The Effect of Reviews on Best Picture Winners

Dante Testini
dtestini@ucsd.edu
UC San Diego
San Diego, USA

Abstract

This paper looks into whether reviews alone can be used to predict Best Picture Winners throughout history. Utilizing machine learning models, namely Decision Trees, with different amounts of features (all relating to reviews and opinion), this paper compares accuracy metrics to make conclusions on the effect of reviews on a movie's chances of receiving Best Picture. Additionally, this paper also looks at the factor of competition, using calculated features to conclude whether the reviews of the other movies in a given year have an impact on a movie's success at the Oscars.

1 Introduction

With the 2025 Oscars happening a couple weeks before the writing of this paper, everyone tries to predict who is going to win before it even happens. Best Picture, the most coveted award, is the biggest target of this trend, as every person has an opinion on who could, should and will win. There are many aspects that make a movie have the quality to win, but most of these aspects can be opinion and immeasurable on the surface. On the other hand, there are many aspects that make a movie more likely to win, based on history and the structure of voting. These factors are much easier to quantify, such as the film length, rating or notoriety of the actors, directors or studio. This paper is going to look into those first set of things - the opinion-based factors of a movie - and see whether these are good predictors of a movie's success at the Oscars, namely the best picture category. This will be done by looking at the reviews of movies, namely via IMDb and Rotten Tomatoes scores, to predict the Best Picture Winner for every year.

Additionally, the *competition* of movies in a given year can vary. One movie that people agree upon should win in one year, might have tough competition in its given year and lose anyway. On the other hand, some movies that win in one year might only win because its just better than the other movies in its year, but would not even be considered in other years. In sports terms, this is considered a *strength of schedule*, and is frequently considered when formulating brackets for tournaments. This paper will also utilize this concept in its feature creation, analyzing whether this is a real phenomenon in movies, and a good predictor of Oscar success.

DSC148, San Diego, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Literature

Predicting award results is not a new thing. In fact, numerous people and publications, of both machine learning and movie backgrounds, have been attempting this same thing with various degrees of success. All of these reports utilize every feature possible, trying to predict correctly year after year. One of the best examples of this is in a blog by Mike H. White III,[1], where he predicts the best picture with 77% accuracy. He utilizes xgboost, and has cast a very large net in grabbing features, from 'if the cinematographer won an award' to 'whether the movie is a book adaptation'.

Another publication worth noting is from Sonkaya and Yalcin, [2] who used the same datasets this paper will use. They utilize similar machine learning models that are used in this paper, achieving applaudable accuracy. The goal of this paper's models is not to compete with those, as that would be redundant. The goal of this paper is to look at the connections between reviews and strength of schedule as predictors, removing other features that might increase accuracy, but take apart from the answers to those questions. Despite others using the same dataset or trying to predict the same outcome, this goal is unique, and will analyze these trends on its own.

3 Predictive Task

The predictive task of this paper is to predict which movie wins Best Picture at the Oscars utilizing its reviews both by themselves, and in comparison to its competition. The models presented will be compared to a baseline *random* model, which predicts the Best Picture winner each year at random. The models will be evaluated based on Accuracy of Correct Guesses. The models are set to always pick one winner per year, and the correctly predicted years will be divided by the total years in the set. Both train and test accuracies for each model will be recorded based on randomly splitting the years with 80% and 20% splits. These models will not be compared to existing models extensively, as their test accuracies will all score less than the 77% from White III [1].

4 Datasets

4.1 Dataset 1

The first dataset used is from Martin Mraz at https://www.kaggle.com/datasets/martinmraz07/oscar-movies/data. This dataset contains 571 rows, each containing a Best Picture nominatated movie from 1928 to 2020. It also contains 29 columns, including statistics from IMDb and Rotten Tomatoes. A quick glance through this dataset reveals some non-ideal results, as there are 23% of rows without Rotten Tomatoes data. This, along with the restraints of only using Review-related data, brings only a handful of relevant columns: Movie Name, Oscar Year, IMDb Rating, IMDB Votes and whether or not the movie won Best Picture.

