Predicting Movie Success using Regression Techniques

Faaez Razeen

***Abstract*—Hollywood is the largest and most profitable movie industry in the world. In 2018 alone, it generated a massive global box office of over $42 billion. With its size comes the need to analyze a movie's eventual revenue. A single production company with multiple movies may benefit greatly from knowing which movies are likely to succeed- it would help them focus their resources on the required advertisement and promotion campaigns. Furthermore, theaters would get a preference on which movies to run for a longer duration based on the success rate. Large-scale investments come with large risks. Using machine learning to predict revenues may help investors mitigate these risks.**

The algorithms in this paper aim to recognize historical patterns in the movie industry to try and predict the success of upcoming movies using a variety of machine learning algorithms. The success metric used is the box office, i.e. the commercial success of a film in terms of overall money earned.

Keywords—regression, machine learning, movie revenue.

I. Introduction

The movie industry is one of the most lucrative industries in the world. Including the global box office and home entertainment revenue, the global film industry is worth $136 billion as of 2018. Marketing and advertising costs take up to 50% of the entire budget allotted for a movie. As a big chunk of the budget, spending needs to be done wisely. Here's where machine learning comes to play. A movie production company often has many movies and TV shows under production at the same time. Since advertising is a huge expense, it would help the company to know which movie to put their money in. This is better supported by the fact that not all movies become blockbusters, i.e. movies that are of great commercial success. In this paper, the movies used are of only U.S. release in hopes of better performance due to homogeneity.

Successful prediction of a movie's success is complicated, as it depends on multiple factors. Success here can be defined in two ways- the revenue it earns or a quantification of how well it is received, i.e. user or critic score. Revenue is a much better definition of success in this case, as the amount of money a movie earns compared to its budget is a more rewardable measure of success.

The data used in this paper was a combination of data taken from IMDb (Internet Movie Database) and Facebook. Due to the fact that this paper uses metadata about a movie, predicting the revenue before it releases was not possible. Other papers which had done this used data like Twitter tweet rate, number of theatres the movie was going to release in, number of Wikipedia edits at the time before release, etc. These were possible to retrieve before release. This paper uses Facebook data. The disadvantage here is that Twitter is more representative of the hype surrounding a movie. Someone excited about a movie who uses both Facebook and Twitter will obviously be using Twitter to talk about it, due to the fact that Twitter tweets are meant to be real-time and to share trending information, while Facebook is more about connecting with friends, sharing photos and videos, etc.

The approach presented in the paper mainly focuses on regression algorithms. This paper aims to see if using machine learning algorithms can help predict movie revenue accurately. It also aims to find the features that are most important in determining a movie's revenue.

II. Literature Review

Nahid Quader et al. [1] used multi-class classification with 5 classes to predict whether a movie would be a 'flop' or a 'blockbuster', along with 3 other categories in the middle. Sentiment analysis was done by calculating the sentiment score for each review of a movie using Microsoft Azure's Text Analytics API and multiplying it by the total number of reviews. Using a variety of classification techniques, they found that IMDb vote count and number of screens were the best predictors. They have used two types of predictions: one an exact match which rewards correct classifications, and the other being a 'one-away' prediction, where the predicted class is considered correct if it is one away from the actual class. Considering the fact that there are only five classes, this method of classification is not indicative of actual accuracy. The inaccuracy is evident by the fact that using one-away prediction gave accuracies in the high 80s, while using normal prediction gave accuracies hovering around 50%. Such a difference is significant and should not be considered for real usage scenarios.

Steven Yoo et al. [2] found that budget had the most correlation (~0.63) with revenue. They have used numerical and textual features and a combination of both, "sentiment" features, which are taken from movie reviews and given a positive or negative weight based on a subjectivity lexicon obtained from the University of Pittsburgh OpinionFinder project, although it only leads to a minuscule increase in accuracy. They have theorized that this was due to two factors: the sentiment is already captured by the 'rating' feature, and the number of people who watched the movie matter more than the number of people who think the movie is good. When using linear regression, they managed to achieve a correlation of 0.7479 between the predicted

revenue and the actual revenue, which they consider not accurate enough to be used in practice.

