Predicting Movie Popularity on IMDB

Laura Hoffman, Bilal Kudaimi, and Erez Sarousi

Data Science, Bellevue University

DSC 630: Predictive Analytics

Dr. Fadi Alsaleem

August 3, 2021

Executive Summary

Ever wondered why Hollywood, despite employing 13,500+ actors and netting a $150 billion profit this year, can still be capable of producing mediocre movies? The reason behind this is because Hollywood relies on marketing to make movies successful. Marketing doesn't always work, though, with each mediocre film costing hundreds of millions in lost revenue. What if there was a way to predict how well a film would fare before it was released to eliminate this risk? After our team searched through possible projects that could maximize commercial profits, it was decided to take on predicting movie popularity! A machine learning model can be built, which is a program that looks at old data to find a relationship, then uses it on new data. In this case, we will program a model to look at old movie data to determine the relationship between those movies' information (director, actors, budget, etc) and how well they did on their IMDb ratings. It will then use this discovered relationship to predict how successful new movies will be using their IMDb based on their data as well. This way, Hollywood could potentially eliminate producing mediocre movies and focus solely on the ones predicted to be good ones. In the end, our model accurately predicted the IMDb scores of recently published movies as well as new ones.

Abstract

Data can provide us with answers to almost any question in any field of business, including the entertainment industry. With a data set that originally incorporated 28 features for over 5000 movies, this project explored the relationship between a few of these variables and the IMDb ratings variable in order to build a machine learning model that would predict the movie's score. The project used the IMDb rating column as a target variable for predictive analytics, and built the machine learning model using the other features as the predictors. The model was deployed to predict IMDb score as a measure of success or failure, which would allow the producers of the movie a better idea of what direction to take for their films. Ridge regression was used on the original data set as well as a one-hot encoded data set, alongside k-fold cross-validation and *train_test_split* to produce several models that predicted IMDb scores. Of the four models that were developed using different combinations of the methods listed above, the model using the one-hot encoded data with *train_test_split* (without using cross validation) had an average error of .93 (the highest) where the model without one-hot encoded features that used cross validation had the lowest average error of .69. Because the highest root mean squared error obtained was 0.93, the project can predict movie IMDb scores with a good amount of accuracy.

Introduction/Background

Hollywood is known to be a high-yielding and profitable business, worth over forty-two billion dollars as of 2019 (Watson, 2020). Just like any other business, their goal is to increase their profits by continuously producing high-quality content. Hollywood employs just under one million people (Sakoui, 2019), and roughly 13,500 of those people are actors in the United States (Gleeson, 2018). The financial side of the business is constantly poring through the data and are interested in generating more revenue while limiting their expenditures, resulting in optimized profits. Hollywood does not know in advance which movies would succeed, so in the interest of commercial profitability, it becomes necessary to predict which films would lead toward higher success. The work being conducted with this project will increase profits by predicting the quality movies and while saving money.

Based on prior data, it's easy to see how prior films fared, but is it possible to determine how they will do in the future? This certainly seems to be the case, yes. Using predictive analytics, this project seeks to discover how well future films will rank based on films from the past by building a machine learning model to draw a relationship between film statistics and their scores on the International Movie Database (IMDb). A movie's score out of 10 on IMDb will be used as the target variable, which tells the model how well a movie performed. In addition to optimizing profits, Hollywood can use this model to prioritize good movies and shift their focus away from movies that are predicted to have low IMDb scores. This ensures that Hollywood consistently produces entertaining movies that are well received by audiences and by extension, result in heightened profits.

Experimental Methods

Our methods will consist of three parts: Data preparation to make the data ready for analysis, exploratory data analysis (EDA) to gain insight into the data, and model building/evaluation, the final goal of the project. The first part, data preparation, will be completed using the R programming language, and the latter two through the Python programming language. R was deemed to be better suited for data preparation, and so it was decided to be used for the first step. To transfer the prepared data to Python, the dataset was saved as a local CSV file that was later imported into Python.

The first step was the data preparation and to conduct that, rows that had missing or incomplete values for the IMDb scores would be removed to ensure that the predictions would be. Whitespace, nonsensical and otherwise incompatible characters from the string columns were removed; this was to ensure that the model could read the data without any errors.Finally, one-hot encoded onto each movie genre was completed in order to ensure that the numerical vectors would be preserved and computed for the model input. Furthermore, duplicate movies were removed so that over and under-representation of our data would not occur.

With the data cleaned and prepared through R, the resulting data set was transferred to Python to begin the exploratory data analysis, which consisted of a few visualizations. This was done to better understand the data. First, a correlation heatmap was generated to determine which variables had little to no correlation with each other, then those variables will be removed. Next, histograms and bar charts were generated for the key continuous and categorical variables, respectively, to determine if there were any skewness present. The variables *color, language,* and *plot_keywords* were not

used, as the former two are overwhelmingly one element while the latter contains too many keywords to be useful for model building. Once these were constructed, a bar chart of the ten highest scoring film directors was made to determine if most good scores belonged to many directors or just a handful. Plots of IMDb score vs budget spent on the top fifty scoring movies and IMDb score versus movie facebook likes for all movies were also generated to determine if spending more on movie filming and marketing increases its score.

