Movie Success Prediction using Machine Learning

Aniruddha Das1 - 12113509, Dimple Nagpal2

*1Student, Department of CSE, Lovely Professional University, Punjab.*

*2Professor, Department Of CSE, Lovely Professional University, Punjab.*

anirda17@gmail.com

***Abstract* ~ The film industry, currently, is a powerhouse of fame and revenue generation, it has stayed, consistently, at the top, among the most lucrative businesses/industries in the world, this has notably made it a riskier business too as a failed movie may cost shareholders and producers alike millions of dollars to recoup, it would be really helpful to know if a movie is going to succeed or not based on its various details and this can be especially useful for movie producers to make a financial decision. A movie’s success depends on a lot of factors such as budget, actors, direction, genre, reviews, marketing, hype etc. -- these factors have a direct impact on the revenue a movie generates, nowadays even movies with less than favorable reviews are making a ton of profit because of effective and shrewd marketing practices and the hype it generates among crowd, another factor to consider is the presence of an upcoming movie on social media, as that generally means that movie is getting noticed by the collective conscious of humanity, This paper gives a brief exploration of machine learning models that can be used to predict overall success of a movie based on it’ s various aspects.**

***Keywords ~* Machine Learning, ML, Supervised, Random Forest, Ensemble, Boosting.**

I. INTRODUCTION

Watching movies is a favorite pastime among many, In India alone, Indian box office sales have soared past 12, 460 crores [2023], which makes it the second most lucrative industry, losing only to tech industry, which is pretty impressive, it is no slouch either in other countries however.

This lucrative industry is a massive sector for making profitable investments, however naturally there are risks associated with it, if a movie bombs hard, it may generate no investment at all, effectively costing the investors and producers way more than they paid for.

Past three decades have seen a significant rise in data being generated about movie details. Sites like Metacritic, Imdb etc. are great sources of gaining information and trivia about movies, this gives us a clearer picture about the success factors of a specific movie, and there are some data points more important than the others here.

Predicting a movie’s success however is still a difficult procedure, as despite all the advancements in data collection and aggregation, human behavior is the key factor in determining success, as a rule of thumb, a good movie may not always be successful, but a successful movie is one that will click with its core audience the most and that will, hopefully, generate the most amount of revenue it needs to be considered successful by the investors.

The idea of the project is to develop several models to predict the revenue and determine its success based on that revenue, for the sake of simplicity.

A movie can be considered successful if it makes up all of its budget plus 50% percent of that budget as revenue, as a standard measure of success, but it can differ from investor to investor.

*Final Implementation*

Based on the core conceit of the problem at hand, relevant dataset was fetched and explored to see which features affect the revenue of the movie the most, based on that, various models were trained, like Random Forest, Voting Regressor, CatBoost and XGBoost. This exploratory project aims to compare and contrast the performance and accuracy of predicted output (in terms of variance) and then give relevant conclusion for each type of regressor that is used in this project.

II. RESEARCH CATEGORIZATION

The model made for this kind of research denotes success in terms of revenue, and this model is no different, Research could be classified into two distinct categories – Quantitative & Qualitative. Qualitative research focuses on success in of raw heuristics and numbers, while Qualitative research focuses on success of a movie in terms of reviews and watch-base and other such non-quantifiable things. Predicting a movie’s success should be helpful to investors or producers during the production or before the production of a movie when they are merely guessing and applying the various cost factor of the upcoming movie they are going to make, this model can make help them in decision making tasks related to budget concerns over that film.

III. STEPS & METHODOLOGY

The project has Four phases, as is the case with any machine learning project, they are:

* Data Extraction
* Exploratory Data Analysis and Feature Engineering
* Model Training/Building
* Conclusion and Prediction of Data

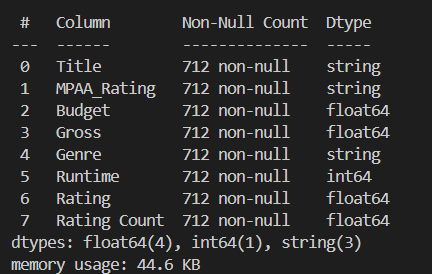
1. **Data Extraction**

The dataset was downloaded from UCI Data Repository, it has about 1000 entries and it consists of 10 columns, they are: MovieID, Title, MPAA\_Ratings, ratingscount, ratings, summary, actor-credit list, Director. The dataset is completely raw, so it was needed to be preprocessed before it’s eventual use into model training, however point to be noted is that, XGBoost and CatBoost have their own data preprocessing abilities, so raw datasets can be easily fed into, for training purposes, but preprocessing is required for other, simpler algorithms.