Nithin Vr et al. [3] had used movies released in the USA and in the English language, in hopes of getting higher accuracies due to the fact that reviews are written in English. They implemented Linear Regression, Logistic Regression, and SVM Regression, and had accuracies of 51%, 42%, and 39%. The supposed cause for this was the small size of their data set (1050 samples). The overall low accuracies would not prove useful in practice.

Ting Liu et al. [4] had used purchase intention based on tweets to find out movie revenue. They found that the best correlated factor to box-office revenue was how many people are willing to see the movie. They took tweets from Twitter and split them based on purchase intention. They used this to predict a movie's first week income and eventual revenue. They used Linear Regression and Support Vector Machine as their models. Their best results are an adjusted R^2 value of 0.95 for predicting first week revenue, and 0.74 for predicting eventual gross. Their experiment verified their assumption that purchase intention is a better indicator than both popularity of the movie and sentiment analysis.

Carl Jernbäcker et al. [5] have used classification to predict success of movies, using both revenue and rating as metrics. They have taken inflation into account in order to reduce the number of possible outcomes from millions to a little more than a hundred. They have used three classification algorithms: decision trees, support vector machines, and variants of nearest neighbour classifiers. Although the accuracies for predicting rating were on par with other papers, the highest accuracy they achieved on predicting the gross revenue was 15%.

Asur and Huberman [6] have shown how social media can be used to predict future outcomes. They have used tweets referring to movies prior to their release to predict the box-office revenue generated by a movie in its opening weekend. Their research shows that the tweet-rate (the average number of tweets per hour) of any movie is highly indicative of the success of a movie's opening week revenue. Using only the average tweet-rate, they got an R^2 value of 0.80. Additionally, they have used 7 variables, created using time series values of the tweet-rate for the 7 days before release, each corresponding to a particular day, along with another variable which contains the number of theatres the movies were released in. Using sentiment analysis to calculate the ratio of positive tweets to negative tweets yielded a slight improvement in performance, however these sentiments were not as important as the tweet-rates themselves. Maximum prediction accuracy was achieved using both tweet-rates time series and theater count, with an R^2 value of 0.97 in predicting opening weekend gross. They have demonstrated that a simple linear regression model with only tweet-rate as a feature can perform better than artificial money markets like Hollywood Stock Exchange (HSX).

Hammad Afzal et al. [7] used various classification algorithms and managed to achieve an accuracy of 84% using logistic regression. However, they have not mentioned the number of output classes, which is important in gauging the efficacy of the algorithm. Classification would not prove useful in practice if a low number of classes were used.

Lopamudra Pal et al. [8] have used a combination of movie meta-data and sentiment analysis to predict movie revenue. They extracted sentiment scores from movie reviews and applied various regression techniques with and without the sentiment scores to see if they actually helped improve regression accuracy. The sentiment scores were calculated using Naive Bayes and VADER (Valence Aware Dictionary for Sentiment Reasoning). Using the RMSE (Root Mean Square Error) scoring metric did not seem appropriate for this problem, due to the fact that RMSE is very sensitive to outliers. Thus, they decided to convert their regression problem into a classification problem. A leeway of 15% was given, which means that if the prediction lies in this range, it is counted as an accurate prediction. The highest accuracies were achieved using Decision Tree Regression and Random Forest Regression. Although the inclusion of sentiment scores improved accuracy and decreased RMSE, the improvement was minor and deemed low enough to be not significant. Additionally, because they have converted their regression problem into a classification one, there is bound to be a loss in accuracy when used in real-time applications.

