ML Project – TMDB Box Office Prediction – Kaggle Competition

Talha Hanif Butt
Department of Computer Science
FAST-NUCES
Lahore, Pakistan
L18-1864

I. OVERVIEW

In a world... where movies made an estimated $41.7 billion in 2018, the film industry is more popular than ever. But what movies make the most money at the box office? How much does a director matter? Or the budget? For some movies, its You had me at Hello. For others, the trailer falls short of expectations and you think What we have here is a failure to communicate. In this competition, metadata on over 7,000 past films has been provided from The Movie Database to try and predict their overall worldwide box office revenue. Data points provided include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries.

II. DATASET

In this dataset, we are provided with 7398 movies and a variety of metadata obtained from The Movie Database (TMDB). Movies are labeled with id. Data points include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries.

We are predicting the worldwide revenue for 4398 movies in the test file.

Many movies are remade over the years, therefore it may seem like multiple instance of a movie may appear in the data, however they are different and should be considered separate movies. In addition, some movies may share a title, but be entirely unrelated.

E.g. The Karate Kid (id: 5266) was released in 1986, while a clearly (or maybe just subjectively) inferior remake (id: 1987) was released in 2010. Also, while the Frozen (id: 5295) released by Disney in 2013 may be the household name, dont forget about the less-popular Frozen (id: 139) released three years earlier about skiers who are stranded on a chairlift...

III. EVALUATION

Submissions are evaluated on Root-Mean-Squared-Logarithmic-Error (RMSLE) between the predicted value and the actual revenue. Logs are taken to not overweight blockbuster revenue movies.

A. Submission File Format

The file should contain a header and have the following format:

id,revenue
1461,1000000
1462,50000
1463,800000000
etc.

IV. LITERATURE REVIEW

Many different solutions to the problem of Box Office Prediction have been tried including the analysis of Wikipedia Activity [6] to build a minimalistic predictive model for the financial success of movies based on collective activity data of online users. [4] proposes a decision support system to aid movie investment decisions at the early stage of movie productions. The system predicts the success of a movie based on its profitability by leveraging historical data from various sources. Using social network analysis and text mining techniques, the system automatically extracts several groups of features, including who are on the cast, what a movie is about, when a movie will be released, as well as hybrid features that match who with what, and when with what. [5] attempts to address this question by using machine learning techniques to predict film box office using Pseudo Inverse, SVM with Naive Bayes and Neural Network. [3] tries to predict the gross revenue of a movie from publicly available data and by using one independent method (Factor Analysis) and one dependent method (Multiple Linear Regression). In the exploratory factor analysis (EFA) phase, eight latent factors of the twenty binary genre variables were identified to be used in the regression modeling phase. [1] proposes a way to predict how successful a movie will be prior to its arrival at the box office. Instead of listening to critics and others on whether a movie will be successful, they applied machine learning algorithms to make this decision. A total of five machine learning algorithms (K-nearest Neighbor (KNN), Gaussian Nave Bayes (GNB), Decision Trees (DT), K-means Clustering, and Graphing Theory) were applied to a dataset comprised of movie data from 2 different sources (IMDB and YouTube). This dataset contained 500 randomly selected movies and 28 features. [2] makes extensive use of the number of views a

films Wikipedia page has attracted as a predictor of box office success in the United States and test their method using 325 films from the United States and then apply it to films from four foreign markets: Japan (95 films), Australia (118 films), Germany (105 films), and the United Kingdom (141 films) and find the technique to have inconsistent performance in these nations. While it makes relatively accurate predictions for the United States and Australia, its predictions in the remaining markets are not accurate enough to be useful.

V. PROPOSED METHODOLOGY

Methods tried to find a solution to the problem under consideration are as follows:

EDA
Linear Regression
Decision Tree
Random Forest
Light GBM
XGBoost
CatBoost
Neural Network

A. EDA

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

B. Linear Regression

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

C. Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

D. Random Forest

Random Forest is a supervised learning algorithm. It creates a forest and makes it somehow random. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

E. Light GBM

Light GBM is a gradient boosting framework that uses tree based learning algorithm.

Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm.

Light GBM is prefixed as Light because of its high speed. Light GBM can handle the large size of data and takes lower memory to run. Another reason of why Light GBM is popular is because it focuses on accuracy of results. LGBM also supports GPU learning and thus data scientists are widely using LGBM for data science application development.

F. XGBoost

The name xgboost, though, actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms. Which is the reason why many people use xgboost.

It is an implementation of gradient boosting machines created by Tianqi Chen, now with contributions from many developers. It belongs to a broader collection of tools under the umbrella of the Distributed Machine Learning Community.

Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. A popular example is the AdaBoost algorithm that weights data points that are hard to predict.

