CSE343/ECE343: MACHINE LEARNING PROJECT FINAL REPORT (Monsoon 2021)

BOX OFFICE PREDICTION

Utkarsh Dubey 2019213 | Robin Garg 2019092 | Bhaskar Gupta 2019237 | Krishna Jalan 2021001

ABSTRACT

The motion picture industry is a multibillion-dollar business, and there is a massive amount of data related to movies available over the internet. Predicting society's reaction to a new product in the sense of popularity and adaption rate has become an emerging field of data analysis. Box office revenue prediction is an important problem in the film industry that governs financial decisions made by producers and investors. Predicting the box office profits of a movie prior to its worldwide release is a significant but difficult problem that needs an advance of intelligence. Github repository for the code is Robin-garg23/ML_Project (github.com)

1. INTRODUCTION

The definition of success of a film is relative, some videos are called successful based on their worldwide gross income, and some movies may not shine in the business part but can be called successful for good critics' review and popularity. Here we are considering a movie's box office success based on its profit only. This project proposes a decision support system for the movie investment sector using machine learning techniques. We have collected valuable data regarding movies and have tried to find correlations between certain features that can have impacts on the total revenue generated by the movie. Proper graph analysis of these features' dependencies has also been undertaken. Once relevant features were extracted, we applied different machine learning approaches namely linear regression, decision tree and random forest so far in order to minimize the RMSE, hence finding a model that fits the collected dataset well.

2. LITERATURE SURVEY

[1]Predicting Box Office Revenue for Movies by Matt Vitelli used two different models to predict Box office revenue, first was a linear classifier with a softmax activation function and the second was two layers neural network with tanh activation function. The author used given features as well as extracted his own features to predict the revenue. By extracting new features he was able to predict data with more accuracy.

[2]A Machine Learning Approach to Predict Movie Box-Office Success used sentiment analysis (Microsoft Azure text analysis of IMDB reviews), Support Vector Machine, Neural network analysis to predict revenue. They found pre and post-release, both the features are important for prediction. Budget, number of screens where a movie is released dominated. Finally figuring out that budget, IMDb votes and no. of screens are the most important features.

3. DATASET DESCRIPTION

We used the official API of TMBd to fetch the details of movies to form our dataset. Our dataset contains over 10,000 movies having details of each movie like title, genre, budget, cast, crew, Release_Date, etc. There were around 22 features we were able to fetch. The target variable for our study is the revenue generated by the movie.

3.1 Data Cleaning

Initially, there were around 10,000 rows in our dataset initially. There was much discontinuity and wrong entries in the dataset collected by us, as the budget of many movies was set to zero, huge mismatches between budget and revenue, many ratings of the movie were zero, some duplicate entries were also present. We removed all such rows and created cleaned data for preprocessing.

3.2 Data Preprocessing and Feature Selection

1. We tried finding relations between the collected features and revenue.
2. Many features were present in a grouped format like production_countries, languages. These groupings needed to be separated accordingly and then further coded to make it a proper input for the regression problems.
3. We found a high correlation of budget and popularity and a decent correlation of runtime with revenue**.**

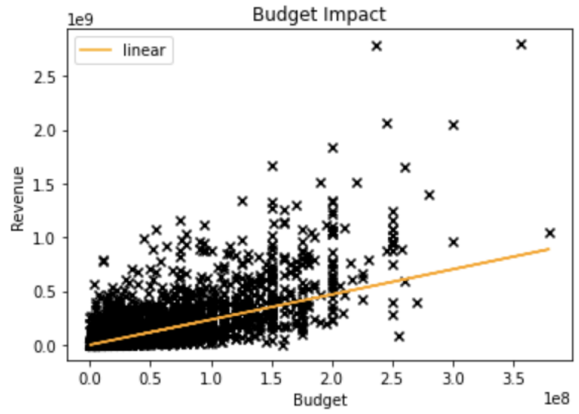

Fig 1: Budget impact on revenue

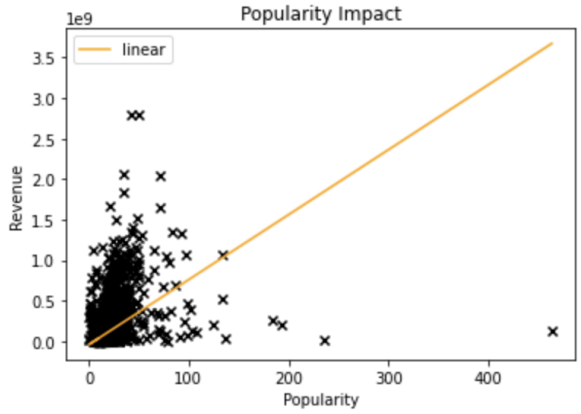

Fig 2: Popularity impact on revenue

4. Further feature extraction and reduction -

   a. Selected features to work with are -

Budget
Popularity
Runtime
Origin Language
Genre
Production Companies
Countries
Cast
Crew
Status
Title
Rating
Part of Collection(binary)

