Predicting Academy Award for Best Picture Winners Using Logistic Regression and Classification Trees

Will White, Tiffany Duong, Andi Zenku, Muhammad Ibrahim Mian, Muhammad Hashmi

Abstract

This paper seeks to investigate whether Best Picture winners for the Academy Awards, otherwise known as the Oscars, can be predicted via classification techniques. We used a dataset from Kaggle consisting of the Best Picture nominees and Winners ranging from the 1st through 93rd (1927-2021) Academy Awards with attributes such as director and production company, as well as their rated scores from RottenTomatoes and IMDB. After observing the relationship between the predictor variables via a Spearman correlation matrix, we made a classification model using logistic regression and classification tree. The data set had an imbalance in the target variable, and Synthetic Minority Oversampling Technique (SMOTE) was utilized to remove it in the training data. For the logistic model, we make use of backward elimination and narrow down our variables to a set of four. Another model was prepared using classification trees. As this is a problem where the minority class is of interest, we use the parameters of precision and recall to judge the usefulness of our model. Using classification trees, we found our model had a precision of 71 percent and recall of 86 percent for the testing data.

Keywords: data analysis, Python, logistic regression, classification trees, film studies

Introduction

The Academy Awards, also known as the Oscars, are an annual awards ceremony honoring movies and individuals within the film industry. Every year, the most coveted award is the “Best Picture” award. During the months leading up to the ceremony, thousands of websites, betting markets, data analysts, and casual film fans seek to predict which film will win this award. There are a litany of factors that observers use in order to do so. Film databases like IMDB and review aggregators like RottenTomatoes, both of which allow users to post film reviews, are often referenced for those seeking to gauge a movie’s Oscars chances. Other factors relating to the film itself, such as director, genre, lead actors and production company, can also potentially be predictive of future Oscars success.

Our paper seeks to determine whether a film’s potential Oscars success can be predicted using data analysis techniques. Using data ranging from the 1st to 93rd (1927-2021) Academy Award for Best Picture winners and nominees, we have used Python to analyze whether RottenTomatoes and IMDB ratings might directly correlate to a film’s chances of winning the industry’s highest award. We hope that our results could be used to develop a classifier for future awards based on multiple predictors such as IMDB and Rotten Tomatoes ratings.

Literature Review

Academics within data science and statistics have written publications on which methodologies are most predictive for Oscars success. Haughton, McLaughlin, Mentzer and Zhang’s research (2015) seeks to determine whether text-based, non-professional reviews of a film can predict its likelihood of Oscar nomination. Their research entailed text mining via SAS Text Miner of both tweeted film reviews and IMDB review data of nine films with Oscar buzz surrounding them. They then used sentiment analysis to classify film reviews expressing positive and negative sentiments, as well as phrases indicating controversial themes. Their findings suggest that in addition to positive reviews, controversy surrounding the movie can be predictive of success.

Jacqueline R. Carlton (2021) attempts to identify different variables (genre, gender, number of previous nominations, etc.) that can potentially predict an Oscar winner for a nominee in each category successfully. Her research uses ANOVA analysis and binary logistic regression to model the data, which she retrieves primarily from IMDB. She did studies to see if age of leading actors/actresses have a higher chance of Oscar, if runtime of the film led to an increase in Oscars, if how often a winner was previously nominated, and if there is a relationship between genre and winning in the top 6 categories. Carlton states that this study is useful for many people because it can increase the chance of a film being more popular and making investors more money.

Like many, D. E. Peacock and G. Hu (2013) use multiple logistic regression and multiple linear regression modeling; however, they also use Maximal Information-based Nonparametric Exploration (MINE) statistics (p. 75). Their logistic regression analysis suggests a high influence from IMDB score and the film being a musical winning the award (D. E. Peacock and G. Hu, 2013, p. 77). In addition to MINE, they use a maximum information coefficient (MIC), which is a new exploratory data analysis indicator about the strength of the relationship between two variables as measured between 0 and 1 (D. E. Peacock and G. Hu, 2013, p. 75). Of all of the data analysis they conducted on the Academy Awards, their MIC and MINE analysis reveals an exponential relationship between the number of screens a film opened on and the films opening-weekend box-office revenue, which is surprising due to it not being a linear relationship (D. E. Peacock and G. Hu, 2013, 78).