An Exploratory Data Analysis does show that there are differences in IMDb data between winners and nominees (or non-winners). Looking at the histograms in Figure 1 and Figure 2, differences between the distribution of IMDb data in Winners and Nominees are apparent, with the Winner's distribution of Ratings peaking past 8.0 while very few non-winners getting past 8. Number of Votes also show trends, (with most of the movies being on the high end of votes being winners), but appear to be less of a predictor than Ratings.

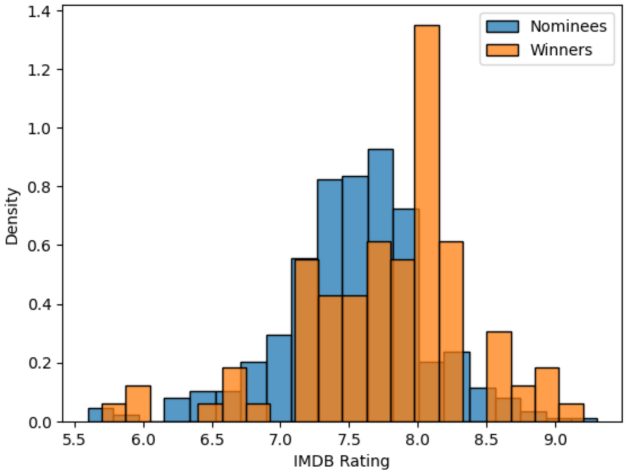

Figure 1: Histogram of IMDb Ratings

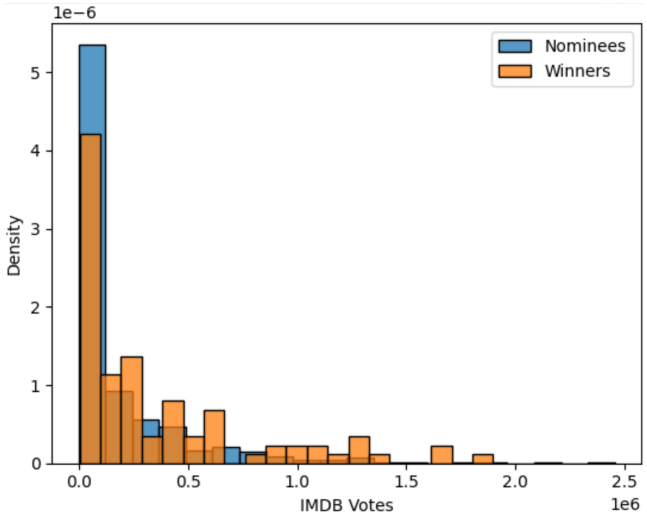

Figure 2: Histogram of IMDb Votes

The other focus of this paper is strength of schedule. Can winners win based on having an *easy* year of movies against them? To look at this visually, line graphs were made to show the average review statistic over time, and then the winners statistic layered on top. The results from Figure 3 are hard to read, as the Winner's IMDb Rating appears to fluctuate a lot, but there are some trends that can be made out. For instance, the IMDb ratings, especially after 1970, are generally higher than the average ratings of the nominees. In Figure 4, this trend is even more clear, with almost no Winner receiving a less than average number of votes for its given year. The concept of *easy* and *hard* years can be seen in both of these graphs too, as the average statistic in both of these fluctuate a lot, especially when looking at older movies versus newer movies.