Table 1: Features for prediction

For all grouped data, we decided to encode them binarily based on their presence for the given movie.

   b. Found the set of production companies, languages, production countries, cast, directors and then hot encoded for each movie.
   c. For our gathered dataset we found -
      i. 5150 unique production companies
      ii. 32 unique languages
      iii. 82 unique origin countries
      iv. 81384 unique cast members
5. Due to the high number of production companies, cast members, and crew members, we decided not to include these features due to computational shortcomings.
6. Further preprocessing - only movies with status released were present in the dataset for better training and testing.
7. We used Normalisation and Standardization to scale the data around the mean.
8. We incorporated another feature of the cast using the top 10 casts by revenue of their movies.

4. Model Details and Methodology

Our objective is to find a model that can most precisely and effectively map the features that play a vital role with appropriate coefficients to predict the revenue generated by the movie. We are training different models available to us and testing them to get the best among them.

4.1 Linear Regression

Assuming a linear relationship between the input and single output variables “Revenue”), we applied linear regression on the dataset with 66-33 train-test data split. We have trained the model with 2857 input samples with 135 features and tested it with 1539 test samples. Applied following regularisation:

- Ridge Regression (alpha= .1)
- Lasso Regression (alpha= 10)

4.2 Decision Tree

Applied Decision Tree Regressor on the given dataset and used 66-33 train-test split. For the hyperparameter, we used max_depth=3 to avoid overfitting of the dataset. We used the decision tree as a baseline for future comparison with different models.

Fig 3: Training iter vs r2 Score (Decision Tree)

4.3 Random Forest

Applied Random Forest Regressor on the given dataset with 66-33 train-test split. With hyperparameter max_depth=5 and criteria=’mse’ to reduce the overfitting problem and achieve better results.

4.4 KNN Regressor

Applied K Neighbors Regressor on the given dataset and used a 66-33 train-test split. For the hyperparameter, we used n_neighbors=10 to best fit the dataset (obtained by testing multiple values of n).

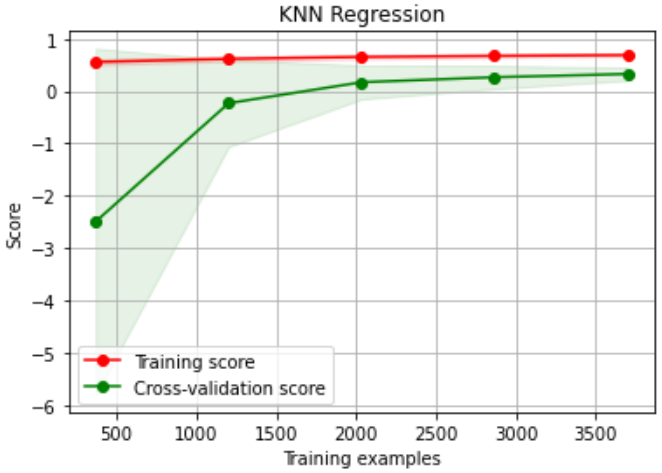

Fig 4: Training iter vs r2 Score (KNN)

4.5 MLP Regressor

Applied MLP Regressor with one input-layer, two hidden-layers [147, 74] (number of input feature, mean of input feature and output feature) and one output layer with a number of neurons = 1 (regression model) with activation function as “linear”. Model is trained for 500 epochs.

5. Results and Analysis

R2 score, RMSE and MAE are used as metrics to determine the performance of each model. And all the metrics corresponding to each model are as follows.

5.1 Linear Regression

The linear regression model fits the dataset with the testing score of 0.71 and the training score of 0.74; after applying the regularisation on techniques such as ridge and lasso, the model started to underfit the data with 0.70 and 0.60 r2 scores for test and training sample. RMSE and MAE scores obtained are (106268722, 60120427).