Methods

To approach this problem, we utilize two classification techniques among those which we have learned in class: logistic regression and classification trees. There are pros and cons to each method, and the selection of one over the other is done after keeping in mind the problem statement. A disadvantage of classification trees is that they are susceptible to overfitting. This can result in very high-performance metrics for training data sets, but a remarkable decrease for testing data sets and in the real world. This can be solved by using different techniques such as pre-pruning a tree, or post-pruning. Later on we will see if we face this problem or not. On the plus side, they are a powerful and easy tool to solve classification problems. One important distinction to make for this data set is the class imbalance in our response variable. This implies that the number of Award Winner instances (1) is much higher than the Award Nominee (0) class. This imbalance can lead to our model giving us accuracy numbers which are not representative of the result.

As would be evident in the next section, even though our model would give us a very high accuracy, it would not be fulfilling its purpose of correctly classifying award winning films. This high accuracy can give an untrained eye a false sense of security. This dilemma can be catered to with a few different methods which we will discuss in the next section.

Discussion and Results

We performed a Spearman correlation analysis on the variables in our data set. We can see the relationships between the different predictor variables themselves. See Fig 1

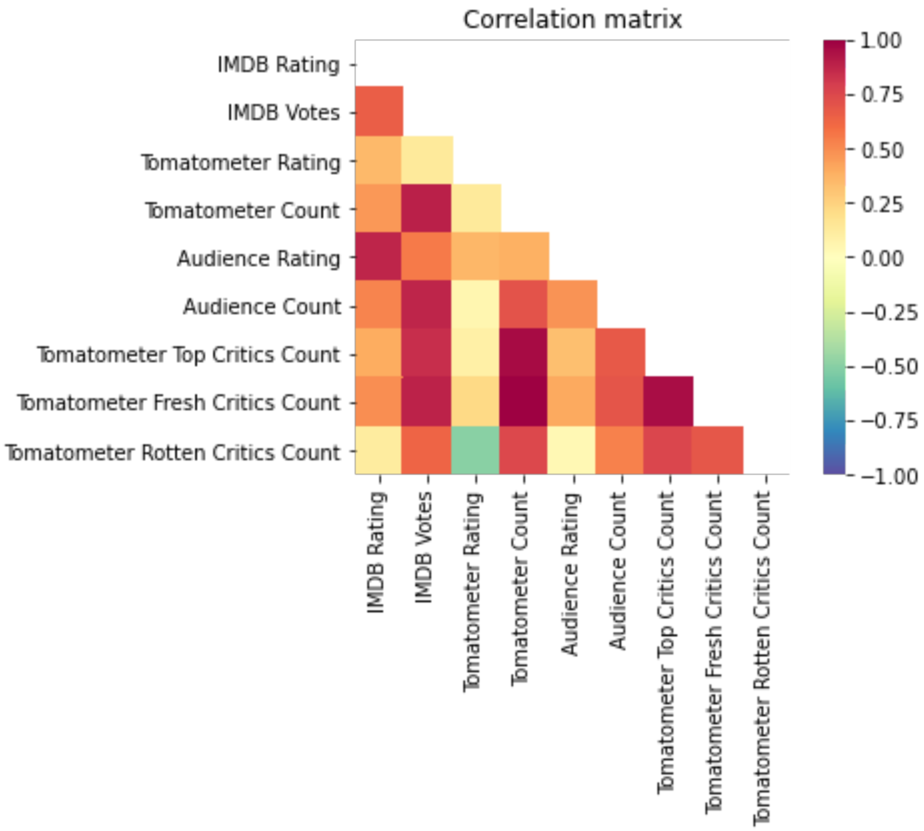
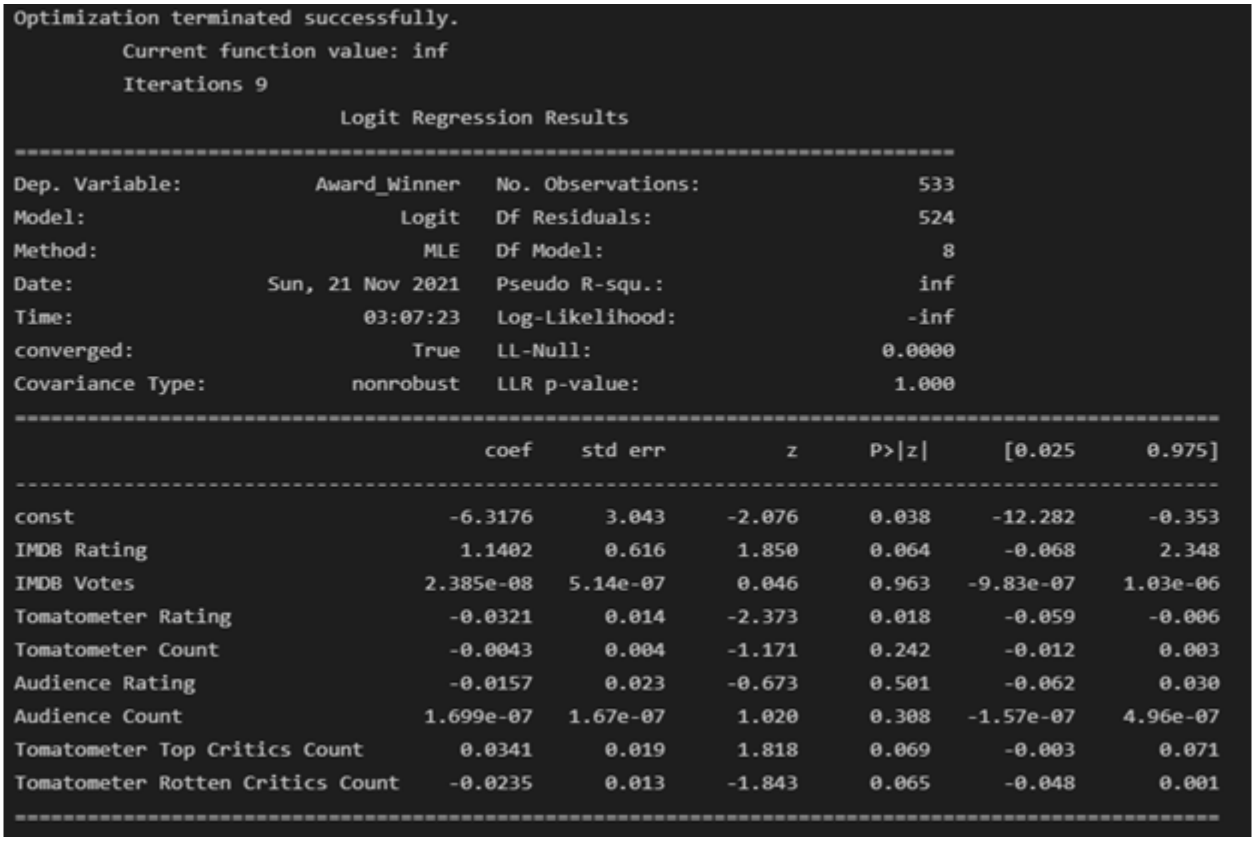


Fig1. Spearman correlation analysis

The first method we utilized is logistic regression. As a first step, we have to resolve the class imbalance problem. This can be done in one of two ways. First is to assign class weights to each class in the response variable. With this method, the model assigns a penalty term to each target class equal to the inverse of its ratio. This means that a class with lower instances will be weighted more heavily as compared to a class which occurs much more frequently. The second approach is to utilize the Synthetic Minority Oversampling Technique (SMOTE), a type of oversampling. In this technique, the minority class is replicated in the training data set to achieve an equal balance between both the classes. We will be using SMOTE as it provides better results and can easily be replicated for our decision trees.

