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

BOX OFFICE PREDICTION

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

1. 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.

2. 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.

3. 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.

4. 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.

4.1 Data Cleaning

Initially, there were around 10,000 rows in our dataset initially. There were many 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.

4.2 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 in order 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.

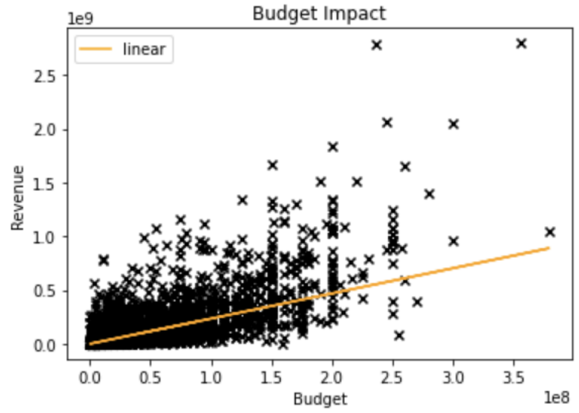

Fig 1: Budget impact on revenue

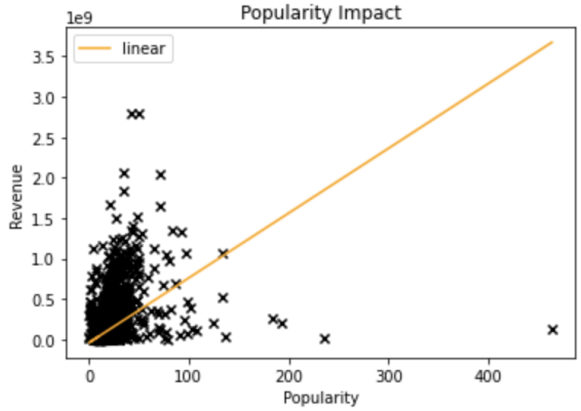

Fig 2: Popularity impact on revenue

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, due to computational shortcomings, we decided not to include these features
6. Further preprocessing - only movies with status released were present in the dataset for better training and testing.
7. Normalization and Standardization are used to scale the data around the mean.

5. Model Details and Methodology

Our objective is to find a model which can most precisely and effectively map the features which play a vital role with appropriate coefficients in order to predict the revenue generated by the movie.

5.1 Linear Regression

Assuming a linear relationship between the input variable and single output variable("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.

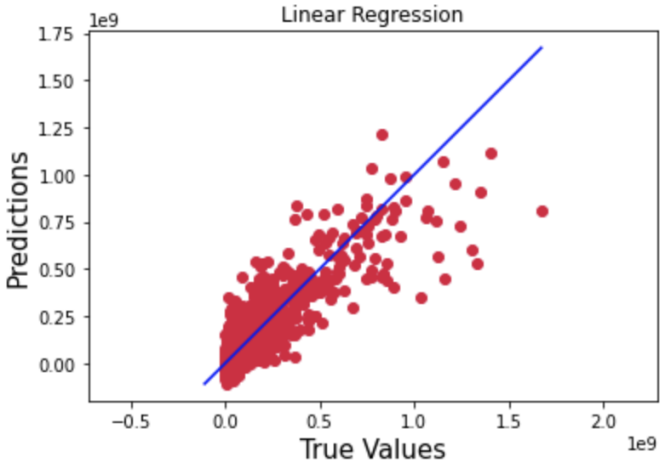

Fig 3: Prediction on Linear Regression

5.2 Decision Tree

Applied Decision Tree Regressor on the given dataset and used 66-33 train-test split. Decision tree is used as a baseline for future comparison with different models.

Fig 4: Prediction on Decision Tree

5.3 Random Forest

Applied Random Forest Regressor on the given dataset with 66-33 train-test split. With max depth of 5 to reduce the overfitting problem to achieve better results compared to the decision tree regressor model.

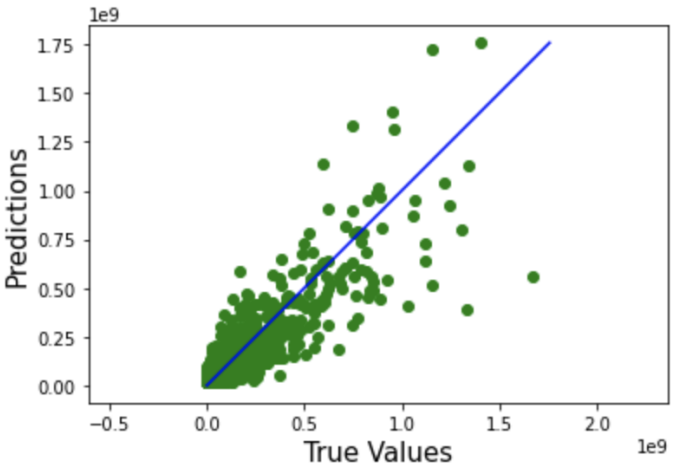

Fig 5: Prediction on Random Forest

6. Results and Analysis

The linear regression model under-fitted the dataset as the testing score is 0.73 and the training score is 0.70. The decision Tree regressor overfitted the whole dataset as the testing score is 0.51 and the training score is 1.0. Last but not least the best fit model comes to be random forest with 0.74 with testing data and 0.84 with training data. Some features have a positive strong

correlation with revenue like “rating_count” and “budget” respectively and some features show a negative correlation with revenue like “IsTitleDifferent” (is multiple titles available for a movie depending upon geographic location).

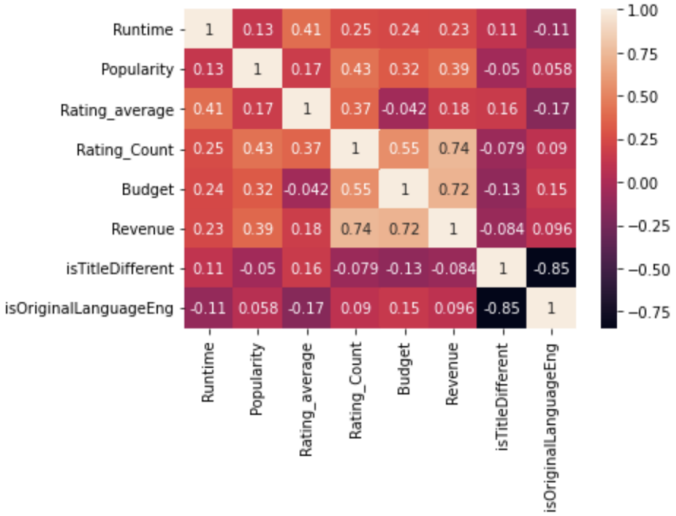

Fig 6: Heat map of features

7. Conclusion

So far the best model so far comes to be random forest with the r2 score of 0.74 followed by linear regression with an r2 score of 0.73 then followed by a decision tree with an r2 score of 0.51 with RMSE, MSE of (100858863, 62278929), (135634568, 77226702) and (98377092, 56722760) respectively. Need to apply hyperparameter tuning to avoid overfitting and underfit to improve the performance measure of each model.

8. Work Left

I. Trying more regression algorithms like Lasso regression, Ridge regression, KNN, SVM.
II. Parameter tuning and refinement for better results.
III. Find a way to incorporate cast and crew member details in our features list.

9. 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, analysis of data on graphs, Parameter Selection for models.

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

10. 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 Link