Factors Affecting Movie Grossing and Prediction

Jingtian Zhang
jz3500@columbia.edu

Weirui Peng
wp2297@columbia.edu

Hanlun Wang
hw2839@columbia.edu

1. Introduction

The movie industry is of intense interest to both economists and the public because of the high profits and entertainment natures. Predicting the pre-release movie gross is an intriguing subject in this area since investors are very interested in making informed judgments for their investments in the movie business.As the rapid development of data-driven algorithms, machine learning models have been widely applied into various fields to cope with different tasks. Traditionally, researchers have attempted to predict movie gross based on numerical and categorical movie data from different public movie datasets, but received limited success probably because the dataset contains some meaningless data or the number of movie features is limited. As a result of which, in this project, we aim to initiate a movie dataset, which contains multiple important movie features,and develop an optimal machine learning model to predict the success and failure of the upcoming movies based on several attributes. Ultimately, we expect to create a website which could enable users to view the past movie data and make movie gross predictions based on inputting specific features.

2. Related Work

Ahmad et al. [2] have selected 36 relevant articles based on defined inclusion and exclusion criteria and thereby made a comprehensive literature review on the application of Machine Learning Techniques in the Movie Revenue Prediction field. It is illustrated that cast, number of screens, and genre, are the most widely used features in movie revenue prediction. They also found that the most popular assessment criteria were mean absolute percentage error, root-mean-square error, and average percentage hit rate, while the most used prediction techniques were multiple linear regression and support vector machines. Apart from that, Rijul and Anand [3] state that Random forest gave the best prediction accuracy for movie gross prediction task, while number of voted users, number of critics for reviews, number of Facebook likes, duration of the movie and gross collection of movie affect the IMDb score strongly.

Moreover, many other movie-relevant features have also been explored to check if these factors could benefit the prediction of movie gross. For instance, Wenbin and Steven [5] has proposed a way to forecast movie success through utilizing the quantitative news data generated by Lydia, their system for large-scale news analysis. They have well proved that the prediction models achieve better performance when using the combination of IMDB data and news data. Ibrahim et al. [1] incorporated YouTube trailer reviews into the task of movie revenue prediction prior to the movie's release, and significantly improved models' prediction accuracy with such method. Last but not least, Kyuhan et al. [4] develop and add a new feature derived from the theory of transmedia storytelling to not only increase the forecast accuracy, but also enhance the interpretability of a prediction model. Despite many researchers having brought out relatively good results on movie revenue forecast tasks, there is a potential problem for their dataset: most of them insist on using the IMDB datasets which include many incomplete movie data. Therefore, we would like to have further consideration for the dataset selection and look forward to addressing such issues to achieve better performance on movie gross prediction.

3. Methodology

In this project, we would like to collect and form a unique and qualified movie dataset containing complete and important movie features and information. Data preprocessing tasks have been involved in the process of building up our movie dataset. Then, we will look deeper on the dataset, conducting data visualization and analysis to see which factors affect movie gross more. After that, we would like to apply and train several types of machine learning models through scikit-learn to evaluate and compare their performances on movie gross prediction tasks. What is more, we also intend to make use of the ensemble approach, which has rarely been adopted in the research on predicting box-office performance, to test if such method could bring improvement on prediction accuracy. In the end, we hope to develop a website that will let people see historical movie data and forecast movie box office revenue using a variety of inputs.

