CSE343: MACHINE LEARNING INTERIM PROJECT REPORT(Winter 2022)

BOX OFFICE PREDICTION

Neha Goel
IIIT, Delhi
neha19066@iiitd.ac.in

Raghav Nakra
IIIT, Delhi
raghav19083@iiitd.ac.in

Ramit Gupta
IIIT, Delhi
ramit19086@iiitd.ac.in

1. INTRODUCTION

The definition of success of a film is relative, some are called successful based on their 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 minimise the RMSE, hence finding a model that fits the collected dataset well. We have tried several hyperparameter tuned models to finally reach our results.

2. RELATED WORK

2.1. *A Survey on Prediction of Movie's Box Office*

Collection Using Social Media Study to list the factors upon which the success of a movie can depend.

2.2. *Predicting Box Office Revenue for Movies*

The study provides a computational model for movie revenue prediction using a combination of features extracted from movie database metadata.

2.3. *A Machine Learning Approach to Predict Movie Box-Office Success*

The study predicts an approximate success rate of a movie based on its profitability by analysing historical data from different sources like IMDb, Rotten Tomatoes. depending on the analysed trend.

3. DATASET & EVALUATION

We used this Kaggle dataset which contains details of over 45,000 films released before July 2017. It contains 7 files, containing movie metadata (45K+ entries containing features like *adult, belongs to collection, budget, genres, homepage, id, imdb id, original language, original title, overview, popularity, poster path, production companies, production countries, release date, revenue, runtime, spoken languages, status, tagline, title, video, vote average, vote count for each film.), credits (movie ids and corresponding crew members), keywords (movie ids and corresponding words used to describe their plot), ratings (subset of ratings for movies)*. There are multiple attributes in the files which are common to more than one file, we used pandas inner join on data frames to access relevant details. We have used R^2 score and RMSE to evaluate our models. RMSE gives the root of mean squared error between actual and predicted value. R^2 score is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

4. METHODOLOGY

4.1. Exploratory Data Analysis:

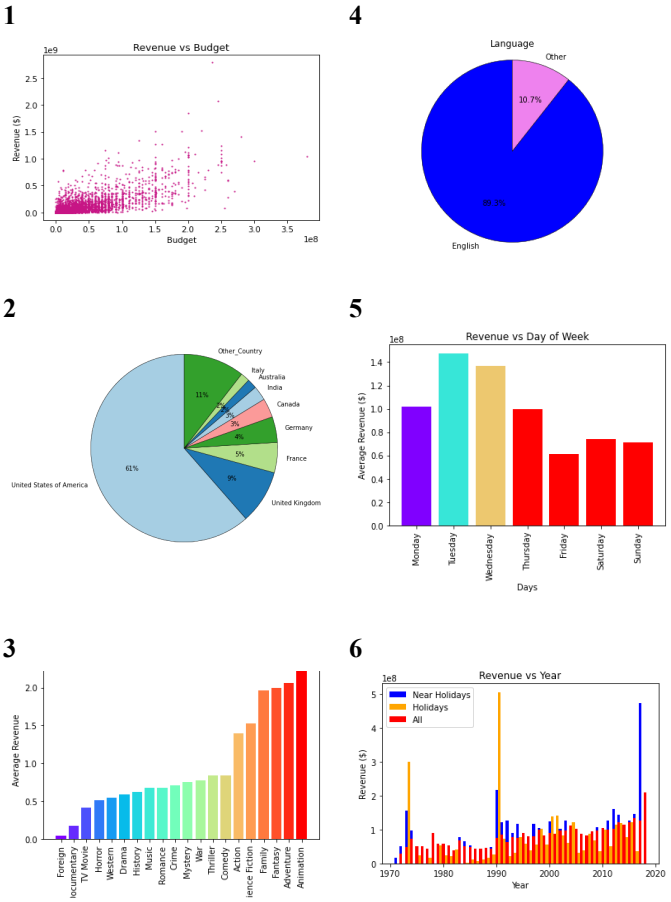
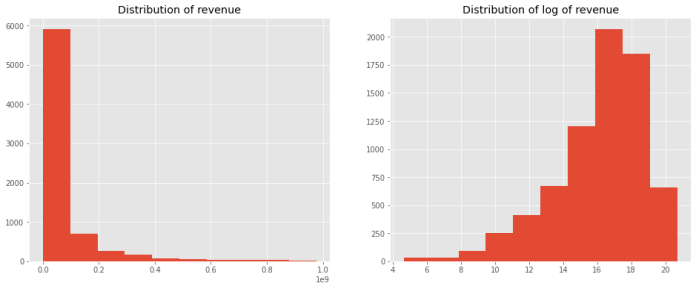