The dataset has about 1000 entries, a small portion of them consist of null values. summary, MovieID and actor-credit list cannot be used as features as they don’t really represent a relationship with other feature-sets in this dataset, hence they were removed during the preprocessing phase.

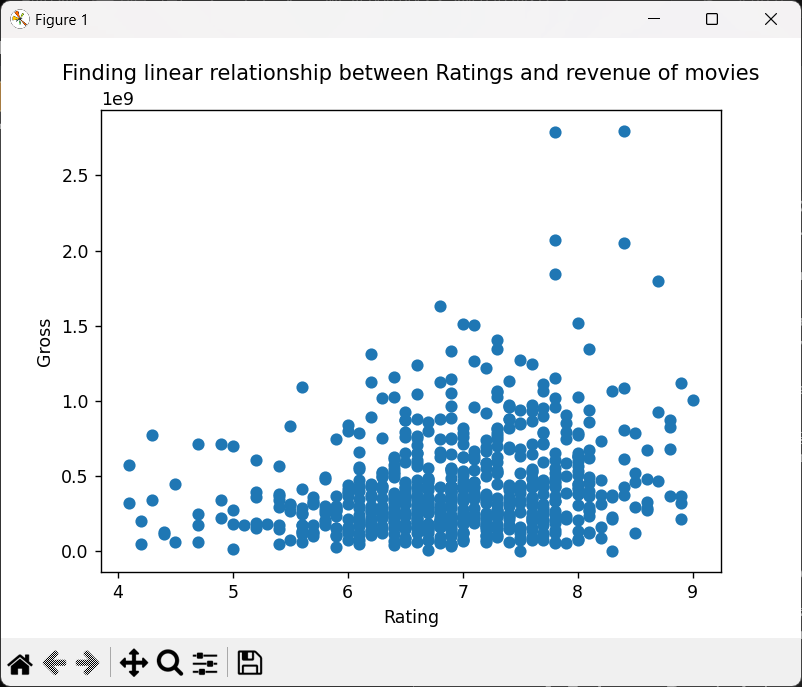
1. **Exploratory Data Analysis & Feature Engineering**

Once the data was preprocessed, A better look at the relationship between various features and feature-sets were taken to better understand which models would be used for training.



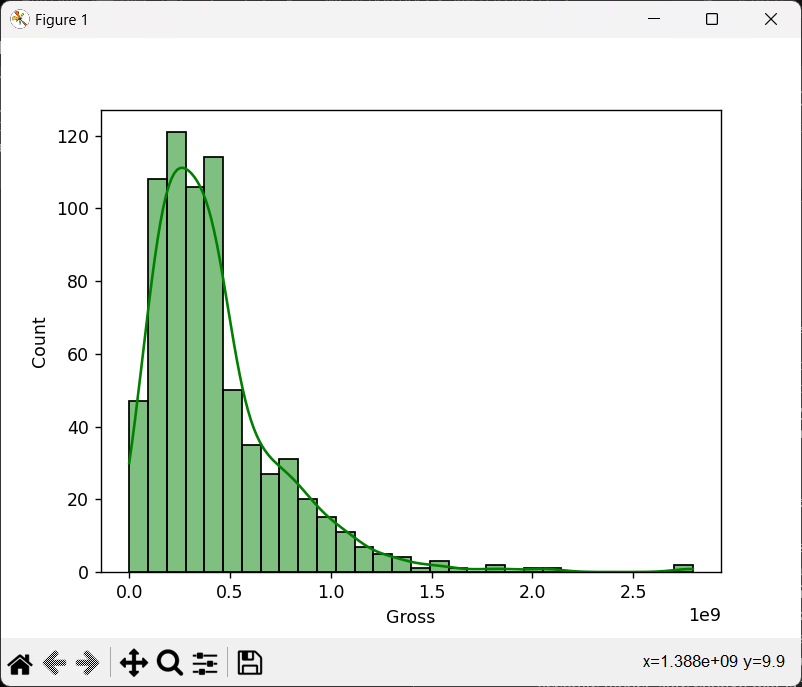
**Fig 3.1 Initial dataset review**

Various feature-sets had non-numeric or ordinal data that simply cannot be used for regression analysis; therefore, they were either converted into nominal datatypes or they were altogether removed because they simply did not contribute to the relations across the dataset.