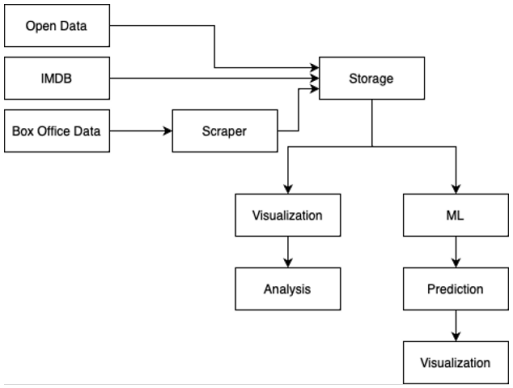

Figure 1. Workflow of the Project

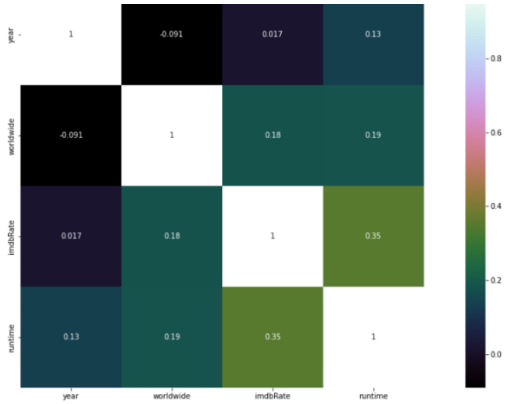

Figure 2. Heatmap of the Data

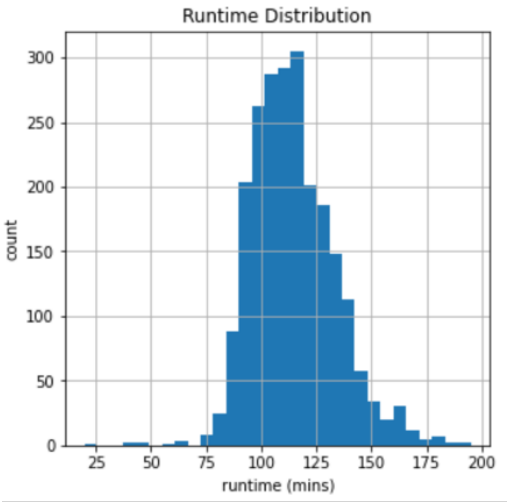

Figure 3. Runtime Distribution

3.1. Data Collection

In order to acquire a complete and clean dataset for training the prediction models, we start from crawling data from websites of those large movie organizations, for instance imdb, and boxofficemojo. For feature engineering, we capture movie features that might affect the movie gross significantly. In our case, movie imdb rate, movie genre, movie duration, movie directors, and movie release year are the features we chose to train the machine learning model. After fetching the official movie data, we start to look at augmenting the datasets, in a word, increasing the volume of our dataset. Therefore, we choose a qualified movie dataset from imdb and preprocess through deleting the incomplete movie data and picking movie data ranging from 2010 to 2022 in order to. Then, by combining the two datasets, we expand the size of our dataset. There are roughly 3000 rows of movie data within our dataset, with all of the relevant feature information clearly presented. Code: crawl.py(A.1) and combine.py(A.2).

3.2. Data Visualization

After successfully forming the training dataset, we plot out some figures to have a better understanding of the data characteristics. The figures are shown below: fig.2, fig.3, fig.4, fig.5, fig.6, fig.7. Code: visualization.ipynb(A.3).

3.3. Model Training

For the model training part, we execute the training programs using scikit-learn. We tested several prediction models: linear regression, decision tree, random forest, XGboost, Gradient Boost Model, and K-Means. Below are the figures of model performance. Code: prediction.ipynb(A.4).

3.4. Performance Comparison

Having got the performance metrics for each prediction model ,we compare each of them and obtain some conclusions for model adaptability and interpretability. table1, table2, table3, table4, table5.

Mean Absolute Error		
Rank	Model	Error
1	RandomForestRegressor	85350848.77
2	XGBRegressor	86341849.71
3	GradientBoostingRegressor	96514381.52
4	LinearRegression	100099030.61
5	SVR	104308704.00
6	KNeighborsClassifier	115191292.56
7	LogisticRegression	132274878.51

Table 1. Mean Absolute Error

In summary, we analyze the performance by using 6 error types. The LogisticRegression model has the best overall performance.

