Preliminary Analysis: Predicting Movie Popularity

Laura Hoffmann, Bilal Kudaimi, Erez Sarousi

Abstract

The answers data can provide us with spans across almost 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 features and the IMDb ratings variable in order to build a machine learning model that would predict the movie's score. The project used the rating 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, in order to 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 constantly producing high-quality content. Hollywood employs just under a million people (Sakoui, 2019), and roughly 13,500 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 producing more revenue while limiting their expenditures, resulting in optimized profits. The work being conducted with this project will increase profits from quality movies 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? Using predictive analytics, this project seeks to discover how well future films will do 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 telling the model how “well” a movie performed. In addition to optimizing profits, Hollywood can use this model to prioritize good movies and shift focus away from movies predicted to have low IMDb scores.

Experimental Methods

Our methods will comprise 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 former will be done using the R programming language, and the latter two with the Python programming language. We felt that R was better suited for data preparation, and so decided to use it for our first step. To transfer the prepared data to Python, we saved the dataset as a local CSV file then we imported it into Python.

For the data preparation step, we first checked for and removed any row with missing/incomplete values to ensure our IMDb score predictions would be accurate. We then removed whitespaces and nonsensical characters from our string columns to ensure our model could read our data without fail. Finally, we one-hot encoded each movie genre and removed all duplicate movies to avoid over- and under-representation of our data.

With our data prepared, we transferred the data to Python to begin our EDA, which consisted of a few visualizations 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. Then, histograms and bar charts were generated for the key continuous and categorical variables, respectively, to determine if there is 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 10 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 50 scoring movies and IMDb score vs movie facebook likes for all movies were also generated to determine if spending more on movie filming and marketing increases its score.

Once we'd gained our insight from the EDA, we constructed our multiple regression model. To determine which type of regression to use, we constructed a correlation matrix to tell us how many of our variables were highly correlated; those variables would be removed from the dataset. Because there were several high correlations between many of our variables, we chose to utilize ridge regression. We examined ridge regression using two datasets: one using our default set and one using our 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, resulting in four models total (see Table 1).

For models 1 and 3, we chose to utilize a 10-fold cross-validation with 3 repeats to eliminate noise in the model, and for models 2 and 4, we used 80% of our data for training and 20% for testing. Model evaluation was conducted using 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 1 below. A lower error value indicates a regression model that outputs IMDb scores closer to the test values.

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. Our models that used 10-fold cross-validation had lower average errors than the *train_test_split* using models; 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, but increased it. Switching to the one-hot encoded dataset while using KFCV caused a much greater jump in error (26%) compared to using the *train_test_split* function (13%), however. Because of these differences, Model 1 had the lowest error value. This means that the IMDb score values predicted by Model 1 most closely matched the actual values of the test dataset than any of the other models.

Discussion and Conclusion

Following the data evaluation, exploration, and analysis, the models were completed in predicting the target variable, IMDb scores. Model 1, which used ridge regression with KFCV and no one-hot encoding, showed us a root mean squared error of just over 0.69, which is very 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 little error. Deploying this model on films for the future will be relatively simple and allow for producers to see their potential ratings.

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 can be an important advancement for the entertainment industry because it allows for movie producers to view the prediction on how well their projects will perform and can help them change course or direction for their films if it is needed. Now that the model is developed and confirmed to be accurate, it can be implemented toward the films yet to be released. Producers of said movies can now rest assured that an accurate predictive analytic algorithm can guide their decision making and increase their profits.

Acknowledgments

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/.