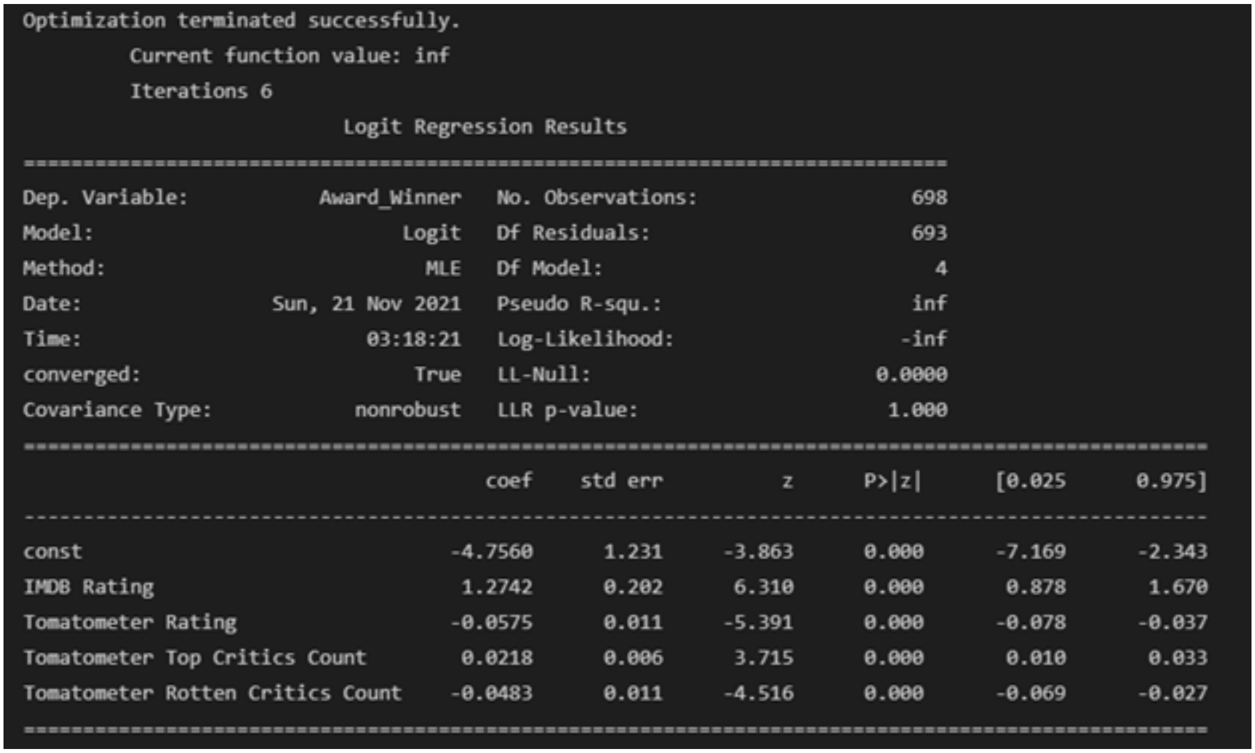
For logistic regression, we adopt a backward elimination approach with our variables. We include all our variables of interest after the exploratory analysis and then observe through the summary statistics which of them are statistically insignificant. Those which are statistically insignificant, i.e. have a p-value greater than the significance level of 0.05 are eliminated, with the variable having the highest p-value eliminated first. The first iteration of the model is shown in Fig 2.



```
Optimization terminated successfully.
         Current function value: inf
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:           Award_Winner   No. Observations:                  533
Model:                          Logit   Df Residuals:                      524
Method:                           MLE   Df Model:                            8
Date:                Sun, 21 Nov 2021   Pseudo R-squ.:                     inf
Time:                        03:07:23   Log-Likelihood:                   -inf
converged:                       True   LL-Null:                        0.0000
Covariance Type:            nonrobust   LLR p-value:                     1.000
=================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                             -6.3176      3.043     -2.076      0.038     -12.282      -0.353
IMDB Rating                        1.1402      0.616      1.850      0.064      -0.068       2.348
IMDB Votes                      2.385e-08   5.14e-07      0.046      0.963   -9.83e-07    1.03e-06
Tomatometer Rating                -0.0321      0.014     -2.373      0.018      -0.059      -0.006
Tomatometer Count                 -0.0043      0.004     -1.171      0.242      -0.012       0.003
Audience Rating                   -0.0157      0.023     -0.673      0.501      -0.062       0.030
Audience Count                  1.699e-07   1.67e-07      1.020      0.308   -1.57e-07    4.96e-07
Tomatometer Top Critics Count      0.0341      0.019      1.818      0.069      -0.003       0.071
Tomatometer Rotten Critics Count  -0.0235      0.013     -1.843      0.065      -0.048       0.001
=================================================================================================
```