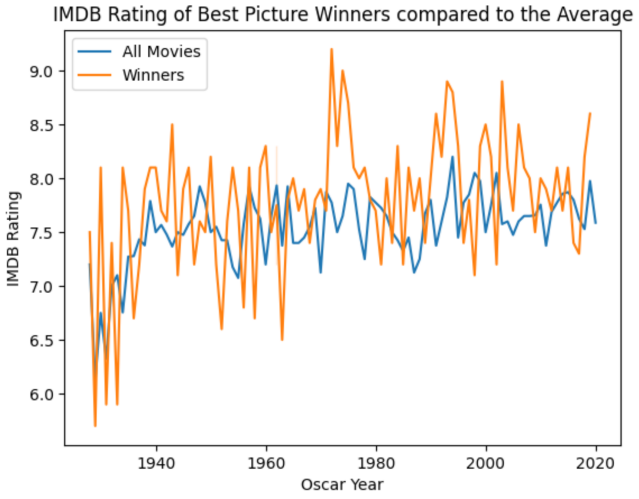

Figure 3: IMDb Ratings over Time

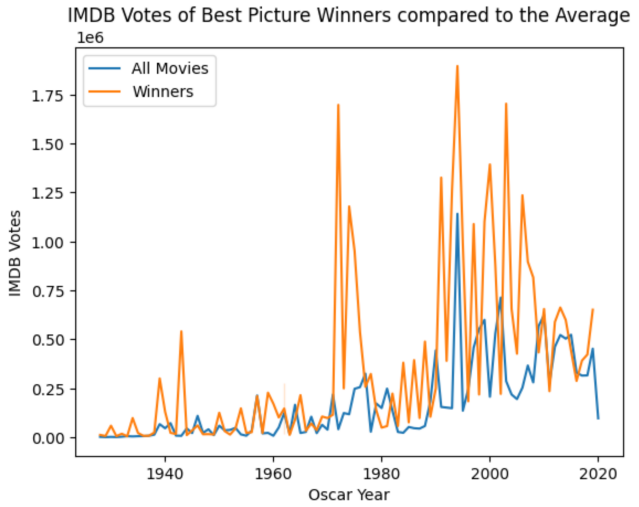

Figure 4: IMDb Votes over Time

4.2 Models

For this classification problem, this paper uses regression models. This appears counterintuitive, but the regression results are sent into custom functions to pick the highest value (or probability) for each year, similar to a classification cutoff in Logistic Regression, except constrained to years. This is to ensure that every year has

one, and only one, predicted winner. The train and test accuracies are then calculated with this data, counting how many years the model guessed correctly.

When testing with classification models, the model would look at each movie and say whether the movie is or is not a winner, with no context of the year the movie is in. This would result in many years with multiple predicted winners, and some with none.

The main model utilized in this analysis is a Decision Tree Regression model. This model was chosen as a middle ground between too complex (compared to a Random Forest) and not complex enough (compared to Linear Regression). Hyperparameter tuning was done manually, and not extensively selecting the `max_depth` hyperparameter based on which received similar train and test accuracies. This rudimentary hyperparameter tuning is not ideal for achieving the best possible overall fit, but it works to compare the models amongst each other.

4.2.1 Model 1. The first model was a Decision Tree Regression model with depth of 5, using just the IMDb Ratings and the IMDb votes of each movie.

4.2.2 Model 2. The second model was a Decision Tree Regression model with depth of 5 utilizing the features from Model 1, along with competition features. The competition features were the Opponent's Average IMDb Rating and the Opponent's Average IMDb votes. These, as the name suggest, were calculated by looking at the opposing movies in each year and calculating the average of that statistic. This is slightly different than the value looked at in the line graphs, as these values do not take in the movie's scores itself into the average. These values are supposed to represent how good or bad the competition is for each year.

4.3 Dataset 2

The first dataset was limited, mainly due to the missing values making some of the data unusable, even if they would have been imputed. Sonkaya and Yalcin found this dataset limiting too [2], so they found fuller one by Máté Váradi at https://github.com/MateVaradi/OscarPrediction/blob/main/data/oscardata_bestpicture.csv. This is the dataset used for all subsequent models. It contains 380 rows for each nominated Best Picture movie from 1961 to 2024. It also contains 60 columns, all already quantified to be used in a model, including no relevant columns with missing values. Columns containing statistics about the movie itself will be ignored for this paper, such as Genre and Rating. Columns with regards to non-Oscar awards will also be ignored and, for simplicity sake, the only Oscar-related feature will be number of Oscars nominations. Although the number of nominations is not a review per se, it is a quantitative representation of the opinions of a group of people, which is what this paper is studying. In final, the relevant columns are film name, Oscar year, IMDb rating, Rotten Tomatoes Audience Score, Rotten Tomatoes Critics Score, Oscar Nominations and whether or not the movie won Best Picture.