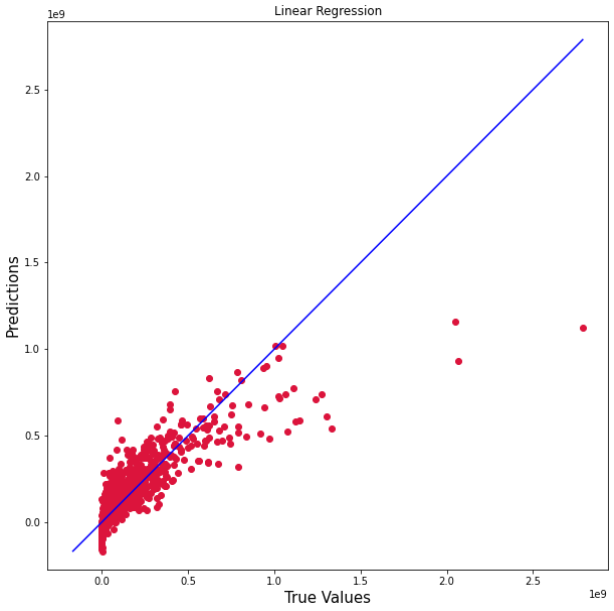

Fig 5: Prediction on Linear Regression

5.1.1 Ridge Regression

Ridge Regression is used with alpha=0.1; regularisation is causing the model to underfit with an r2 score of 0.70 and 0.60 for test and training, respectively.

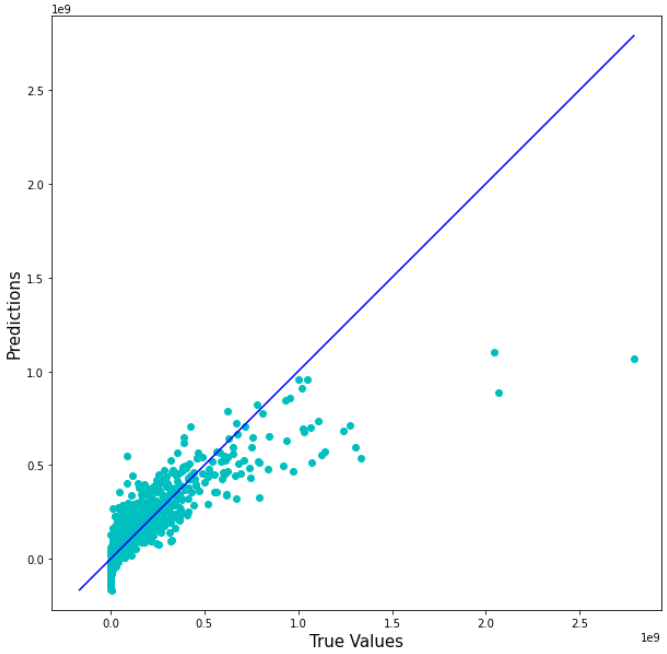

Fig 6: Prediction on Ridge Regression

5.1.2 Lasso Regression

We used Lasso Regression with alpha=10, causing the model to underfit similar to Ridge with an r2 score of 0.70 and 0.60 for test and training, respectively.

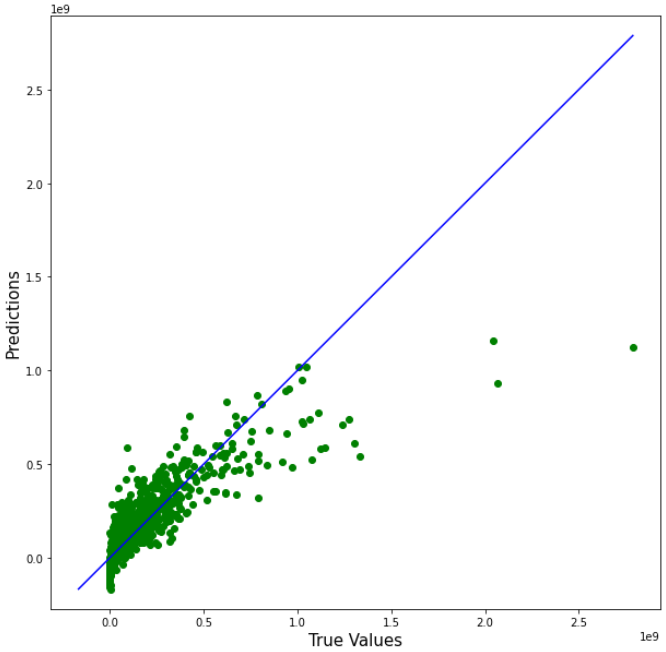

Fig 7: Prediction on Lasso Regression

5.2 Decision Tree

We used the decision tree as a baseline for future comparison with different models. The decision tree scored 0.95 and 0.96 for testing and training dataset resp. With RMSE, MAE score: (4207113, 22058687)