Also, we draw some pictures to illustrate the difference between the prediction value (blue line) and actual value

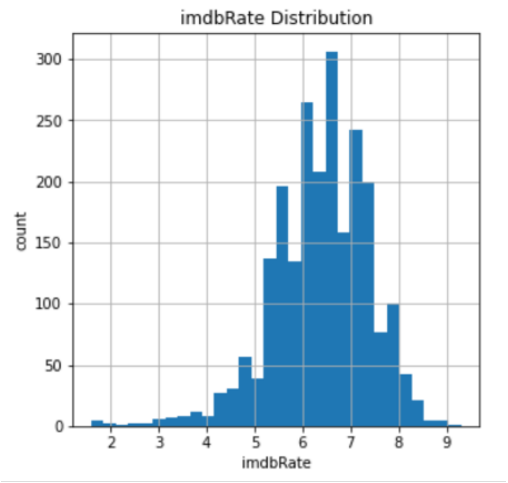

Figure 4. IMDB Rate Distribution

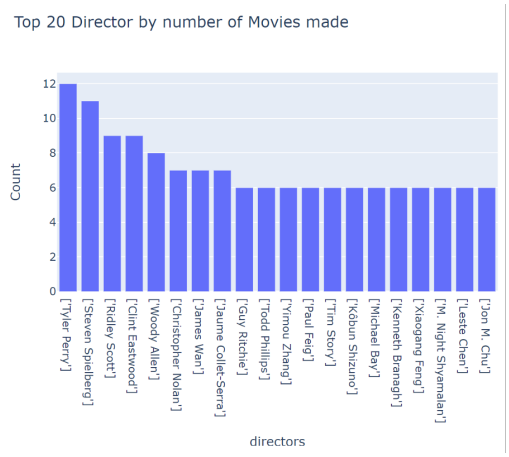

Figure 5. Top 20 Directors by number of Moviews Made

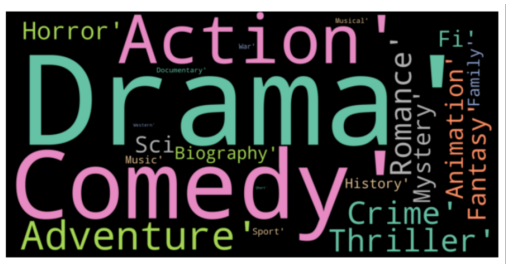

Figure 6. Tag Cloud of Movie Genres

(red line). fig.8, fig.9, fig.10, fig.11, fig.12, fig.13, fig.14.

4. Future Prospective

In the next few weeks, we will aim at using javascripts to build a website which contains data visualization functions as well as the prediction functions. We look forward to creating an effective website for users to make predictions for their desired unreleased movies and view useful figures which well describe movie characteristics.

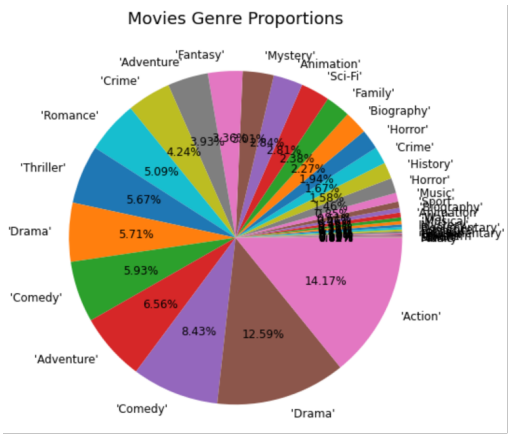

Figure 7. Pie Chart of Movie Genres

Mean Square Error		
Rank	Model	Error
1	XGBRegressor	2.50e+16
2	LinearRegression	2.69e+16
3	GradientBoostingRegressor	2.74e+16
4	RandomForestRegressor	2.76e+16
5	SVR	4.83e+16
6	KNeighborsClassifier	5.46e+16
7	LogisticRegression	5.67e+16