No extensive EDA was taken when looking at Best Picture Winners, as it is similar to the first dataset, but additional features adds the element of between-features correlation. The correlation matrix can be seen in Table 1, where the IMDb and Rotten Tomatoes scores are all positive correlated to each other. They are also relatively strong correlations, especially IMDb and Audience scores, showing that these reviewers share similar thoughts on these movies. Interestingly, the number of nominations correlation is very weak, and even negative in some instances, showing that this opinion feature might be more unique than the others.

Table 1: Correlation Matrix of New Features

	IMDb	rtcritic	rtaudience	totalnoms
IMDb	1.000000	0.349038	0.717632	0.161437
rtcritic	0.349038	1.000000	0.367754	-0.082430
rtaudience	0.717632	0.367754	1.000000	0.048464
totalnoms	0.161437	-0.082430	0.048464	1.000000

NOTE: If others try to use this dataset, note there is an error in the 2018 results, as this dataset has no winner. The Shape of Water won Best Picture that year, and this was fixed in the data before models were run on it. Other data was not fact-checked, but is assumed to be correct, despite this minor error.

4.4 Models

4.4.1 Model 3. The third model mirrors the first model, using a Decision Tree Regressor with depth of 3 and the selected review features. These features include IMDb rating, Rotten Tomatoes Audience Score, Rotten Tomatoes Critics Score and Number of Oscar Nominations.

4.4.2 Model 4. The fourth model mirrors the second model, using a Decision Tree Regressor with depth of 3 and the features from the third model added onto calculated competition features. These features, being Opponent's Average IMDb Rating, Opponent's Average Rotten Tomatoes Audience Score, Opponent's Average Rotten Tomatoes Critics Score and Opponent's Average Oscar Nominations, were calculated exactly like the ones from the previous dataset: by averaging the statistic of the other movies in the given year.

4.5 More Comparison Models

Two other models were created to compare the predictive accuracy of the main four. These models were a Gradient Boosting Regressor and a Random Guesser. Upon completion of the other models, a boosted model was used as it is similar to the model that White III used. [1] This allows that model to be compared more directly with his 77%. The Random model was used to get a baseline, to see the predictive power of reviews of Best Picture winners.

4.5.1 Boosting. This model utilized `sklearn's` Gradient Boosting Regressor with a `n_estimators` of 15 to make these predictions. This parameter was chosen from hyperparameter tuning, which also revealed that this model has overfitting issues. This model also used the same features from the fourth model, as it has the most features.

4.5.2 Random. This model randomly created probabilities for each movie, and then sent those probabilities into the usual processing that finds the highest one for each year and *predicts* that film. This

utilized the second datasets movies, which is only relevant to how many years it predicted, and how many movies were in each year.

