Predicting Academy Award Best Picture Winners

Michael Chen
maichen@ucdavis.edu

Douglas Turner
dcturner@ucdavis.edu

June 7, 2018

1 Introduction

The Academy Awards are a yearly award ceremony honoring the best movies of the year and those who contributed to their making. The penultimate award is the Academy Award for Best Picture, which is given to the best movie of the year from a pre-announced set of nominees. Every year, there is substantial speculation about which movie will win the award. The goal of this project is to predict Academy Award for Best Picture winners from a list of Academy Award for Best Picture nominated movies.

Data is scraped from Wikipedia and IMDb. The features included in this dataset and analysis are: Director, Producer, Written by/Screenplay by, Cast, Budget, Box Office, Language, Runtime, Year, Rating, Genres, and Plot. A number of binary classification methods are considered: logistic regression, support vector machines, k-nearest neighbors, random forests, and AdaBoost. These methods are implemented, analyzed, and compared. While a variety of these variables factor in exploratory data analysis and logistic regression analysis, textual analysis of the plot is the primary component of binary classification efforts, but all relevant features are included in the construction of a similarity measure (kernel).

These classification and prediction algorithms, as well as the general analysis of this project, could be used to predict which movies, particularly movie plots, are most likely to be successful at the Academy Awards. Production company executives could use these algorithms to assess the quality of scripts, and industry analysts could use these methods as a basis for speculation about yearly results.

2 Preparing the Data

First, data is scraped from Wikipedia pages for each Academy Award nominated movie. The variables collected include: Movie Name, Director, Producer, Written By, Screenplay By, Budget, Box Office, Starring, Languages, Plot Summary, Running Time, and a brief description of the film. Additionally, the following variables are scraped from IMDb: Year, Genre(s), and IMDb rating. Note that many movies are categorized as being more than one genre.

After the data is collected and processed into an Excel file, it is split into testing and training portions (70-30 split). Additionally, data vectors are created for each of the variables for later use in analysis. For Plot Summary, the raw data consists of a long string that summarizes the plot of each movie. Term frequency times inverse document frequency is used as a measure of the prevalence of a word (or n-gram of words) in each document. The final result is a matrix where each row is a movie and each column is the prevalence of a given feature (word or n-gram of words; we have set it to examine 1-grams and 2-grams) in that document.

3 Exploratory Data Analysis

3.1 Data visualization

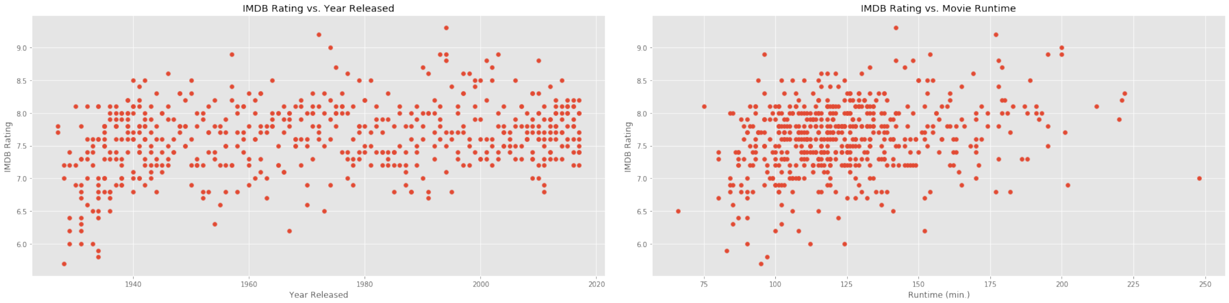

Figure 1: IMDb Rating vs. Year of Release and Runtime

Figure 1 shows IMDb rating plotted against year of release and movie runtime for all movies in our dataset. Ratings are relatively homogeneous over the decades, although there are a few trends that we take note of. Movies prior to 1940 seem to have a lower rating on average than the rest, and movies within the past ten or so years have noticeably less variation in their ratings. The largest amount of variation in ratings occurs roughly between 1955 and 1995. There is a weak correlation ($r = 0.28$) between year of release and IMDB rating - nominated movies from later years tend to have a higher rating. Additionally, there appears to be somewhat of a positive correlation ($r = 0.27$) between movie runtime and rating; among movies nominated for an Academy Award, longer ones seem to enjoy a slight advantage when it comes to rating.