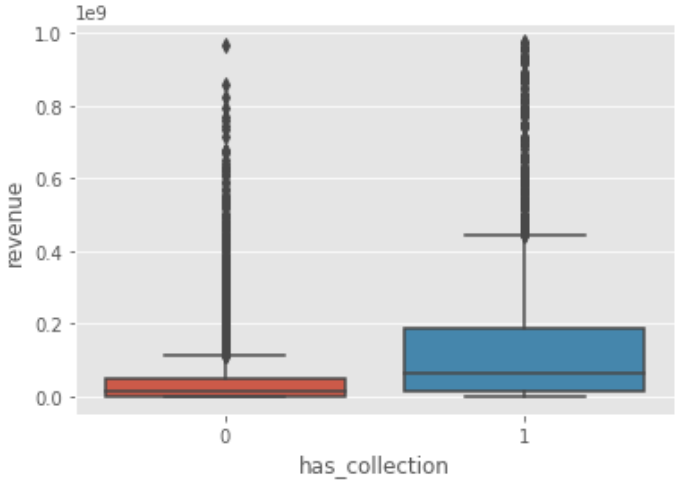
Figure 2: Data analysis done to understand dataset better

4.2. Pre-processing for key features:

a) For revenue: Revenue can be seen to be highly concentrated in a small range, however, taking the log of revenue gives us a better distribution.



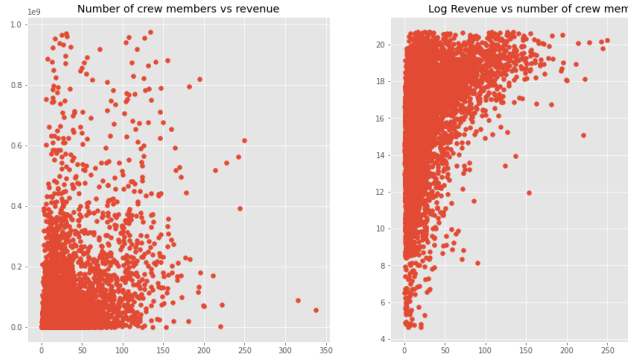
b) For collections: Most of the films didn't belong to any collection, so we removed that column and instead added two columns, one for whether the film has a collection, and another for the name of the collection.



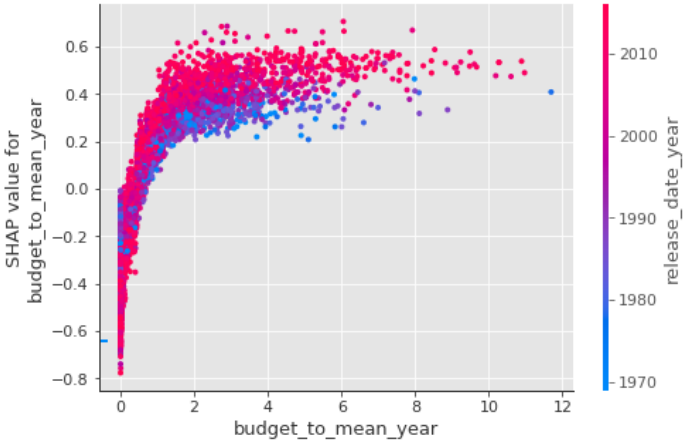
c) For genre: We took the top 15 genres in 15 columns and then one hot encoded the columns on the basis of which genre a film belongs to.

d) For Keywords: We filtered the most commonly occurring keywords and took their count.

e) For cast and crew: Instead of just giving cast and crew unique ids, we also took into account their gender, the characters played by cast members, the department of crew members, the total size of cast and crew etc.



We created some features of our own, for example, budget to year ratio to account for growing inflation, and these were seen to have a significant impact on the revenue output.



4.3. Model training and evaluation:

We followed 10 fold cross validation to generate average RMSE on training and validation set.

First, we tried the LGBMRegressor: We ran it for all ten folds, with 200 early_stopping_rounds, and in most of the cases, we had the model stop training early to minimise training and validation error, and prevent overfitting. The final RMSE obtained on the validation set was 57005177.

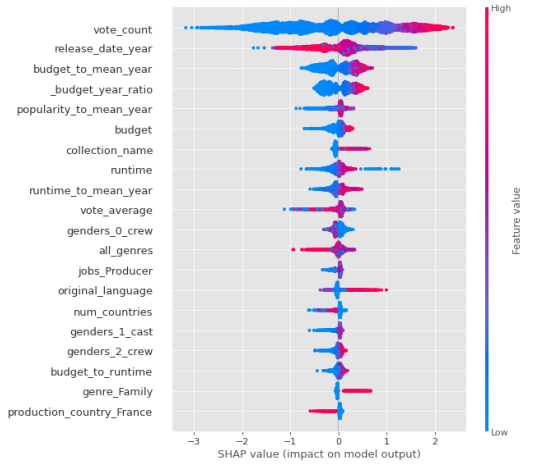
We tried XGBoost which uses regression trees as a base learner, with max depth as 7 to avoid overfitting. We again had 20000 epochs and for almost all folds the model stopped early according to best validation rmse. The final RMSE on the validation set came out to be 56947589.