Once insight from the exploratory data was gained from the exploratory data analysis, the multiple regression model was constructed. To determine which type of regression to use, a correlation matrix was generated to inform the team how many of our variables were highly correlated. Those variables were then removed from the dataset. Furthermore, because there were several high correlations between many of our variables, ridge regression was deemed to be the appropriate choice and was chosen as the predictive analytics model for this project. Ridge regression was examined using two datasets: one using the default set and one using the set one-hot encoded to convert categorical features into binary features. For each model, both k-fold cross-validation (KFCV) and *train_test_split* were used. This resulted in four models total (see Table 1).

For models one and three, a 10-fold cross-validation with three repeats was conducted and utilized to eliminate any noise within the model, and for models two and four, a 80/20 split was used. This means that 80% of our data was used for training and the remaining 20% was reserved for testing. Model evaluation was conducted measuring the root-mean-squared error for all models. The errors between the

predicted and actual values were squared and averaged, and the square root of this value was reported in Table One as shown below. A lower error value indicates a regression model that outputs IMDb scores closer to the test values, and is generally seen to be more accurate.

Table 1:

Model	One-hot encoded features?	10-fold cross-validation used?	Train_test_split used?	Root Mean Squared Error
1	NO	YES	NO	0.69
2	NO	NO	YES	0.82
3	YES	YES	NO	0.87
4	YES	NO	YES	0.93

Results

Depending on what methods were implemented for each model, the average error was higher or lower. Average error correlates inversely to how close the model's predicted values are -- the higher the error, the farther the values are from the true values. The models that utilized 10-fold cross-validation had lower average errors than the models that used the alternate the *train_test_split* methods; this is because cross-validation uses all folds as the testing set to minimize error. Surprisingly, one-hot encoding our categorical features did not reduce the model error; instead, it increased the error. Furthermore, switching to the one-hot encoded dataset while using the KFCV method caused a much greater jump in error (26%) compared to using the *train_test_split* function (13%), however. Because of these differences, Model One had

the lowest error value. This means that the IMDb score values predicted by Model One most closely matched the actual values of the test dataset than any of the other models. It is therefore concluded that Model One is the model used for the predictive analytics project.

Finding out the results to the regression model was completed with the assistance of hyperparameter tuning and it shows that based on the history of previous films according, the next blockbuster hit is likely to be rated G or R, between 150 to 200 minutes long, is directed by Marc Forster, will star Leonardo DiCaprio as Actor 1 and Rory Kinnear as Actor 2, both of which will have high likes on Facebook. The film will also be produced in either the United Kingdom or France, and be any genre that is not comedy, to boost ratings.

Discussion and Conclusion

Following the data evaluation, exploration, and analysis, the models were completed in predicting the target variable, IMDb scores. Model One, which used ridge regression with KFCV and no one-hot encoding, resulted in a root mean squared error of just over 0.69, which is very accurate and successful in predicting the IMDb score values. This means that the model developed by this project is able to predict the values of IMDb scores of future films with relatively little error. Deploying this model on films for the future is only a matter of inputting new movie data into the model, and will allow for producers to see their potential ratings in real time.

The goal of this project was to accurately predict the IMDb scores of future films in order to give insight to the movie producers on how their projects might be received

by audiences. This project could be an important advancement for the entertainment industry if it was implemented, because it allows for movie producers to view the prediction on how well their projects will perform and can help them change or improve the course and direction for their films if it is needed. Now that the model is developed and confirmed to be accurate, it can be put to use on the films yet to be released. Producers and directors of future movies can now rest assured that an accurate predictive analytic algorithm can reveal to them their likely ratings which would help guide their decision making, increase viewership and maintain a loyalty to the production brand. This ultimately results in what Hollywood seems to deem as the greatest barometer of success - increased profits.

Acknowledgements

We would like to thank Professor Fadi Alsaleem and all of our DSC 630 classmates for their continued enthusiasm and support during this MSDS program.

References

Gleeson, P. P. D. (2018, June 27). Statistics on People Getting Famous in Acting. Work - Chron.Com. https://work.chron.com/statistics-people-getting-famous-acting-23946.html.

Sakoui, A. (2019, March 18). Hollywood Employs More Workers Than Mining and Farming, MPAA Says. Bloomberg. https://www.bloomberg.com/news/articles/2019-03-18/hollywood-tops-mining-crop-production-in-employment-mpaa-says.

Watson, A. (2020, November 10). Film Industry - statistics & facts. Statista. https://www.statista.com/topics/964/film/.