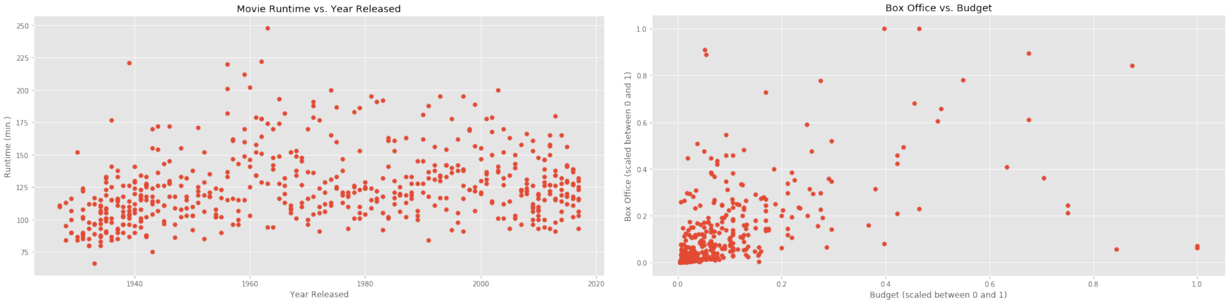

Figure 2: Runtime vs. Year of Release and Box Office vs. Budget

Figure 2 shows movie runtime plotted against year of release and box office return plotted against budget. Movie length (runtime) seems to have some relationship with year of release, with movies prior to 1958 or so having noticeably shorter runtimes on average than those after 1958. From 1970 onwards, the distribution of runtimes relative to year of release appears very homogenous. There is a weak correlation ($r = 0.24$) between year of release and runtime - nominated movies from later years tend to be longer. As one might expect, there is a medium-strength correlation between the movies' budget and box office return ($r = 0.58$). Note that the budget and box office figures have been rescaled to be between 0 and 1 (for each movie, its budget/box office is divided by the maximum budget/box office among all movies). The majority of movies have relatively low budget and box office figures compared to the maximum (less than 20% of the maximum budget and less than 40% of the maximum box office).

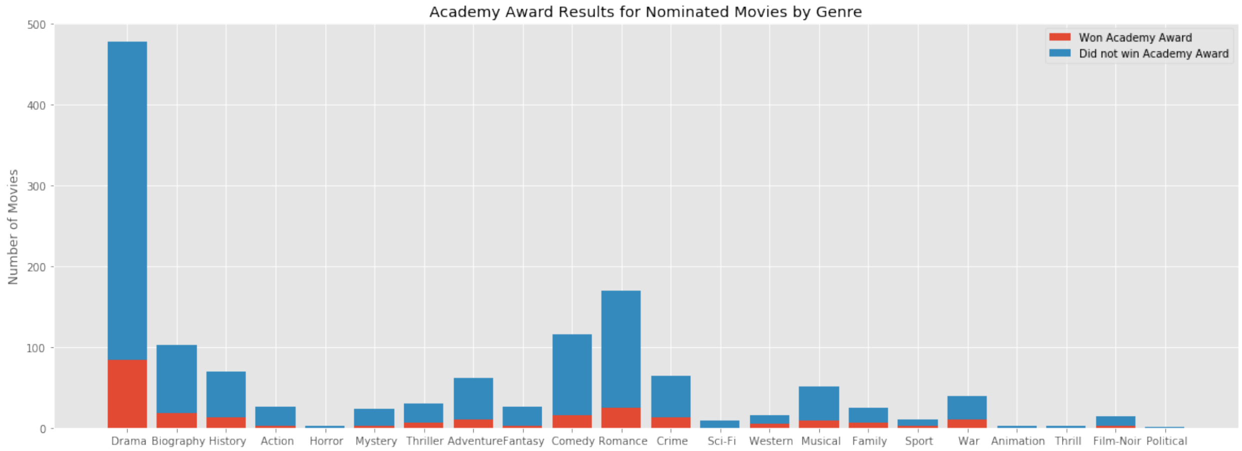

Figure 3: Barplot of Results by Genre

Figure 3 displays the win/loss results for each genre in our dataset. First, we should note that many movies in our dataset have multiple genres. For the purposes of this visualization, every genre of a movie is assessed a ‘win’ or a ‘loss’ depending on whether the movie won the Academy Award that year.