Fig 2. First iteration results

This process is repeated until we are left with a model which includes only statistically significant variables. The summary statistics for such a model are shown below, with only the significant variables. The data is now split into a training and testing split. 20% of the data is used for testing while 80% is used for training. For the testing data, SMOTE is applied to remove the class imbalance. The results of the logistic model fit on the training data are shown in Fig 3



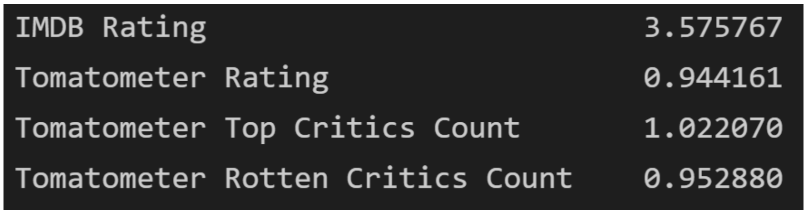
```
Optimization terminated successfully.
         Current function value: inf
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:           Award_Winner   No. Observations:                  698
Model:                          Logit   Df Residuals:                      693
Method:                           MLE   Df Model:                            4
Date:                Sun, 21 Nov 2021   Pseudo R-squ.:                     inf
Time:                        03:18:21   Log-Likelihood:                   -inf
converged:                       True   LL-Null:                        0.0000
Covariance Type:            nonrobust   LLR p-value:                     1.000
====================================================================================================
                                       coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                               -4.7560      1.231     -3.863      0.000      -7.169      -2.343
IMDB Rating                          1.2742      0.202      6.310      0.000       0.878       1.670
Tomatometer Rating                  -0.0575      0.011     -5.391      0.000      -0.078      -0.037
Tomatometer Top Critics Count        0.0218      0.006      3.715      0.000       0.010       0.033
Tomatometer Rotten Critics Count    -0.0483      0.011     -4.516      0.000      -0.069      -0.027
====================================================================================================
```

Fig 3. Logistic model fit on training data

The significant variables are:

- IMDB Rating; 95% CI (0.878, 1.670)
- Tomatometer Rating; 95% CI (-0.078, -0.037)
- Tomatometer Top Critics Count; 95% CI (0.010, 0.033)
- Tomatometer Rotten Critics Count; 95% CI (-0.069, -0.027)

The odds ratios for the significant variables are shown in Fig. 4.



```
IMDB Rating                         3.575767
Tomatometer Rating                  0.944161
Tomatometer Top Critics Count       1.022070
Tomatometer Rotten Critics Count    0.952880
```

Fig 4. Logistic model odds ratios

The data shows that increasing the IMDB rating by 1 unit increases the log odds of a movie being an Oscar winner by 1.2742, considering all else is the same. The predicted classification from our model, using the testing data set, is then compared with the observed classification in the testing data set, using a 0.5 threshold for classification. The confusion matrix on Fig 5:

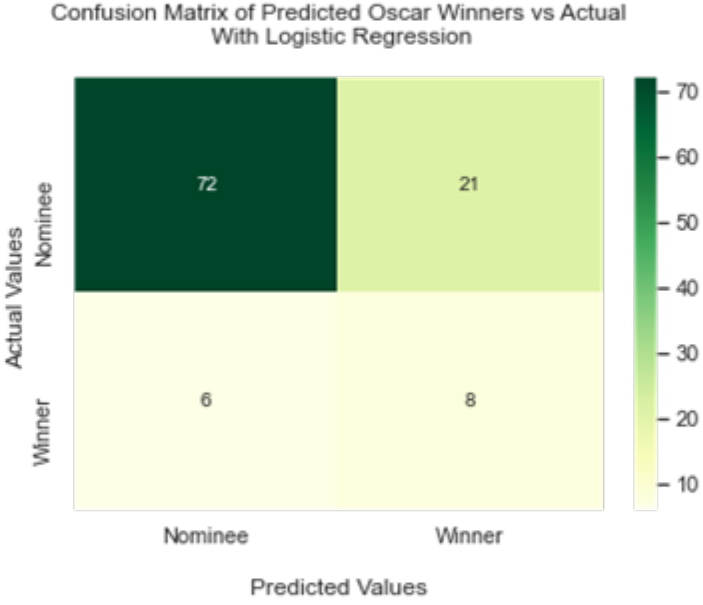


Fig 5. Confusion Matrix

The accuracy score for this model is 74%, but that does not give the whole picture. Even though the accuracy is high, we are still not classifying winners as much as we would want to with our model. This is because while our model has a True Negative Rate, it does not have high precision or recall. Precision in this case tells us the number of movies that we correctly identify as having won an Oscar out of all the movies we predicted of winning it. Recall on the other hand gives us a measure of how many movies we correctly identified as having won an Oscar out of those that actually did win it. It is also known as True Positive Rate.

In cases of class imbalance, and where the minority class is the class of interest, precision, recall, and F1 Score are the important metrics to judge a model by. The metrics for target class 1 for this model are shown in table 1:

Precision	0.27
Recall	0.57
F1 Score	0.37

Table 1. Metrics for target class 1

Going by these statistics, when our model predicts that a movie will win an Oscar, it is correct only 27% of the time. Evidently, this model has sub-par performance. The exact decision whether we want a higher precision or higher recall depends on the nature of the problem and purpose of the classifier. Do we want to bet on a movie winning an Oscar? If so, then it would be wise to achieve high precision, as we do not want to waste our money on bets where there is only a 27% chance of our prediction being correct. In another setting, such as in the field of medicine, we would want higher recall, since we do not want to miss anyone having a disease but not being classified as such.

Now we attempt this problem using a classification tree. For the impurity criterion, we use the entropy. Applying the data classifier on our data set, we get 100% precision and an F1 Score of 1. This means our model far outperforms the logistic model, but this is for training data.

This can very well mean that our model is overfit and would perform poorly on testing data. Let us now apply the model on testing data. The results are shown in table 2:

Precision	0.71
Recall	0.86
F1 Score	0.77

Table 2. Results of model on testing data

Now, whenever our model predicts that a movie will win an Oscar, it is correct 70% of the time. That is a remarkable improvement. Comparing our testing results vs training, we can see that the difference between the training and testing data set is not as drastic as it could have been. Hence, we can generalize this model to real world data and gain meaningful results from it.

The confusion matrix is shown on Fig 6:

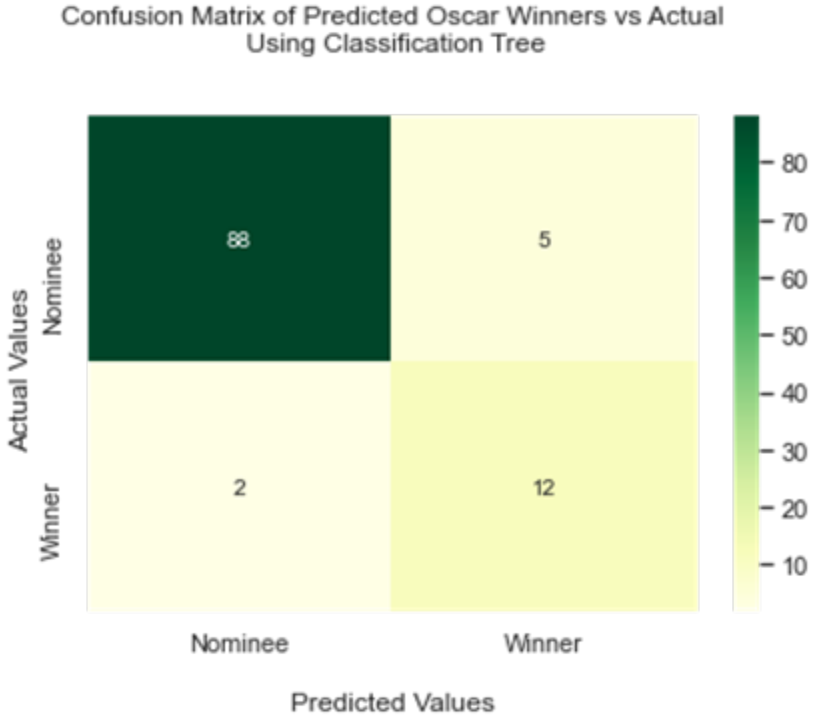


Fig 6. Confusion Matrix

For a decision tree, there are some tradeoffs and modifications that need to be applied. These are mainly to reduce overfit of our model, and also to reduce its complexity. In this model, we have chosen a max tree depth of 15, while minimum sample split is set as 2. The max tree depth is how far our model splits the internal nodes. An extremely high value leads to overfitting, and lack of model generalization, not to mention the complexity. A low number results in underfitting.

The combination of these hyper parameters gives us the best results on our testing data set. This is achieved through an iterative process where multiple hyperparameter combinations are tested to achieve the best combination for our purposes. The precision vs recall curve shown in Fig 7, where AP is the average precision of the model. It is a parameter akin to the AUC of an ROC.

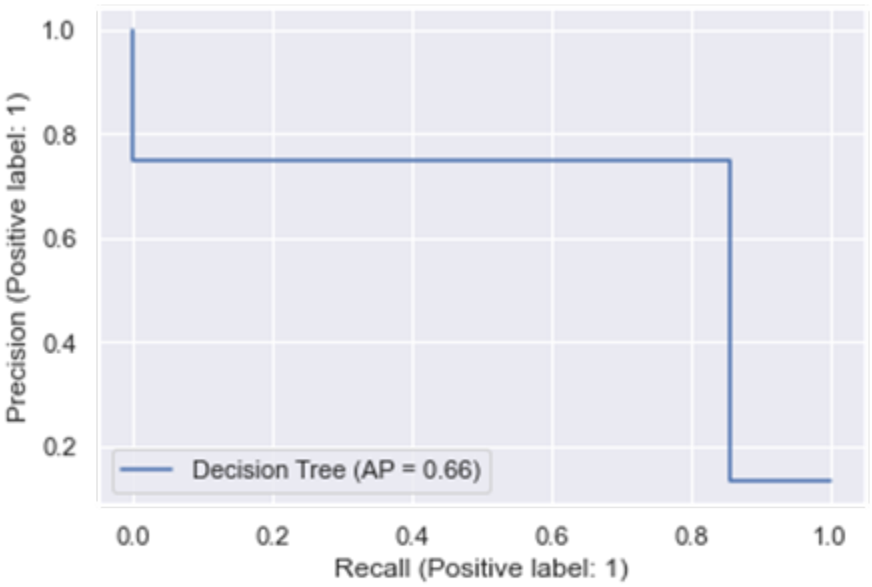


Fig 7. Precision vs Recall Curve

Some limitations of our work include the availability of data and the innate selection bias that using review systems as part of our judgment criteria entails. Alone, there have been a finite number of Academy Awards and only one winner per year. This leaves the actual dataset to be small and presents a bias toward a more revisionist perception of films versus how they may have been reviewed or reviewed when first released. We also ran into incomplete data at times in our dataset, and thus, we had to clean and delete 38 rows of data in order to successfully complete our data analysis. Future work on this model could account for prior awardees or pivot to predicting the awardees of other awards ceremonies such as the Golden Globe Awards or the British Academy of Film and Television Arts (BAFTA) Awards. Accounting for the Golden Globes winners especially could be helpful as they are seen as precursors to the Academy Awards in terms of the awards season in Hollywood.

Conclusion

Based on our research, it can be concluded that several factors can be used to predict Oscar "Best Picture" awardees. Generally speaking, critic ratings differ from audience ratings but are also more highly correlated with Oscar winners. This is according to expectations, since critics rate films independent of the audience perceptions and those in the film industry determine the nominees and winners of the Academy Awards. We hope that our work will shed light onto methodologies for both researchers and the general public to more accurately predict Oscar winners in the future. In terms of the application of this research, betting markets could use our research to predict which movies would win.