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

G. CatBoost

CatBoost is an algorithm for gradient boosting on decision trees. Developed by Yandex researchers and engineers, it is the successor of the MatrixNet algorithm that is widely used within the company for ranking tasks, forecasting and making recommendations. It is universal and can be applied across a wide range of areas and to a variety of problems.

While training time can take up longer than other GBDT implementations, prediction time is 1316 times faster than the other libraries according to the Yandex benchmark.

In addition to classification and regression, Catboost supports ranking out of the box.

Catboost introduces two critical algorithmic advances - the implementation of ordered boosting, a permutation-driven alternative to the classic algorithm, and an innovative algorithm for processing categorical features. Both techniques are using random permutations of the training examples to fight the prediction shift caused by a special kind of target leakage present in all existing implementations of gradient boosting algorithms.

H. Neural Network

A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes.[1] Thus a neural network is either a biological neural network, made up of real biological neurons, or an artificial neural network, for solving artificial intelligence (AI) problems. The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be -1 and 1.

VI. Data Exploration

Fig. 1. Word Cloud of Genres

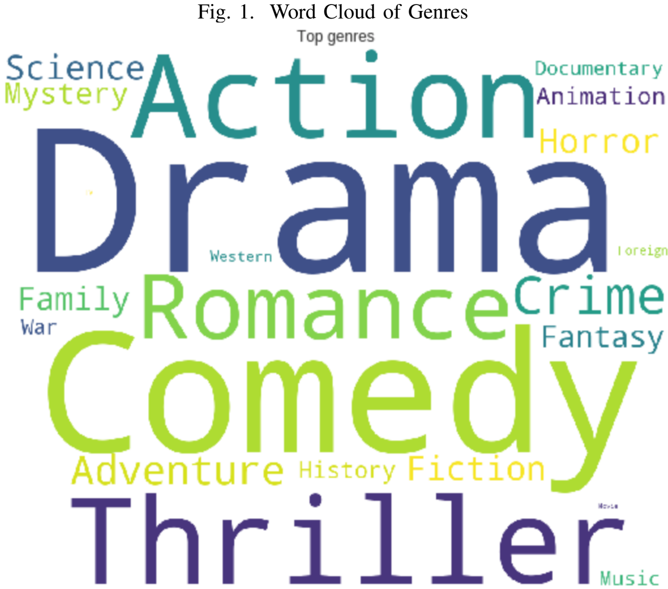


Fig. 2. Word Cloud of Taglines

Fig. 3. Top 10 Production Companies with respect to releases

```
[('Warner Bros.', 202),
 ('Universal Pictures', 188),
 ('Paramount Pictures', 161),
 ('Twentieth Century Fox Film Corporation', 138),
 ('Columbia Pictures', 91),
 ('Metro-Goldwyn-Mayer (MGM)', 84),
 ('New Line Cinema', 75),
 ('Touchstone Pictures', 63),
 ('Walt Disney Pictures', 62),
 ('Columbia Pictures Corporation', 61)]
```