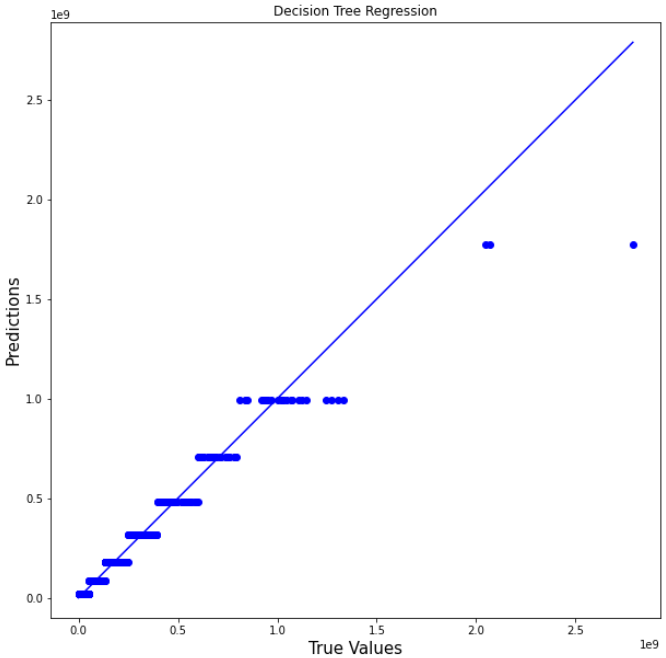

Fig 8: Prediction on Decision Tree

5.3 Random Forest

For RandomForest, we get the best r2 score compared to all the other models, i.e. 0.98 and 0.99 for test and train resp.

RMSE, MAE scores are as follows: (27441463, 3361622).

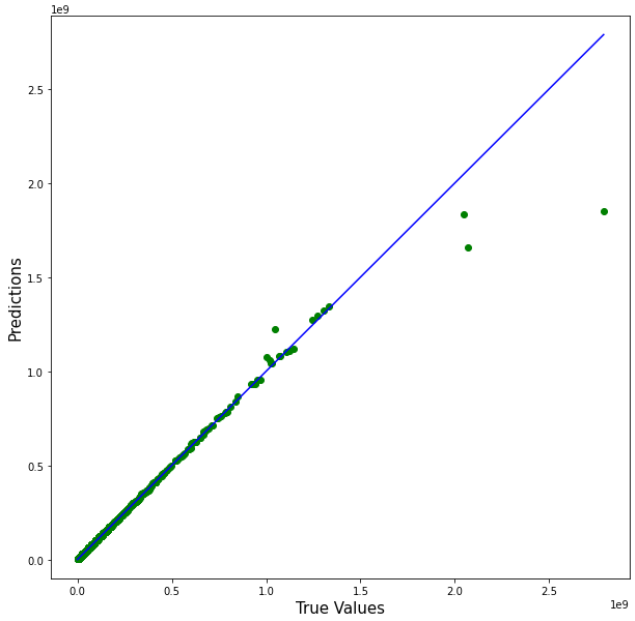

Fig 9: Prediction on Random Forest

5.4 KNN Regressor

KNN with ten neighbours results in an R2 score of 0.60 and 0.70, which shows that the model is overfitting the data. RMSE and MAE for the KNN is (126378561, 64130231) resp.

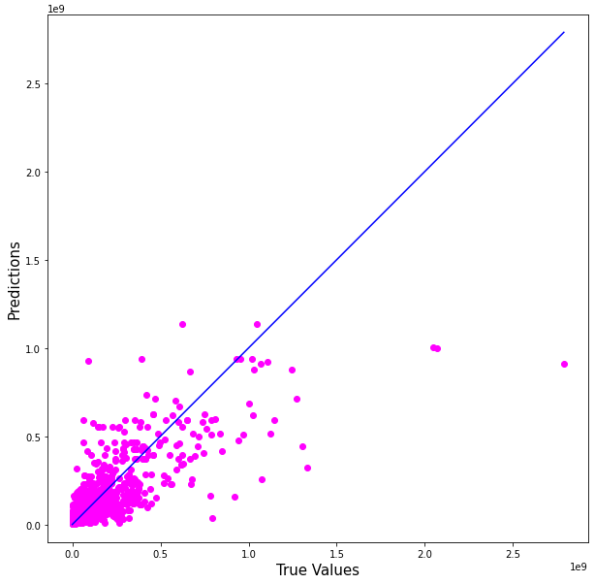

Fig 10: Prediction on KNN Regressor

5.5 MLP Regressor

MLP yields the r2 score of 0.64 and 0.66 for test and training, indicating that the model can not fit the data with the given configuration and needs more tuning and training. RMSE and MAE score for the model is (11668440, 66859636).