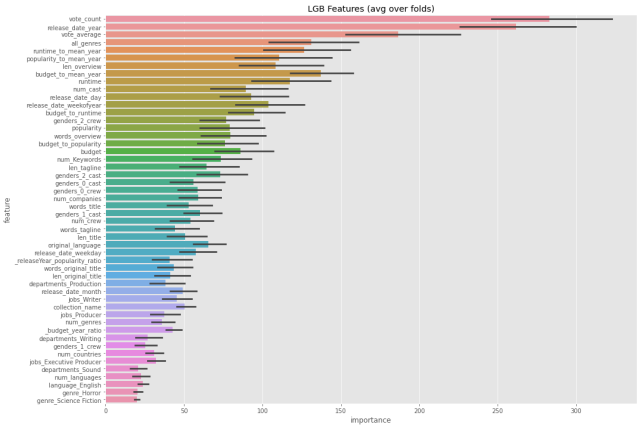
We also tried CatBoostRegression with 0.002 learning rate for 20000 epochs. The final RMSE on the validation set was 60061557.

5. RESULTS & ANALYSIS

The following observations have been made.

1. Budget of a movie does not independently determine revenue as evident from the movies with similar budgets generating various revenues.
2. The pi-chart helps in recognizing countries that have produced a significant number of movies. We observed most movies were produced by 8 countries.
3. On average, animated movies make the highest average revenue while foreign movies make least revenue.
4. English dominates distribution of language with roughly 89.3% of movies in English.
5. Releases made early in the week(from Monday) tend to earn more revenue.
6. Movies released during holidays after the 2000s tend to make more revenue.





7. The importance map shows how features like release date, year, vote count, and even artificially generated features like budget to year ratio had a significant impact on the output.

The following results were obtained:

Model	R2 score	RMSE
Linear Regression	0.6494	98579406
Quadratic Regression	0.5191	115455181
Cubic Regression	0.4951	127107421
Regression with degree 4	-0.0633	167279329
MLP	0.5765	108382092
Neural Network	0.5764	115585076
PCA	0.6397	99928132
KNN Regressor	0.5787	113613112
RNN Regressor	0.5791	195982681
LGBM Regressor	–	57005177
CATBoost Regressor	–	60061557
XGB Regressor	–	56947589

The RMSE of 56947589 is equivalent to 8.7% error over the mean revenue (as opposed to the 48% error in the interim project phase). The benchmark accuracy of the 2nd Paper was 61.7%, since the following paper converted the problem into the classification one, we can't fully compare it with our results.

Earlier, before trying the boosting, none of the models, regression or neural networks seem to give an optimal rmse, even though the R2 score is decent for linear regression and PCA. The best performance of PCA indicates that we need better feature selection, or possibly newer features.
On trying boosting and doing significant changes in selecting the features, there was a significant improvement in the accuracy of the models and the RMSE score showed a drastic decrease. LGBM, XGB and CATBoost methods were tried. XGB Regressor gave the best results, by bringing down the RMSE score 10 times.

6. CONTRIBUTIONS

6.1. Deliverables

Deliverables promised during proposal are as follows:

TASKS	TEAM MEMBERS
Data scraping and designing	Raghav and Ramit
Constructing dataset and transformation	Ramit and Neha
Data Visualisation	Raghav and Neha
Feature Extraction	Raghav and Ramit
Researching and Training Models	Ramit, and Neha
Analysis and Performance of Models	Raghav and Neha
Hyperparameter tuning	Raghav, Ramit, and Neha
Report Writing	Raghav, Ramit, and Neha

All mentioned tasks were achieved by the corresponding team member.

6.2. References and Citations

The models and techniques used were taken from many sources. The links to the sources are as follows:

1. XGBoosting and Importance: https://www.youtube.com/watch?v=GrJP9FLV3FE
2. Boosting and Blending in case of regression analysis. https://www.kaggle.com/code/kamalchhirang/eda-feature-engineering-lgb-xgb-cat/notebook
3. EDA, Feature Engineering, and model interpretation https://www.kaggle.com/code/artgor/eda-feature-engineering-and-model-interpretation
4. Movie Analysis and Comparative Modelling https://github.com/IndraP24/Sem-3-Project-TMDb-Movie-Analysis-and-Modeling
5. EDA and feature selection https://github.com/sibeltan/disney_movies_analysis

6.3. Individual Contributions

Ramit - Performed Boosting, Blending and Feature Selection which helped in further improving the accuracy of the model and led to better revenue prediction.
Raghav - Performed EDA, Boosting and Feature Selection which lead to removal of redundant features and improving the model. Relations between different features were also observed.
Neha - Performed Hyper parameters tuning for almost every model and method followed, setting params in Boosting and Feature Selection. This helped in improving the results in a better way and further contributing to better performance.

The EDA for our project was very large and involved various tasks, the files for the individual components can be found via the given link.