References

Carlton, J. R. (2021). *The statistics of the Oscars: What type of nominee will win?* [Thesis, Florida Southern College]. https://repository.flsouthern.edu/handle/11416/548

Haughton, D., McLaughlin, M.-D., Mentzer, K., & Zhang, C. (2015). Can We Predict Oscars from Twitter and Movie Review Data? In D. Haughton, M.-D. McLaughlin, K. Mentzer, & C. Zhang (Eds.), *Movie Analytics: A Hollywood Introduction to Big Data* (pp. 41–54). Springer International Publishing. https://doi.org/10.1007/978-3-319-09426-7_6

Peacock, D. E., & Hu, G. (2013). Analyzing Grammy, Emmy, and Academy Awards Data Using Regression and Maximum Information Coefficient. *2013 Second IIAI International Conference on Advanced Applied Informatics*, 74–79. https://doi.org/10.1109/IIAI-AAI.2013.14.

Page Break

Executive Summary

We’re currently starting to enter Oscars season, with Hollywood beginning to release films vying for awards at the upcoming 93rd Academy Awards next February. Much of the buzz on social media revolves around who will win the Best Picture, the most prestigious award in the film industry. Oscars prediction market can be a lucrative business, with betting market and awards prediction websites like VegasInsider, GoldDerby and PlayUSA. Betting on the Oscars is currently legal in Michigan, Colorado, Indiana and New Jersey, and will likely be expanded to other states in the future as we’ve seen with sports betting. In recent years we’ve seen academics within data science and statistics hypothesize which factors may give films seeking an Oscar the edge, ranging from production company and content rating to user-generated reviews on websites like IMDB and RottenTomatoes. Using the programming language Python, our team has worked on creating a model for predicting Oscar winners that has an accuracy rate of 71 percent. We believe that our research could potentially be used towards developing a tool for placing educated bets in the growing film betting market.

We used a dataset with information on every Best Picture nominee and winner from the first to most recent Academy Awards from the dataset publishing website Kaggle. The data includes basic information about each film like year, director and production company, alongside RottenTomatoes and IMDB reviews. We primarily used two common data modeling techniques: logistic regression and classification trees. Logistic regression lets us model the probability of an event happening, in our case winning an Oscar. Our model using this method assigned Best Picture winners a 1 and nominees a 0. We then used a classification tree, which lets us make predictions with our data by dividing it into a series of hypothetical pathways, to see what percentage of the time we can correctly predict a Best Picture winner.

Our results are straightforward in nature, showing that by training a model as such on a random subgroup of our data, we were able to create a model that predicts Best Picture winners for the rest of the data with 71% precision and 86% accuracy. Furthermore, our work reveals that critic opinion on movies more directly correlates with Academy Awards success than that of audience or public opinion, which includes the presence of "top critic" opinion on a movie at all. As such, our results were expected as the Academy Awards nominees and winners are determined by those in the film industry versus the general public.

Some limitations of our work include the availability of data and the innate selection bias that using review systems as part of our judgment criteria entails. Alone, there have been a finite number of Academy Awards and only one winner per year, leaving the actual dataset to be small. Additionally, the Academy Awards have been ongoing since the late 1920s, but review systems such as Rotten Tomatoes and IMDB have only existed since the 1990s. This presents a bias toward a more revisionist perception of films versus how they may have been reviewed or reviewed when first released. We also ran into incomplete data at times in our dataset, and thus, we had to clean and delete 38 rows of data in order to successfully complete our data analysis. Future work on this model could account for prior awardees or pivot to predicting the awardees of other awards ceremonies such as the Golden Globe Awards or the British Academy of Film and Television Arts (BAFTA) Awards. Accounting for the Golden Globes winners especially could be helpful as they are seen as precursors to the Academy Awards in terms of the awards season in Hollywood.

Using logistic regression and classification trees, we were able to successfully generate a model that can predict Best Picture winners with high precision and recall based on background information and ratings of models. This model can be used in order to bet or predict winners for recreational as well as financial hedging reasons. Further exploration may expand this model with inputs from other machine learning techniques or to more heavily value other factors in addition to precision versus recall.