**Fig 3.2 Scatterplot between Gross & Revenue**

Two of the key factors in determining the success of a movie and their relationship was explored using graphical exploratory data analysis tool called a scatterplot, which plots every datapoint with a circle(point) on a cartesian plane which allows us the see the relative distance between each and every datapoint for these two feature-sets, in this specific case we can see that there is high overlapping between data points which shows that simple regression models will likely fail to produce accurate results for this kind of relationship, therefore decision trees would be a far better tool to make a fairly robust and accurate model, sure enough that is the case, as we will see later on.



**Fig 3.3 histogram for gross revenue generated**

Next, a histogram was plotted, which showed that the data is somewhat lop-sided, if we check the dataset, using a .csv editor like MS Excel or any exploratory data tool, we can see that fairly recent movies generate a lot more reviews and revenue than old movies, as older movies were supplanted at the end of the dataset.



**Fig 3.4 Heatmap showing the correlation between feature-sets**

Next, a heatmap was generated that gave us a glimpse of correlation between various feature-sets within the data, we can see that the correlation values between them are fairly weak, a better dataset should have more values and feature-sets with stronger correlation between them, as regression hinges on this singular concept the most, when it comes to predicting future values based on previously included data.

1. **Model Training & Hyperparameter Tuning**

Based on the EDA done earlier, four algorithms were chosen to be used for model training, they are:

1. **Random Ensemble Forest**: This supervised machine learning algorithm uses the concept of regression with ensemble learning, that is predicting values based on past input using multiple CART (Classification & Regression Trees)

/Decision trees. It is robust to outliers and overfitting and gives a much better result for highly overlapping and complex dataset as compared to something simpler, such as linear regression, each decision tree can be considered a model in its own, therefore it is considered an ensemble ML algorithm.

1. **Voting Regressor**: This machine learning algorithm actually takes other ML algorithms as estimators which are then trained on the same dataset, the average of the output given by various algorithms is then taken as the final outcome, the estimators used in this project are, Linear Regressor, Support Vector Regressor, Random Ensemble Forest Regressor, due to its ensemble nature, it can be just as good as other more robust algorithms but, that depends on which algorithms are used as estimator to make a model.
2. **XGBoost Regressor**: XGBoost is an ensemble machine learning algorithm which uses the concept of Gradient Boosting during the model training process, It essentially combines several weak learners or, in this case ensemble forests with varying properties, which are then combined to create a strong learner, the output created by the first model trained, is bound to have errors in prediction, these erroneous data entries are then given to the next model in tow for training, which hopefully learn from the mistakes of the previous learner to give us a better outcome, but it is also bound to make some mistakes of its own, which are then given to the next model in tow, and the cycle repeats until we can maximize the accuracy of the prediction, it has many hyperparameters such as eta – learning rate, depth – tree height, colsample\_byheight(n) – inputs to be taken from n number of columns, n\_jobs – number of CPU threads to be used for model training. It is also much faster and robust than the default boosting model given in scikit library, due to its parallelized execution.
3. **CatBoost Regressor**: CatBoost is a robust, high performance, gradient boosting regressor developed by a Russian research group called Yandex. It can be used for both classification and regression tasks, and it doesn’t require pre-processed data, as it can pre-process data on its own, making it simpler to use, as compared to previously mentioned frameworks, it works by setting the means of the target variable and then starts removing residuals and inaccuracies by building a robust decision tree, it does this procedure that spans across multiple iterations, then it takes the best iteration and it’s corresponding decision trees as the model to be used for future prediction tasks.

All four model were trained using the feature-sets previously made. Now let’s see the conclusion to the model training…

1. **Conclusion**

For the evaluation metrics, I used R2 score along with RMSE (which penalizes overshooting values as compared to undershot prediction) as well as accuracy score for each of them. Based on this initial few runs of the program, Random Ensemble Forest individually did better than Voting Regressor as it has two estimator, namely Linear regressor and SVR which are not fit for this kind of feature-set, therefore taking into account that in Voting Regressor average of outputs of each estimator is taken, it is easy to see that it failed to give a better output.