Fig. 4. Top 10 Production Countries

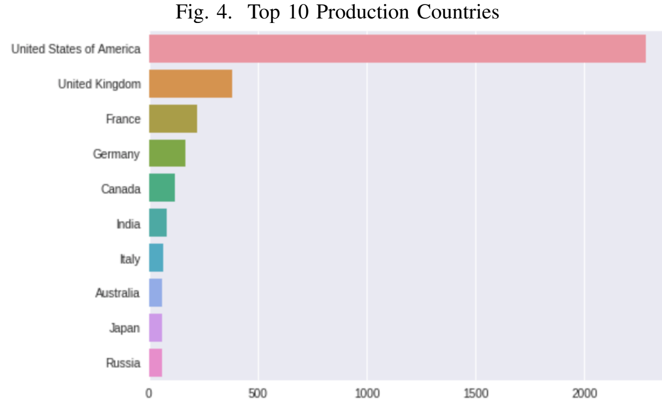


Fig. 5. Top 10 Genres

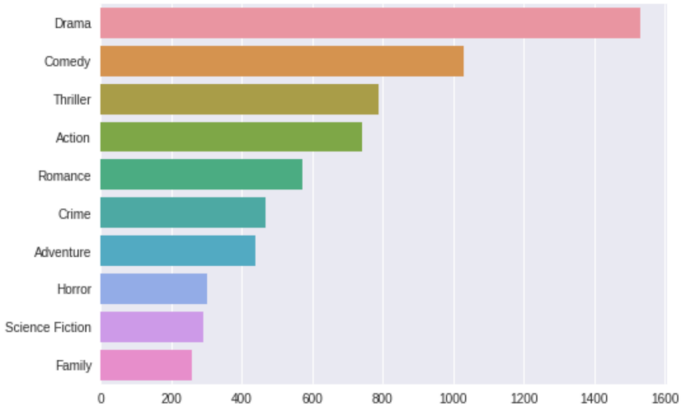


Fig. 6. Budget vs Revenue

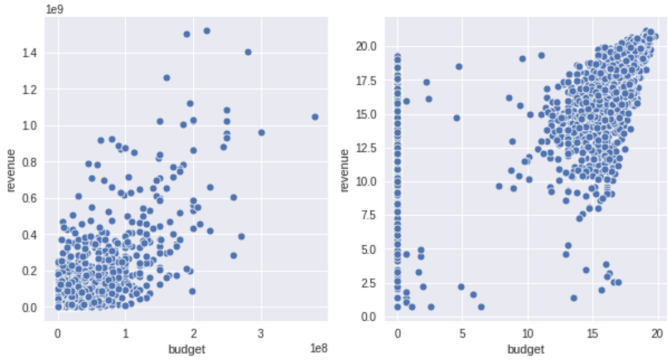


Fig. 7. Popularity vs Revenue

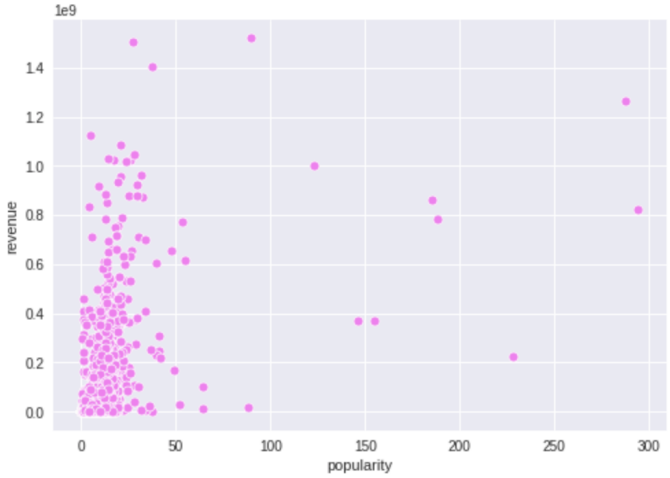


Fig. 8. Release day vs Revenue

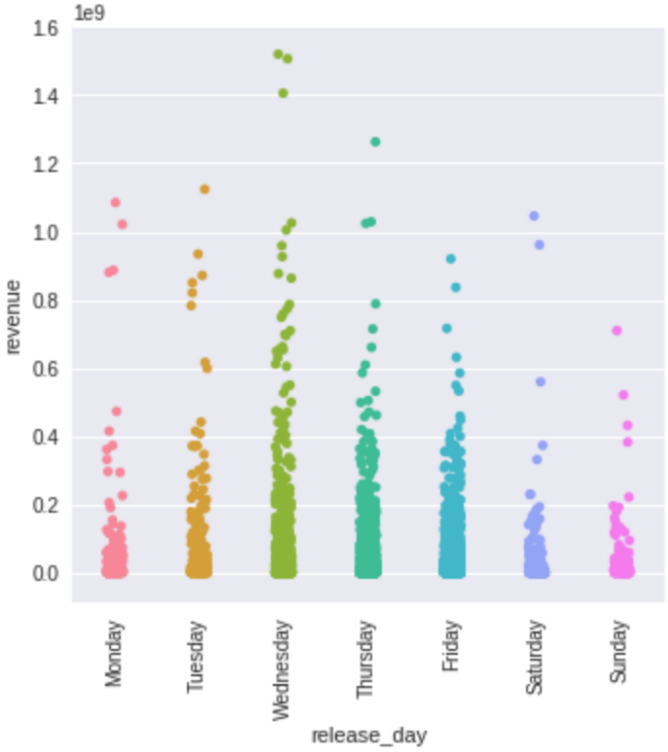


Fig. 9. Release month vs Revenue

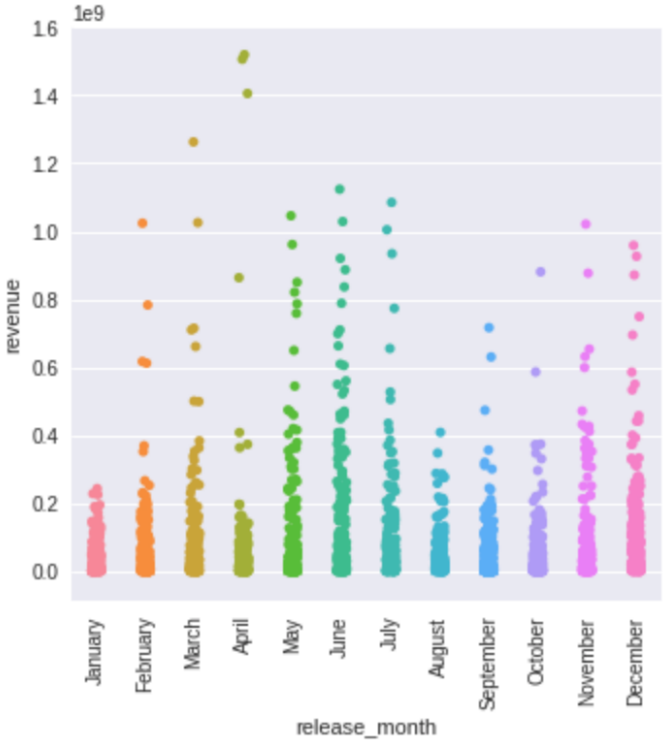


Fig. 10. Year vs Revenue

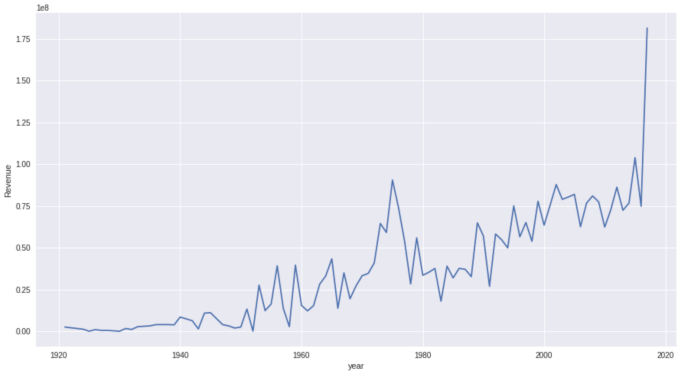


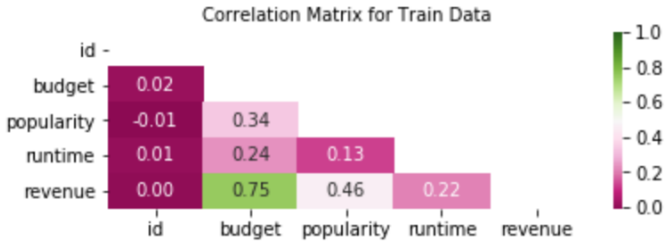
Fig. 11. Top 10 Crew members

```
[('Avy Kaufman', 50),
 ('Robert Rodriguez', 44),
 ('Deborah Aquila', 40),
 ('James Newton Howard', 39),
 ('Mary Vernieu', 38),
 ('Luc Besson', 37),
 ('Jerry Goldsmith', 37),
 ('Steven Spielberg', 37),
 ('Francine Maisler', 35),
 ('Tricia Wood', 35)]
```

Fig. 12. Top 10 Actors

```
[('Robert De Niro', 30),
 ('Samuel L. Jackson', 30),
 ('Morgan Freeman', 27),
 ('Susan Sarandon', 25),
 ('Bruce Willis', 25),
 ('J.K. Simmons', 25),
 ('Liam Neeson', 25),
 ('John Turturro', 24),
 ('Bruce McGill', 24),
 ('Willem Dafoe', 23)]
```

Fig. 13. Correlation Matrix



VII. Results

Method	Error
Linear Regression	2.71166