Márton Mestyán et al. [9] have used data from Wikipedia to predict movie box office success. To estimate the popularity of a given movie, they measured 4 types of activity on a page: number of views, number of users, number of edits, and collaborative rigor (quality of thoroughness of the article). They also included another feature which they considered essential: number of theatres. A multivariate linear regression model was used to predict revenue. Using all 5 features, the R^2 value achieved was 0.77.

III. Data Preprocessing and Feature Engineering

A. *Preprocessing*

Two separate datasets were merged using an inner join on the IMDB ID column. Since they were merged, there were columns with the same attributes. Columns with either lesser information or which contained a lesser degree of relevancy were discarded. Columns which played no statistically significant role in prediction were also removed.

1) Revenue: The merged dataset had two of the same columns. However, the difference between them was found to be 43%. The cause of this was because one dataset contained revenue for the region of USA alone while another had global revenue. The column with the global revenue was dropped in favour of USA revenue in order to keep the scope of the project homogeneous.

Feature Name	Description
vote_average	TMDB (The Movie DataBase) average vote rating
vote_count	TMDB total vote count
num_critic_for_reviews	Number of critic reviews on IMDb (Internet Movie Database)
num_user_for_reviews	Number of user reviews on IMDb
director_facebook_likes	Number of likes on Facebook for the director of the movie
actor_1_facebook_likes	Number of likes on Facebook for actor 1 of the movie
actor_2_facebook_likes	Number of likes on Facebook for actor 2 of the movie
actor_3_facebook_likes	Number of likes on Facebook for actor 3 of the movie
cast_total_facebook_likes	Number of likes on Facebook for actor 3 of the movie
revenue	Box office earnings of the movie inside the U.S.A. (in dollars)
num_voted_users	Number of people who voted for the movie
num_voted_users	Number of actors features on the movie poster
duration	Duration of the movie in minutes
budget	Total budget allocated for the movie in dollars
imdb_score	IMDb score of the movie
movie_facebook_likes	Number of likes on Facebook for the page of the movie
release_year	Year of release

B. Feature Engineering

Feature Engineering is a process of creating new features by utilizing available features and domain knowledge in order to improve the performance of machine learning models.

1) Genre: The revenue across different genres was also studied on. Upon visual analysis, it was found that Action and Adventure movies on average earned more revenue when compared to other genres. Thus, a new column was created, which contains a 1 if the genre of the movie is either action/adventure, else contains a 0.

2) Top Director: The top ten directors sorted in descending order by average revenue per movie were used to create a new column, called 'top_director', which contains a 1 if the director of that particular movie is a top director, else contains a 0.

3) Collection: If a movie was part of a series, this column contained a value of 1. If the movie was a standalone movie, this movie contains a 0.

IV. Feature Selection

Feature selection is the process of selecting only a set of useful features from the available lot. It is used to prevent overfitting, a condition where an algorithm captures too much noise of the data, rendering it inefficient on data it has not seen before. The removed features usually have little to no effect on prediction accuracy.

A. Numerical Attributes

Since only regression techniques are used, all nominal features were either dropped or converted to categorical features.

B. Backward Elimination

Backward Elimination is an iterative process starting with all candidate variables and in each iteration, deleting the variable whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit. In this paper, the probability value (p-value) was used. For a given statistical model, the p-value represents the probability that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two groups) would be equal to or more extreme than the actual observed results. The smaller the p-value, the greater statistical incompatibility of the data with the null hypothesis. [10]

In simpler terms, if the p-value of a certain feature is 0.05, then it means that there is a 5% chance that the results obtained were due to pure chance rather than due to the statistical features of the data. The industry standard for p-value is 0.05 or 5%, and the same is used here. Using the statsmodel library, in each iteration of calculating the p-values, the feature with the highest p-value is removed until all features have p-values < 0.05.