XGBoost is the second worst algorithm which is just two points lower than the best algorithm – CatBoost. Due to its highly iterative and greedy approach to solving this particular problem, it gives us the best values of R2\_score, RMSE score and Accuracy as compared to other frameworks.

|  |  |
| --- | --- |
| Algorithm | R2\_Score |
| Random Ensemble | 0.589 |
| Voting Regressor | 0.506 |
| XGBoost | 0.561 |
| CatBoost | 0.599 |

|  |  |
| --- | --- |
| Algorithm | RMSE |
| Random Ensemble | 144.9736438724035 |
| Voting Regressor | 151.7917065935185 |
| XGBoost | 147.3696053709494 |
| CatBoost | 144.6606257296153 |

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| Random Ensemble | 58.9% |
| Voting Regressor | 50.6% |
| XGBoost | 56.2% |
| CatBoost | 60.0% |

**Fig 4.1, 4.2, 4.3 Conclusion**

1. **Comparatory Analysis**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy  (Higher the better) | RMSE (My Score)  (Lower the better) | RMSE ([Paper I](C://Users/anird/Downloads/Movie_Box-Office_Success_Prediction_Using_Machine_Learning.pdf)) | Accuracy ([Paper II](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8281839)) | Accuracy ([Paper III](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8275242)) |
| Random Ensemble | 58.9% | 144.97364387240345 | 357.068054 | 47.38% | 55.36% |
| Voting Regressor | 50.6% | 151.7917065935185 | 351.252695 | - | - |
| XGBoost | 56.2% | 147.3696053709494 | 348.237095 | - | - |
| CatBoost | 60.0% | 144.66062572961533 | 344.352039 | - | - |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

IV. REFERENCES

[1]. V. Subramaniyaswamy, M. Viginesh Vaibhav, R. Vishnu Prasad and R. Logesh, ”Predicting movie box office success using multiple regression and SVM”, 2017 International Conference on Intelligent Sustainable Systems (ICISS), Dec. 2019, pp. 182-186.

[2]. Gaurang Velingkar , Rakshita Varadarajan , Sidharth Lanka and Anand Kumar M, “Movie Box-Office Success Prediction using Machine Learning”, ieee.explore.org, 2019.

[3]. Dipankar Chaki, Md. Haider Ali, Md. Osman Gani, Nahid Quader, “A Machine Learning Approach to Predict Movie Box-Office Success”, ieee.explore,org, 2017.

[4]. Nahid Quader , Md. Osman Gani and Dipankar Chaki, “Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box Office Success Prediction”, ieee.explore.org, Dec 2017.

[5]. J. Ali, R. Khan, N. Ahmad, I. Maqsood, ”Random Forests and Decision Trees”, IJCSI International Journal of Computer Science Issues, vol. 9, Sept. 2012, pp. 272-278.

[6]. A. V. Dorogush, V. Ershov, A. Gulin, ”CatBoost: Gradient boosting with categorical features support”, arXiv:1810.11363, Oct. 2018.

[7]. S. B. Jha, R. F. Babiceanu, V. Pandey, R. K. Jha, ”Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study”, arXiv:2006.10092, Jun. 2020.

[8]. P. Chaovalit and L. Zhou, “Movie review mining: a comparison between supervised and unsupervised classification approaches,” in Proceedings of the Hawaii International Conference on System Sciences (HICSS), 2005.

[9]. A. Sivasantoshreddy, P. Kasat, and A. Jain, “Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining,” International Journal of Computer Applications, vol. 56, no. 1, pp. 1–5, 2012.

[10]. W. Zhang and S. Skiena, “Improving Movie Gross Prediction through News Analysis,” 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009.

[11]. M.H Latif, H. Afzal "Prediction of Movies popularity Using Machine Learning Techniques," National University of Sciences and technology, H-12, ISB, Pakistan.

[12]. T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo, “Predicting movie Box-office revenues by exploiting large-scale social media content,” Multimedia Tools and Applications, vol. 75, no. 3, pp. 1509–1528, Feb. 2014.

[13]. K. Jonas, N. Stefan, S. Daniel, F. Kai “Predicting Movie Success and Academi Awards through Sentiment and Social Network Analysis” University of Cologne, Pohligstrasse 1, Cologne, Germany.

[14]. P. Chaovalit and L. Zhou, “Movie review mining: a comparison between supervised and unsupervised classification approaches,” in Proceedings of the Hawaii International Conference on System Sciences (HICSS), 2005.

[15]. D. Gregorio, “Prediction of movies box office performance using social media,” Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13, 2013.