Mid-Progress Report: ***Movie Success Analysis***

Nisa Gurung & Kristel Tan

Abstract

The main goal of this project is to identify trends in movie attributes that correlate with a movie's success and genre. Movie attributes include the duration of a movie, its allocated budget, featured actors, title length, director, plot keywords, and number of faces in the movie poster. For the purpose of this project, we define a film's success by how much gross profit it made in the box office. By analyzing this data, we expect to be able to make recommendations for movie attributes and predictions of whether or not a given set of attributes will result in a highly profitable movie launch.

Introduction

The primary application for this project is to benefit and increase competition in two major industries — the entertainment and advertising industries. By the end of our analysis, this project will identify which movie attributes negatively or positively affect a movie release the most. With this information, filmmakers, directors, and actors can determine whether or not they would like to pursue a movie idea or make any adjustments to their planned release. This will also inevitably impact how a film is marketed to their target audience and the general public.

The dataset to support this project was initially retrieved from 'kaggle.com'. It was a CSV file of 5,000 movie attributes scraped from IMDB and 'the-numbers.com'. We cleaned this dataset of entries that were missing attributes and narrowed it down to the top five genres as well. Because this step eliminated a significant amount of data, we scraped more data from IMDB's top box office lists for the past five years. We then merged this data back with the original dataset, totaling roughly 4,500 relevant movies that tended to be successful and had all of the attributes we were looking to analyze.

At a high level, we expect that the attributes that strongly affect a movie release will differ based upon the genre(s) it belongs to. In this study, we will be analyzing the top five genres from the dataset: action, drama, comedy, thriller, and romance. In general, we believe that all movies with a bigger budget will perform better because they can afford more popular actors, a better set, and a quality crew. In particular, we predict for example, that romances short in runtime will do better than longer ones because romances do not typically have a complex plot.

Technique Used in this Project

For this project, we plan to use three different techniques to help identify movie attributes related to its success. These methods include multiple linear regression, k-means clustering, and logistic

regression. Multiple linear regression is a natural choice because it gives us an idea of whether or not certain independent variables are significant to a top genre. In particular, for the mid-progress-report, we looked at the duration, title length, and budget of a movie in relation to the gross profit it made. Because these attributes are all very numerically defined, we believe applying a multiple linear regression model to them can help us easily determine whether or not they affect the gross profit. The OLS linear regression function will provide us with coefficient and confidence interval data to accomplish this.

Furthermore, we plan to use k-means clustering to analyze a different set of attributes. For the mid-progress report, we've chosen to focus on evaluating a movie's plot keywords for the top 1,000 profitable movies of our overall dataset. In other words, we will cluster the top 1,000 profitable movies by the most similar plot keywords that appear, provided from IMDB's website. At a high level, this will give us an idea of what plots form the largest clusters, which also indicate what storylines movie critics and audiences find most interesting. This technique can also be applied to words found in movie titles or synopses for further analysis in the future.

Lastly, we will utilize the logistic regression technique to help us make predictions about a movie's other miscellaneous characteristics such as featured actors, directors, and number of faces in its movie poster. Similar to the multiple linear regression, this model will reveal one of two binary outcomes — the attribute contributes to a movie's success or it does not. This is interpreted from the coefficient, confidence intervals, and probabilities returned from the logit function. This should enable us to make more predictions given a set of information about a movie's attributes.

Datasets & Experiments

The datasets used in our experiments are manipulated subsets of cleaned data from the master set of movies. Because each analysis technique focuses on a different set of movie attributes, we create new, more relevant data frames to help perform the respective experiment. For example, in the multiple linear regression model, we extracted a dataset that only consisted of the movie title, gross profit, genres, duration, calculated title length, and budget. In a similar fashion, for the k-means clustering technique, we used a subset that consisted of only movie titles and plot keywords. Using smaller subsets makes it easy to iterate through the data and focus on only what each experiment needs.

In this mid-progress report, we performed our first two experiments: multiple linear regression and k-means clustering. As stated in the previous section, our goal with the linear regression model was to identify whether or not there was a strong correlation between the gross profit a movie made and its duration, title length, and budget. Before determining this model, we first identified which movies belonged to which top five genres (i.e. action, drama, comedy, thriller, romance). We then calculated the number of words in each movie's title. Finally, we passed the data to the statsmodels api OLS function and plotted the partial regression of the results. We completed these steps for each of the top five genres and recorded our observations.

For k-means clustering, we extracted a list of movie titles and their corresponding plot keywords. We used the TfidfVectorizer to obtain a tf-idf matrix. One of the parameters for this vectorizer was the tokenize function, which returned a filtered list of tokens. Then, we evaluated the number of clusters for our k-means plot, passing the vectorized matrix of the keywords and the max number of clusters as 6. After that we ran the k-means algorithm on the tf-idf matrix. Since we did not have numeric values to visualize the results of the k-means, we used Multidimensional Scaling (MDS). We obtained a distance matrix for the tf-idf matrix using cosine similarity and using this we obtained coordinates for visualization from MDS.

Results & Discussions

Our initial results from the multiple linear regression experiment provided some useful insight for each top five genre. We discovered that the number of words in a movie title does not have any predictive value of success for all of the movie genres because the confidence interval for each contained 0. We believe this observation is supported by the fact that even if a popular movie has a long title, the general public will find ways to shorten it or call it by a common name anyway, thereby having little effect on its success. This was also a similar case for the allocated budget of romance, comedy, and drama films. We suspect that budget may contrastingly have a greater impact on action and thriller movies because they typically have multiple set locations and many actors. Additionally, the runtime of a movie seemed to be one of the most significant attributes for all genres except thrillers. This is no surprise to us as well because the duration of a movie can often times determine how engaged an audience will be and may even affect whether or not someone chooses to see it in theaters.

In regards to k-means, our initial results indicate that among the highest grossing movies, the overarching themes of the clusters are violence and war, love and relationships, sex and nudity, and box office sentiment. This begins to give us an idea of what plots and storylines do best in the box office. Our current algorithm produces a lot of repetitive tokens, so the next step would be to improve the stemming and experimenting further with the parameters for the TfidfVectorizer such as changing the values for the min_df and max_df. Another improvement might be to parse out words pertaining to box office sentiment since they are not representative of a movie's plot.

Conclusion

Although our initial results have given us some insight about a movie's characteristics and its success, we believe that it would be premature to make definite conclusions because there are further improvements and analysis that can be made on our dataset. For example, in our linear regression model, we plan to remove insignificant independent variables and rerun the experiment on other possibly more important attributes. We also plan to run an additional logistic regression analysis on the dataset to make predictions about other movie characteristics.