V. Regression Models

All algorithms were carried out using the Scikit-learn library [11] in Python. The efficacy of each model used would be judged based on four criteria:

- R^2, a value which measures goodness of fit of the model. indicates a perfect fit.
- Mean Absolute Error (MAE), which measures the difference between two variables, the actual value (y_i) and the predicted value ($\hat{y_i}$) . For easy interpretation of MAE, the features are normalized before applying algorithms.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y_i} |$$

- Root Mean Squared Error (RMSE), which is the square root of the Mean Squared Error (MSE). RMSE is a better metric compared to MAE, as it is sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

- Correlation between the predicted revenue and actual revenue. From the final subset of features, vote_count was found to be the highest correlated

feature with revenue, with a Pearson correlation coefficient of 0.7515. This is used as a baseline- all models used in this paper should be able to achieve a correlation equal to or greater than 0.7515.

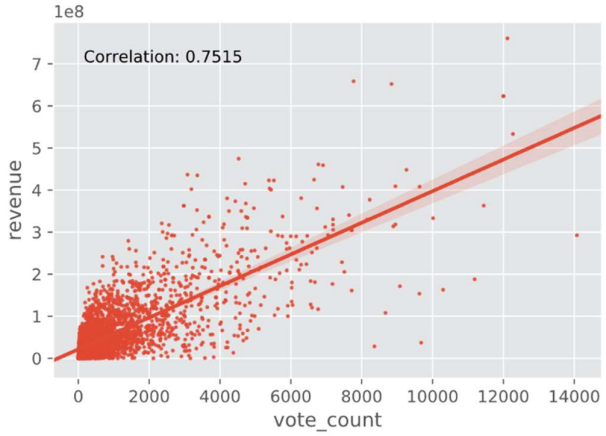

Figure 1. Correlation between vote_count and revenue.

All metrics except correlation were cross validated between 10 sets. Cross validation is a resampling procedure where each subset of the data is used both as a training set and as a testing set. Doing this helps prevent the model from overfitting on the data.

A. *Linear Regression*

Multiple linear regression is a technique that uses multiple explanatory variables to predict the outcome of a single response variable through modeling the linear relationship between them. It is represented by the equation below:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Where y_i = response/dependent variable, x_i = explanatory/independent variables, β_0 = y-intercept (constant term), β_n = slope coefficients for each explanatory variable.

B. *Support Vector Regression*

A Support Vector Machine is a classifier that aims to find the optimal hyper-plane (the separation line between the data classes with the error threshold value epsilon) by maximizing the margin (the boundary between classes and that which has the most distance between the nearest data point and the hyper-plane).

Instead of minimizing the observed training error, Support Vector Regression attempts to minimize the generalization error bound so as to achieve generalized performance. The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function. [12]

C. *Decision Tree Regression*

A decision tree is a supervised classification model that predicts by learning decision rules from features. It breaks down data into smaller subsets by making a decision based on asking a series of questions (the answers are either True or False), until the model gets confident enough to make a prediction. The end result is a tree, where the leaf nodes are the decisions.

The questions asked at each node to determine the split are different for classification and regression. For the latter, the algorithm will first pick a value, and split the data into two subsets. For each subset, it calculates the MSE (mean squared error). The tree chooses the value with the smallest MSE value. After training, the algorithm runs it through the tree until it reaches a leaf node. The final prediction is the average of the value of the dependent variable in that leaf node.

D. *Random Forest Regression*

Random forest is an ensemble method, which means that it combines predictions from multiple machine learning algorithms, in this case, decision trees. The problem with decision trees is that they are very sensitive to training data and carry a big risk of overfitting. They also tend to find the local optima, as once they have made a decision, they cannot go back. Random forest contains multiple decision trees running in parallel, and in the end, averages the results of multiple predictions.

E. *Ridge Regression*

Ridge regression is a technique used to mitigate multicollinearity in linear regression. Multicollinearity is the problem phenomenon where one feature in a multiple regression model can be predicted from other features due to a positive linear relationship. This causes inaccurate estimates of the regression coefficients. While the fulfillment of this assumption is not necessary for prediction, using a method like ridge regression showed vast improvement in performance.