Table 2. Mean Square Error

Root Mean Squared Error		
Rank	Model	Error
1	XGBRegressor	158090892.52
2	LinearRegression	164048579.29
3	GradientBoostingRegressor	165552153.04
4	RandomForestRegressor	166074359.29
5	SVR	219674144.85
6	KNeighborsClassifier	233595098.56
7	LogisticRegression	238166865.86

Table 3. Root Mean Squared Error

R squared training		
Rank	Model	Error
1	XGBRegressor	0.9
2	SVR	0.9
3	RandomForestRegressor	0.9
4	LogisticRegression	0.9
5	LinearRegression	0.9
6	KNeighborsClassifier	0.9
7	GradientBoostingRegressor	0.9

Table 4. R squared training

R sqaured testing		
Rank	Model	Error
1	XGBRegressor	0.342
2	SVR	0.342
3	RandomForestRegressor	0.342
4	LogisticRegression	0.342
5	LinearRegression	0.342
6	KNeighborsClassifier	0.342
7	GradientBoostingRegressor	0.342

Table 5. R sqaured testing

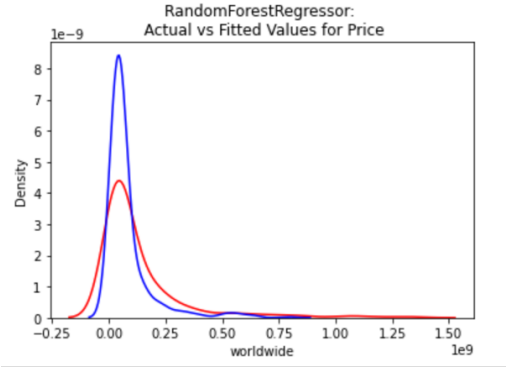

Figure 8. Random Forest Regression Actual and Fitted Values for Price

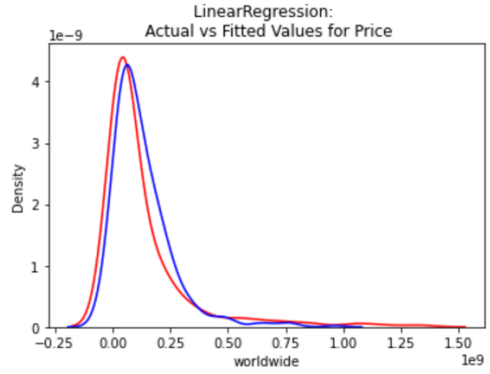

Figure 9. Linear Regression Actual and Fitted Values for Price

5. Conclusion

In conclusion, we have succeeded in constructing a new, qualified movie dataset with high volume and high variety. Apart from that, optimal prediction models have been trained and fine-tuned to predict the movie gross with high accuracy. For the next step, we would like to focus on the user interface for our project, which will further strengthen the usability for this system.