Decision Tree	3.05994
Random Forest	2.69673
Light GBM	2.64155
XGBoost	13.48911
CatBoost	2.64746
Neural Network	2.64746

TABLE I
Experimental Results

VIII. Conclusion

From Table I, it is clear that LightGBM has the best performance on test data while XGBoost has the worst. CatBoost and Neural Network have similar performance. Linear Regression, Random Forest and Decision Tree are close but not good enough.

References

[1] Cary D. Butler. Predicting movie success using machine learning algorithms. http://www.laccei.org/LACCEI2017-BocaRaton/student_Papers/SP499.pdf.

[2] Brian de Silva and Ryan Compton. Prediction of foreign box office revenues based on wikipedia page activity. *CoRR*, abs/1405.5924, 2014.

[3] Sharmistha Dey. Predicting gross movie revenue. *CoRR*, abs/1804.03565, 2018.

[4] Michael T. Lash and Kang Zhao. Early predictions of movie success: the who, what, and when of profitability. *CoRR*, abs/1506.05382, 2015.

[5] Pengda Liu. Machine learning on predicting gross box office. http://cs229.stanford.edu/proj2016/report/PengdaLiu-MachineLearningOnPredictingGrossBoxOffice-report.pdf, December 2016.

[6] Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, 8(8):e71226, 2013.