It is immediately apparent that Drama is by far the most popular genre among nominated movies with over double the number of movies categorized as Drama comapred to the next-most popular genre (Romance). The overall win percentage is $90/546 = 16.5\%$.

Romance (14.8%) and Comedy (13.0%) movies both have a below-average winrate despite fairly large sample sizes (169 and 115 respectively), while Drama (17.6%), Biography (17.6%), History (18.6%), and Crime (20.3%) all enjoy somewhat above-average winrates (sample sizes of 477, 102, 70, and 64, respectively).

Two particular standouts genre are War and Western - war movies enjoy a winrate of 25.6% with a respectable sample size of 39, while Westerns have had a winrate of 31.2%, though with a small sample size (16).

Action movies have performed poorly, with a 7.7% winrate out of 26 nominations. Five genres have movies that were nominated for an Academy Award, but have not yet won - these are Horror, Sci-Fi, Animation, Thrill, and Political. Among these, sci-fi movies have had the most opportunities, with 9 nominations being classified as sci-fi over the years.

3.2 Analysis of the response variable with select predictors

In this section, we use logistic regression to analyze the relationship between the response variable (binary: 1 if won Academy Award, 0 otherwise) and select predictors.

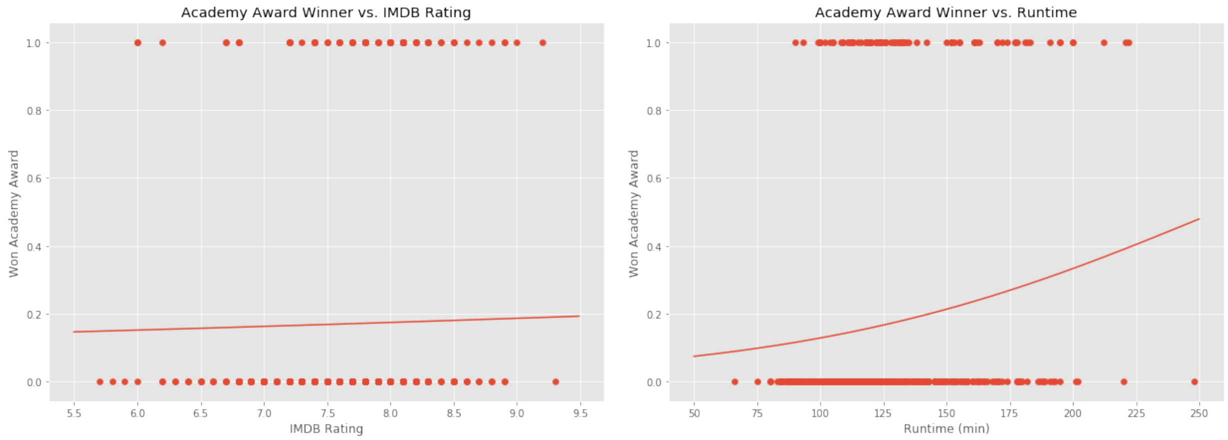

Figure 4: Response vs. IMDb rating and Runtime

Figure 4 shows the response variable plotted against IMDb rating and movie runtime. After fitting a logistic model to the data, we obtain the logistic curve and overlay it on the scatter plot. We see that there is a slight relationship between IMDB rating and whether a movie won the Academy Award, though perhaps not as much as we would expect. Interestingly, there appears to be a noticeable relationship between a nominated movie's runtime and whether it won the Academy Award: longer movies were more often winners.