Ridge regression L2 uses regularization, which is a method used to avoid overfitting by penalizing high-valued regression coefficients through shrinkage, where extreme values are shrunk towards a certain value. Particularly, In L2 regularization, the coefficients are penalized towards the square of the magnitude of the coefficients.

F. *Lasso Regression*

Similar to ridge regression, lasso regression shrinks all coefficients towards a value, in this case, the absolute value of the magnitude of coefficients. This is called L1 regularization, and can sometimes lead to elimination of some coefficients.

VI. Results

A. Comparison of models

Regression Model	R^2	MAE	RMSE	Correlation
Linear	0.7299	0.0331	0.0509	0.8584
Support Vector	0.6893	0.0408	0.2337	0.8532
Decision Tree	0.6893	0.0408	0.2337	0.7626
Random Forest	0.7342	0.0326	0.051	0.8941
Ridge	0.6893	0.0408	0.2337	0.8614
Lasso	0.6893	0.0408	0.2337	0.8614

Results above compare the performance of different regression algorithms based on certain metrics. The usage of a single decision tree was found to be the least effective algorithm for this scenario. The best performing algorithm was Random Forest, while Linear, Ridge, and Lasso fell shortly behind. While the former three algorithms seem to be similarly performing in terms of correlation and R^2, the error metrics are much higher in Ridge and Lasso compared to Random Forest and Linear Regression. Overall, Random Forest is the most suitable algorithm, followed by Linear Regression.

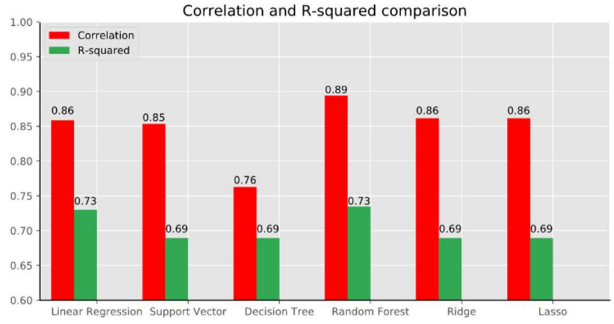

Fig. 2. Comparison of correlation and R-squared between algorithms.

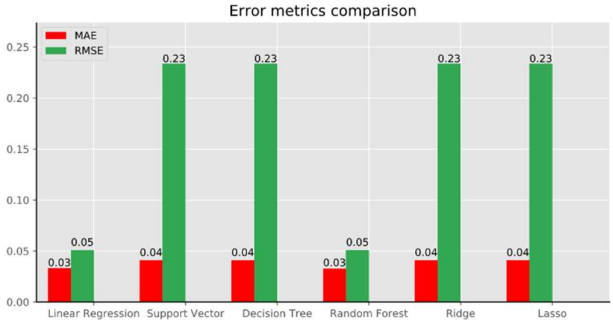

Fig. 3. Comparison of error metrics between algorithms.

B. The most important features

To find out the most important features in determining the success of a movie, random forest regressor was used. In each individual tree, the decision is made based on the MSE (Mean Squared Error). When training individual trees, the degree of how each feature decreases the MSE can be averaged. The features are then ranked accordingly in ascending order. The following bar-plot shows that vote_count and budget are the most important in determining the revenue of a movie. Similarly, vote_count was the best predictor in [1] and budget was the best predictor in [2]. This shows that people's opinion of a movie is more important in determining the success of a movie than the budget it was allotted.