5 Results

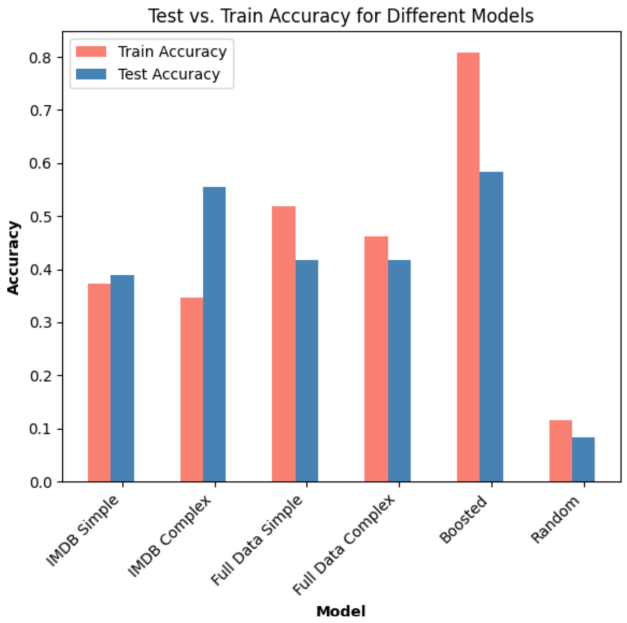

Figure 5: Test and Train Results

6 Discussion

6.1 Limitations

The results are quite inconsistent, with some models even having test accuracies higher than their train accuracies. This is mainly due to the size of the dataset. Since the accuracies were only calculated on a yearly basis, a test set of 20% of the hundreds of rows is really only 20% of the 60 to 90 years (depending on the dataset). For example, the test accuracy of Model 2 is 55%, because it got 10/18 correct predictions in the test set. These small test sets make the variance high. There are many other limitations in this study, but the lack of complete and full data is the biggest.

6.2 Dataset Differences

The first dataset was used on Models 1 and 2 while the second was used on the rest. This can maybe be seen as the accuracies appear to be higher on average on models 3 and 4 than in models 1 and 2. This would make sense, as there are more features in the second dataset, and the first is limited to just one review website. On the other hand, the accuracies do not go significantly up. The simple model's test accuracy goes from 39% to 41%. This is not a stark increase, especially comparing the difference between these models and the boosted or random models. Overall, it is inconclusive on whether these new features had a serious impact on the prediction accuracy, as these scores frequently mirror each other. In the correlation matrix from earlier, the three review scores all had positive and relatively strong correlation between themselves. The opinions on IMDb do not severely change from opinions on Rotten Tomatoes, and therefore, one website's opinion source might be all that is needed to gather a consensus.

6.3 Effect of Competition

To analyze the effect of competition, or strength of schedule, the *simple* and *complex* models can be compared to each other, as the models labeled 'complex' include the added competition statistic. Overall, this does not seem to be a noteworthy effect. In fact, when looking at model 3 and 4, the only noticeable difference between the two is that the train accuracy *decreased*. The competition features appear to have not helped the model at all, instead only muddying up the data to less reliable results.

6.4 Effect of Reviews

To analyze the effect of reviews, the overall success of the first five models can be compared to the *random* model. With this, we see a clear difference, as the random model is getting 11% and 8% accuracies, far below the range of the other models. This shows that reviews are a trustworthy predictor of whether a movie wins Best Picture. Looking specifically at the boosted model, we see a test accuracy of 58%, with only a handful of features. This model does suffer from overfitting due to its complexity as noted earlier, but it is fair to say that perhaps with a fuller dataset, this model could arrive close to the 77% of the model mentioned earlier, which has numerous features from every aspect of the movie's composition. Overall, reviews appear to be a good predictor of a Best Picture winner.

7 Conclusion

Good movies are hard to define, but it is not impossible to predict a movie's success at the Oscars. This paper has proven that opinion, measured in reviews, is a valuable path to quantify quality, resulting in promising accuracies from history. On the other hand, strength of schedule did not get a promising result that would label it as a strong predictive factor. In fact, the competition of a movie appeared to have no effect on its likelihood to be named Best Picture. Nonetheless, the power of reviews alone were able to predict Best Picture Winners despite this. So next time you are curious what movie is going to win big at the Oscars, instead of relying on complicated metrics, assumptions and data, you might be able to just take a look at the reviews.

References

[1] Mike H. White III. 2024. Modeling the Oscar for Best Picture. (March 2024). https://www.markhw.com/blog/oscars2024

[2] Seher Zeynep Sonkaya and Ipek Doga Yalcin. 2025. Predicting Oscar Best Picture Winners. (Jan. 2025). https://www.kaggle.com/code/zzzz07/predicting-oscar-picture-winners#8.-Solution:-New-Dataset