Predicting Movie Revenue from IMDB Data

Abhishek Gowda R M (4NM21MC001)
Ganavi C N (4NM21MC029)
NMAM Institute of Technology
Nitte, Karnataka

Can we predict a movie's total gross revenue based on the information available at the time of its release? Such data would be useful to marketers, theatre operators, and others in the film industry, but it is a difficult problem to solve, even for humans. Given a set of numeric, text-based, and sentiment features from IMDb, we discovered that linear regression outperforms class-based linear regression in predicting gross revenue. However, neither produces precise enough results to be used in practise.

Using a variety of machine learning algorithms, the algorithms in this paper aim to recognise historical patterns in the movie industry in order to predict the success of upcoming films. The box office, or the commercial success of a film in terms of total money earned, is the success metric used.

Keywords—regression, machine learning, movie revenue.

INTRODUCTION

The film industry is one of the most profitable in the world. The global film industry is worth $136 billion in 2020, including global box office and home entertainment revenue. Marketing and advertising costs can account for up to 50% of a film's total budget. Spending must be done wisely because it accounts for a large portion of the budget. This is where machine learning comes in. A film production company frequently has several films and television shows in the works at the same time. Given the high cost of advertising, the company would benefit from knowing which film to invest in. This is further supported by the fact that not all films become blockbusters, i.e. films with enormous commercial success. In this paper, only movies released in the United States are used in the hopes of improving performance due to homogeneity.

Predicting the success of a film is difficult because it is dependent on a variety of factors. Success can be defined in two ways: the amount of money earned or a quantitative measure of how well it is received, such as a user or critic score. In this case, revenue is a much better definition of success, as the amount of money a movie earns in comparison to its budget is a more rewarding measure of success.

The approach presented in the paper is primarily concerned with regression algorithms. The purpose of this paper is to determine whether or not using a machine
learning algorithms can assist in accurately forecasting movie revenue. It also seeks to identify the characteristics that are most important in determining a film's revenue.

DATA PREPROCESSING AND FEATURE ENGINEERING

Preprocessing :

An inner join on the IMDB ID column was used to join two separate datasets. There were columns with the same attributes because they were merged. Columns with less information or with a lower degree of relevancy were eliminated. Columns that did not play a statistically significant role in prediction were also eliminated.

Revenue :

Two of the columns in the merged dataset were the same. The difference between them, however, was discovered to be 43%. This was due to the fact that one dataset only contained revenue for the United States, whereas another contained global revenue. To keep the project's scope consistent, the global revenue column was removed in favour of the revenue from the United States.

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 Movies: The top ten movies sorted in descending order by average revenue per movie were used to create a new column, called 'top_movies', which contains a 1 if the director of that particular movie is a top movie, 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.

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.

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

REGRESSION MODELS

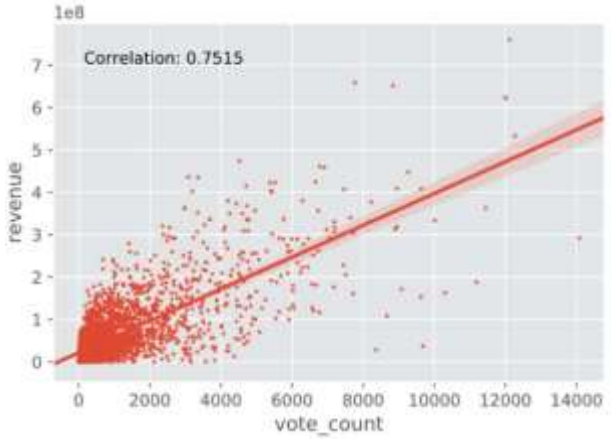
Linear regression algorithm is carried out using the Scikit-learn library in Python. The efficacy of model used would be judged based on four criteria:

- R^2, a value which measures goodness of fit of the model. indicates a perfect fit.