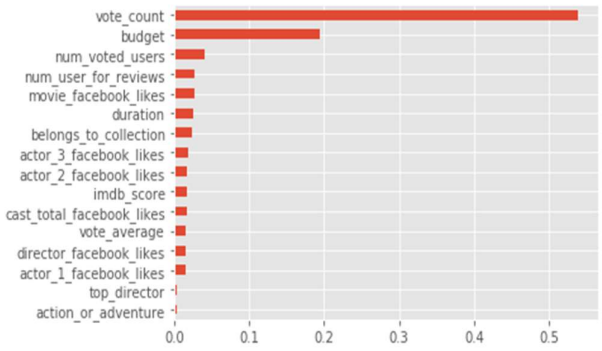

Figure 4. Feature importances using Random Forest Regressor.

The barplot shows also shows that out of the feature engineered columns, the belongs_to_collection column had more importance compared to the other engineered columns. The other two columns (action_or_adventure and top_director) had very less importance, but still contributed to a slight increase performance, proven when they were removed and the performance metrics were measured again.

VII. Conclusion and Future Work

The aim of this paper was to use regression techniques in order to predict movie revenue using movie metadata. Compared to other papers that were reviewed, the R^2 values were higher, except in one case [6], where Twitter data was used. While it is well researched that using real-time, trend based social media data like Twitter tweets can have excellent results, most other papers using movie metadata had lower performing algorithms when compared to the results found in this paper.

Error metrics could not be compared as other papers did not mention them. The best algorithm used was Random Forest, which had an R^2 value of 0.89. Most papers had implemented classification algorithms, which are unsuitable for predicting numerical values, as they have a window within which the actual value lies. If the size of this window is too big, results obtained may vary from reality.

Overall, the methods presented in this paper show that movie revenue can in fact be predicted successfully, although some might argue that the accuracy is not enough for real life use. For future work, an aggregation of trend-based social media data like Twitter, social media data from Facebook, and data from IMDb could be used to improve the prediction results vastly, to the point of deploying the algorithm for real-time use.

REFERENCES

[1] N. Quader, M. Gani, D. Chaki, and M. Ali, “A Machine Learning Approach to Predict Movie Box-Office Success." 02 2018. [Online]. Available: 10.1109/ICCITECHN.2017.8281839.

[2] S. Yoo, R. K. Kanter, D. C. Cummings. and A. Maas, "Predicting Movie Revenue from IMDb Data." 2011.

[3] N. Vr. M. Pranav, P. B. S. Babu, and A. Lijiya, "Predicting Movie Success Based on IMDB Data." *International Journal of Business Intelligents*. vol. 003, pp. 34-3,. 12 2014. [Online]. Available: 10.20894/IJBI.105.003.002.004.

[4] T. Liu, X. Ding, Y. Chen, C. Haochen, and M. Guo, "Predicting movie Box-office revenues by exploiting large-scale social media content." *Multimedia Tools and Applications*, vol. 75, pp. 1-20, 10 2014. [Online]. Available: 10.1007/s11042-014-2270-1.

[5] C. Jernbäcker and P. Shahrivar, "Predicting movie success using machine learning techniques," Stockholm, Sweden, 2017.

[6] S. Asur and B. Huberman, "Predicting the Future with Social Media." *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence*, *WI 2010*, vol. 1, 03 2010. [Online]. Available: 10.1109/WI-IAT.2010.63.

[7] H. Afzal, "Prediction of Movies popularity Using Machine Learning Techniques”, *International Journal of Computer Science and Network Security*, vol. 16, pp. 127-131, 08 2016.

[8] L. Pal, L. H. Bui, and R. Mody, "Predicting Box Office Success: Do Critical Reviews Really Matter?"

[9] M. Mestyán, T Yasseri, and J. Kertész, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data." PloS one, vol. 8, p. e71226, 08 2013. [Online]. Available: 10.1371/journal.pone.0071226

[10] R. L. Wasserstein and N. A. Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose", *The American Statistician*, vol. 70, no. 2, pp. 129-133, .2016. [Online]. Available: 10.1080/00031305.2016. 1154108

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] D. Basak, S. Pal, and D. C. Patranabis. "Support Vector Regression." 2007.

.