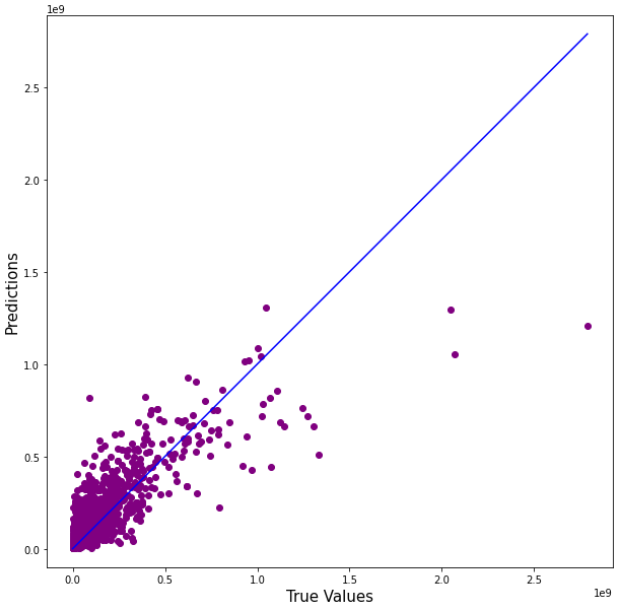

Fig 11: Prediction on MLP Regressor

6. Conclusion

We tried predicting the Box Office Revenue before the movie's release using the dataset of the movie presented to us. We found a positive correlation of revenue generated by a movie with the budget, actors' popularity, and runtime.

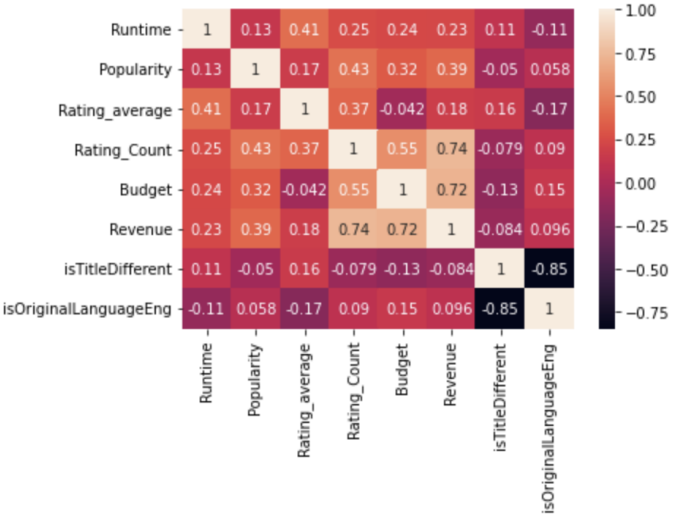

Fig 12: Heat map of features

So far, the best model comes to be a random forest with the r2 score of 0.98 followed by Decision Tree regression with an r2 score of 0.95, Linear Regression with 0.71, MLP with 0.62 and last but not the least, KNN with 0.60 r2 scores.

7. Future Work

Future work can include applying the Deep Learning model to the given dataset for better results. We currently dropped features like Production Company and crew members due to a large number of companies/members in the dataset, and we have only considered the top 10 cast members contributing to the revenue. We can include those features and expand our model to run on more features. The features like Title, Overview got dropped due to textual data; we can apply NLP to process the data. Finally, the dataset was limited, and we further dropped many rows during preprocessing; we can include more movies to improve the training of the current model.

8. Team Contribution

1. Krishna Jalan - Reducing noise from data, Analysis of results/models using different evaluation methods

2. Robin Garg - Finding and analysing Dataset, Literature review, Data Preprocessing of grouped data, Model Selection.

3. Utkarsh Dubey - Data Preprocessing and Data scraping, walking over different available datasets, analysing data on graphs, Parameter Selection for models.

4. Bhaskar Gupta - Data Extraction from TMDb API, Data Preprocessing, Verification of Accuracy/Parameterization.

9. References

[1] Predicting Box Office Revenue for Movies Link

[2] A Machine Learning Approach to Predict Movie Box-Office Success Link

[3] TMDb API for Dataset Link2021001