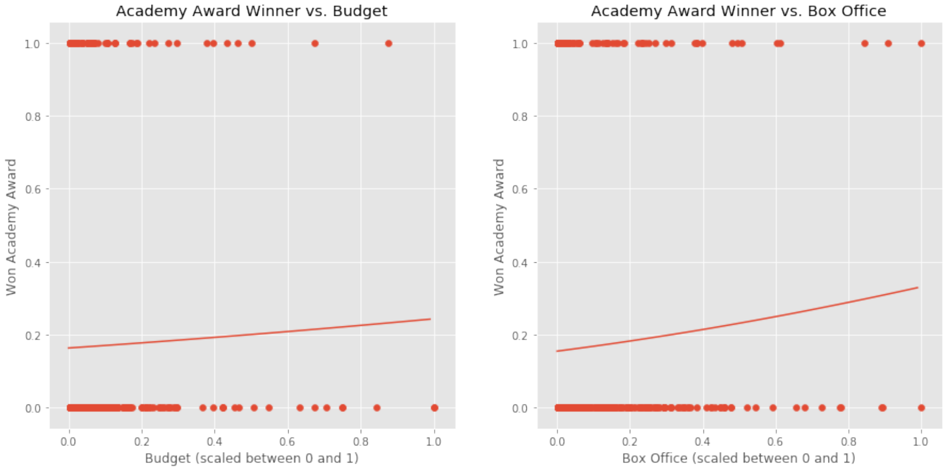

Figure 5: Response vs. Budget and Box Office Return

Figure 5 shows the response variable plotted against budget and box office return. We see that there is a positive relationship between a nominated movie's budget/box office and whether it won the Academy Award; of the two, box office has a stronger relationship. Intuitively, this is a reasonable conclusion since although budget correlates with box office, a movie's box office figure is decided after the movie is released, and is based on the quality of the movie as well as its popularity. Box office appears to also have a stronger relationship with the response than the IMDB rating does.

4 Classification Methods

4.1 Similarity Kernel Construction

To construct the similarity kernel, we use the radial basis function. The kernel between two vectors (which represent movies) is given by

$$k(x_i, x_j) = e^{-\gamma||x_i - x_j||_2^2}$$

with γ set equal to $\frac{1}{\text{no. of features}}$. Similarity submatrices for the training set and test set are created: the training set kernel similarity matrix describes the similarity between movies in the training set, and the test set kernel similarity matrix measures the similarity between the movies in the test set and the movies in the training set.

4.2 Classification results

We consider multiple different machine learning methods to fit and predict with this dataset. Shown below in Figures 6 and 7 are the ROC curves for various methods after tuning relevant parameters (if applicable).