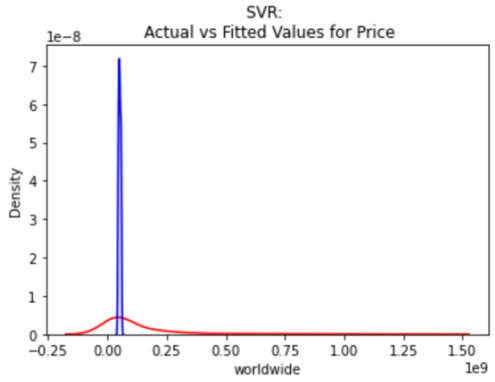

Figure 10. SVR Actual and Fitted Values for Price

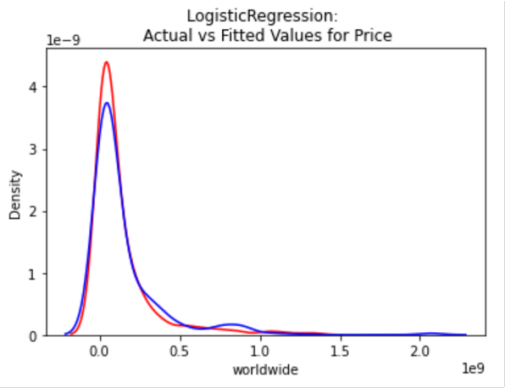

Figure 11. Logistic Regression Actual and Fitted Values for Price

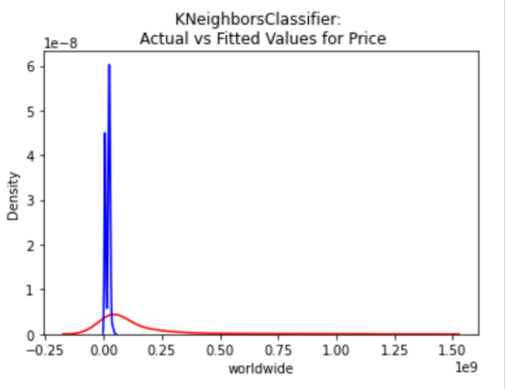

Figure 12. K Neighbors Actual and Fitted Values for Price

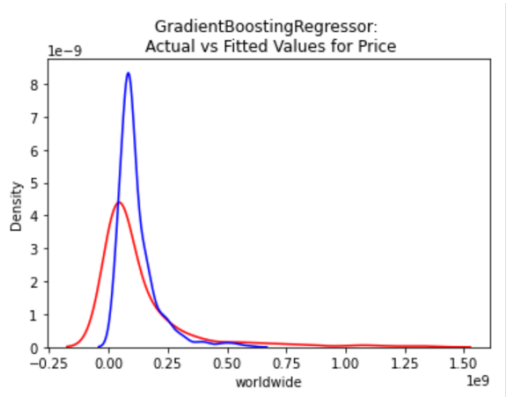

Figure 13. Gradient Boosting Actual and Fitted Values for Price

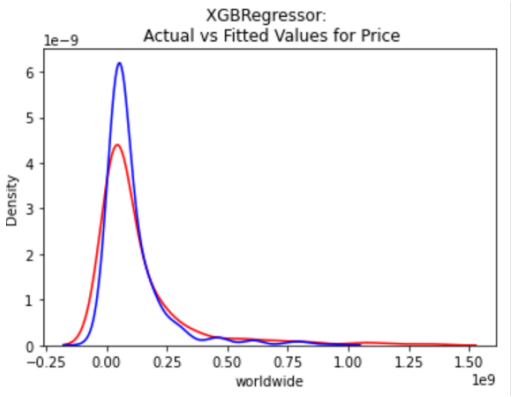

Figure 14. XGB Regression Actual and Fitted Values for Price

References

[1] Ibrahim Said Ahmad, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. Movie revenue prediction based on purchase intention mining using youtube trailer reviews. *Information Processing & Management*, 57(5):102278, 2020. 1

[2] Ibrahim Said Ahmad, Azuraliza Abu Bakar, Mohd Ridzwan Yaakub, and Shamsuddeen Hassan Muhammad. A survey on machine learning techniques in movie revenue prediction. *SN Computer Science*, 1(4):1–14, 2020. 1

[3] Rijul Dhir and Anand Raj. Movie success prediction using machine learning algorithms and their comparison. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 385–390. IEEE, 2018. 1

[4] Kyuhan Lee, Jinsoo Park, Iljoo Kim, and Youngseok Choi. Predicting movie success with machine learning techniques: ways to improve accuracy. *Information Systems Frontiers*, 20(3):577–588, 2018. 1

[5] Wenbin Zhang and Steven Skiena. Improving movie gross prediction through news analysis. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 301–304. IEEE, 2009. 1

A. Codes

A.1. crawl.py

https://github.com/HanlunWang/EECS6893_Final_Project/blob/main/crawl.py

A.2. combine.py

https://github.com/HanlunWang/EECS6893_Final_Project/blob/main/combine.py

A.3. visualization.ipynb

https://www.kaggle.com/code/zjt1027/visualization?scriptVersionId=112880368

A.4. prediction.ipynb

https://www.kaggle.com/code/zjt1027/prediction-model?scriptVersionId=112879881