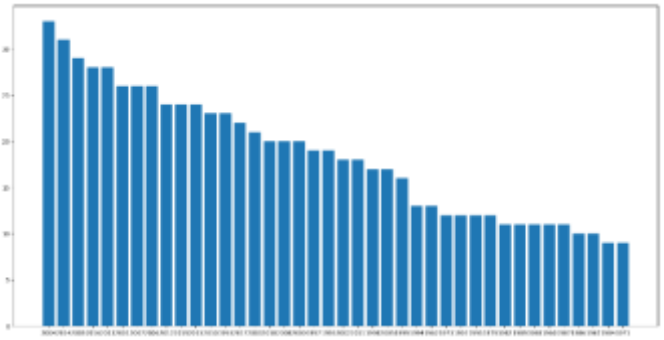
- Mean Absolute Error (MAE), which measures the difference between two variables, the actual value and the predicted value. For easy interpretation of MAE, the features are normalized before applying algorithms.

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

- 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.



Correlation between vote_count and revenue.



Graph between year of the movie and count of the movies released in the specific year

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.

	name	metascore	rating	runtime	votes
0	The Shawshank Redemption	80.0	9.3	142	2394059
3	The Dark Knight	84.0	9.0	152	2355907
9	Inception	74.0	8.8	148	2113984
10	Fight Club	66.0	8.8	139	1892181
7	Pulp Fiction	94.0	8.9	154	1862472
12	Forrest Gump	82.0	8.8	142	1851357
15	The Matrix	73.0	8.7	136	1710438
11	The Lord of the Rings: The Fellowship of the Ring	92.0	8.8	178	1693187
6	The Lord of the Rings: The Return of the King	94.0	8.9	201	1672460
1	The Godfather	100.0	9.2	175	1658439

Top 10 movies according to votes in descending order

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 yi = response/dependent variable, xi = explanatory/independent variables, β0 = y-intercept (constant term), βn = slope coefficients for each explanatory variable.

Mean_squared_error :

70.4781882284528

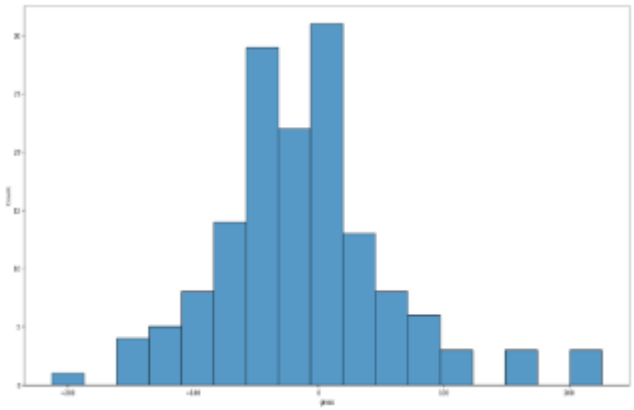
R^2_Score :

0.283954237908681

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 run time and count are the most important in determining the revenue of a movie. Similarly, run time was the best predictor in and count was the best predictor. 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.



CONCLUSION

We framed this problem as a regression and classification problem because we weren't sure which would produce the best results; as a result, we implemented both and devised methods to compare them. In general, we discovered that linear regression performs nearly as well as logistic regression for classification on our data, while having a much stronger correlation with actual gross revenues.

We discovered that the features we used (simple numeric, text, and sentiment features) were insufficient for making strong predictions of gross revenue. Others have had greater success using additional features such as the number of theatres, marketing budget, and so on, but we were unable to include them because IMDb does not contain such data. In the future, in addition to using different feature sets,

We might think about using better regularisation on linear regression to provide a more rigorous safeguard against high-variance models, because we consistently observed decreases in linear regression's test accuracy as the number of features increased.

Another, fundamentally different, data set that could be useful in predicting movie revenue is social graph data: with such data, we could analyse the characteristics of how a movie's popularity propagates through social networks, as well as the propagation tree's speed and extent over time. People's expectation to see a movie is represented by its propagation speed, which we expect to be directly related to its gross revenue.

XI. REFERENCES

[1] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In Proceedings of NAACL-HLT, 2010.

[2] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on, 1:301–304, 2009. This paper use AMAPE (Adjusted Mean Absolute Percentage/Relative Error) for their measurement

[3] Simonoff, Jeffrey S. & Ilana R. Sparrow. Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers. Chance Magazine, 13 (3), 15-24, 2000

[4] Brendan O'Connor, et al. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." Proceedings of the International AAAI Conference on Weblogs and Social Media, 2011

[5] Leonid Velikovic, et al. "The viability of webderived polarity lexicons." NAACL, 2010.