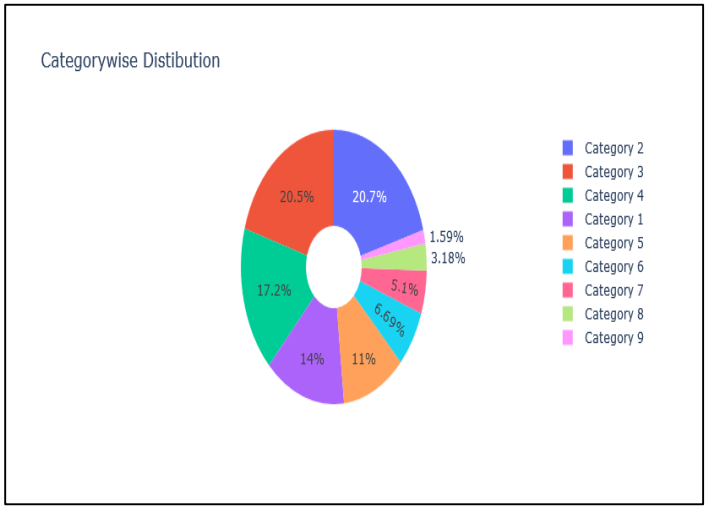

Figure 2: Percentage of each Category (Target)

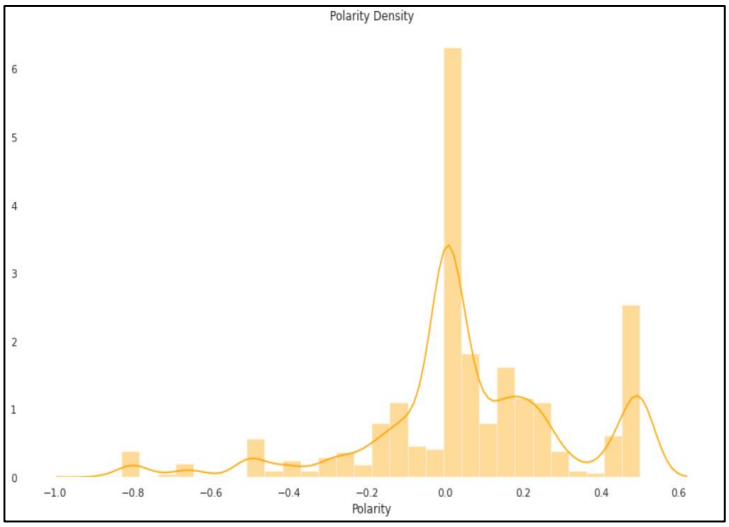

Figure 3: Polarity Density of Board Rating Reason Sentiments

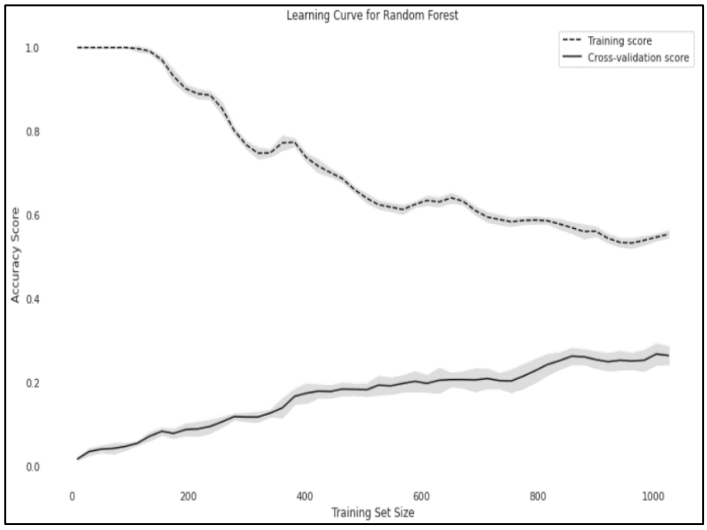

Figure 4: Learning Curve for Random Forest

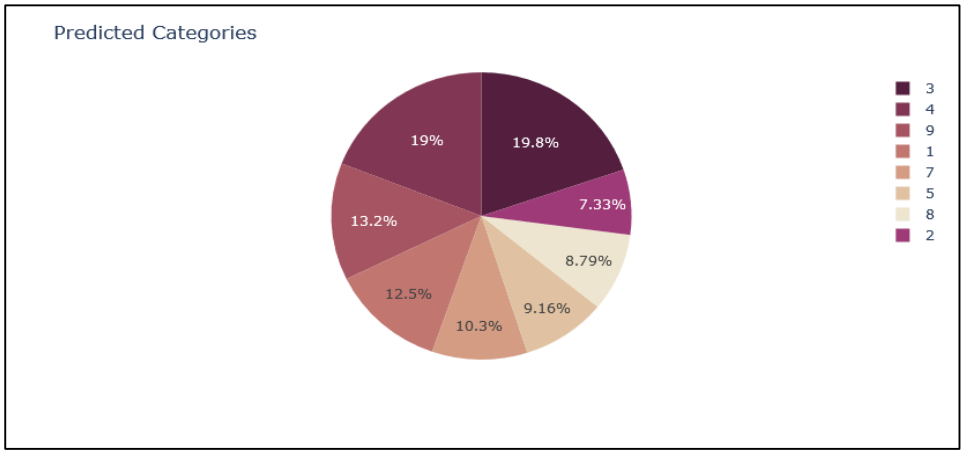

Figure 5: Predicted Categories for Scoring Data (Result)