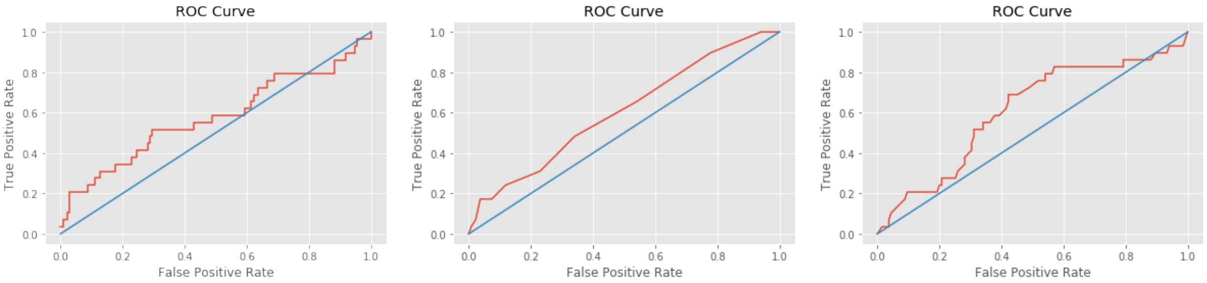

Figure 6: ROC curves for SVM, kNN, and AdaBoost

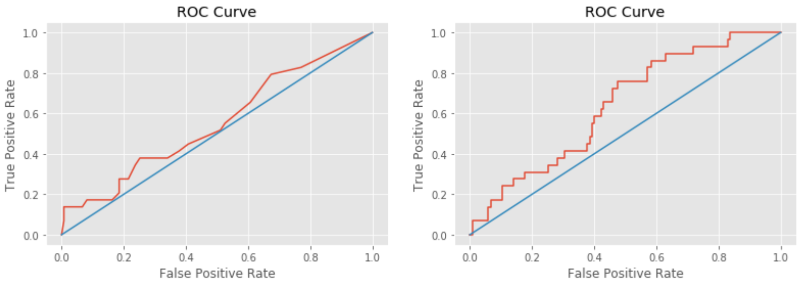

Figure 7: ROC curves for Random Forest and Logistic Regression

4.2.1 Support Vector Machine (SVM)

As seen above, the SVM is not very successful, although it is slightly better than random guessing for most threshold values. Potential challenges that lead to this algorithm's underperformance are noted in the conclusion. Its performance is probably the worst among considered algorithms.

4.2.2 k-Nearest Neighbors (kNN)

The key parameter in this algorithm in the number of neighbors used to determine each movie's class. Most movies in the dataset did not win the Academy Award, so it seems reasonable that this number of neighbors should be relatively large so as to capture a representative group of "nearest" neighbors. This parameter is tuned using leave-one-out (LOO) cross validation; we choose 20 neighbors.

kNN is somewhat successful in predicting Academy Award winners. The algorithm beats random guessing consistently and is more successful than SVM, but of course relies on the tuning parameter.

4.2.3 AdaBoost

The learning rate must be tuned in this algorithm. Empirical risk was not very informative and LOO risk was not computationally feasible, so a 70-30 split of the training set is made. The algorithm is trained on 70% of the trianing set and its performance is evaluated on the remaining 30% by means of mean squared error. It seems that smaller values are superior, and we choose a learning rate of 0.1.

The AdaBoost algorithm is somewhat successful in predicting Academy Award winners. Its ability is comparable to kNN and superior to SVM. The algorithm is consistently better than random guessing, but not by a large margin.

4.2.4 Random Forest

Random Forest algorithms utilize bagging and decision trees to make predictions. The number of trees in this forest must be chosen. Empirical risk is used for this parameter tuning because LOO risk was not computationally feasible. The empirical risk curve indicated that there is little improvement from increasing the number of trees beyond 40.

This algorithm performs better than random guessing. However, like the other algorithms in this project, it only beats it by a small margin. The authors would rank this algorithm's performance as superior to SVM but inferior to kNN and AdaBoost.

4.2.5 Logistic Regression

The authors suggest that logistic regression is the best-performing algorithm out of the algorithms explored here. It does not vastly outperform random guessing, but it is quite successful considering the difficult nature of classification in this setting. As can be seen in the ROC curve, logistic regression outperforms random guessing for all threshold values.

5 Conclusion

The above analysis shows that it is difficult to predict Academy Award winners. Here are some potential challenges in this classification problem:

- Small sample size: Since only one movie wins the award each year, there are few movies to train an algorithm on that are winners. This dataset could be expanded by including non-nominated movies, but these movies would of course not be winners.
- Uniqueness: A primary component of this analysis is the use of a similarity measure. However, often the uniqueness and individuality of a plot or film is the very thing that makes it successful. As a result, it can be difficult to predict whether a movie will be successful by comparing it to other successful movies.
- Emotion: Many variables affect a movie's appeal and success. Many of these are subtle or emotional and therefore difficult to quantify.

Despite these challenges, each algorithm beats random guessing (for most threshold values). There is still some variability in academy award victories that can be explained consistently by these algorithms. However, PR curves reveal that there is still significant room for improvement. Here are some potential areas for improvement in future research and work:

- Larger sample size.
- More advanced algorithms to deal with the class imbalance. Most movies do not win the Academy Award, so there are far more movies that didn't win than those who did win. A more advanced and complex algorithm may deal with this problem more effectively.
- More advanced parameter tuning.
- More detailed plot analysis, perhaps using scripts.

In summary, this classification problem is a difficult one and it present a variety of challenges. Despite this, a variety of binary classification algorithms do show some promise in predicting Academy Award winning movies using a variety of features.

Note: Many figures and commentary were excluded from this document for sake of brevity. A full description of all methodology, figures, and code can be found in the .ipynb file included in our submission.