THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

SCHOOL OF SCIENCE

And the Oscar Goes to… (An Application and Comparison of Models Used to Predict the Winner of the Academy Award for Best Picture)

JOSHUA SLAYTON
SPRING 2021

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Mathematics
with honors in Statistics

Reviewed and approved* by the following:

Michael Rutter
Associate Professor of Statistics
Thesis Supervisor and Honors Advisor

Terry Blakney
Associate Teaching Professor of Statistics
Faculty Reader

* Electronic approvals are on file.

ABSTRACT

Each year, the Academy of Motion Picture Arts and Sciences recognizes exceptional achievements in cinema with its Oscars ceremony. Of all the awards handed out at this annual celebration, the Academy Award for Best Picture is arguably the most coveted. This project involves using historical Oscars data, as well as data from other awards shows such as the Golden Globes, to develop and compare models for predicting the winners of the Academy Award for Best Picture. In particular, models using both logistic regression and decision tree classification will be developed. Model performance will be evaluated using the leave-one-out cross validation procedure to compute prediction accuracies and root mean square errors for each model. These measures will allow direct comparisons between the models to be made, which should lead to some interesting results. For example, models that incorporate more than just Oscar nominations data perform better, unsurprisingly, than those that do not. Also, as will be discussed, the decision tree classification models all have higher prediction accuracies than their corresponding logistic regression models. The predictions that the "best" logistic regression and decision tree classification models got most incorrect will be explored. Finally, the models generated through this project will be applied to make predictions for this year's Best Picture winner, to be announced on April 25, 2021.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank Dr. Michael Rutter for agreeing to work on this project with me. Your guidance and insight were invaluable, and I would not have completed this project without you. You have equipped me with knowledge that I will carry with me for the rest of my life.

I would also like to thank Professor Terry Blakney for giving me a solid foundation in the academic study of statistics and for expressing a level of enthusiasm that cannot be matched. I will always remember walking into your classroom and walking out feeling a renewed sense of energy.

Lastly, I would like to thank my mom, Jennifer, for being there with me through all of the highs and the lows (we both know there were many of them). Your unconditional love and support have allowed me to develop into the person I am today, and I hope I will continue to make you proud for years to come.

Chapter 1

Introduction

This chapter will begin with a discussion of the cultural importance of cinema before exploring the history of the Academy Awards, the ultimate celebration and recognition of cinematic achievement.

The Cultural Importance of Cinema

Cinema, like other artforms, has both reflected and helped to shape cultures and human experience. Though it can never capture all of the complexities of reality, cinema can present a mediated understanding of the world we live in or one that was once lived in. As such, it functions to preserve things of the past and things of the present for consumption by future audiences. Sometimes cinema is meant to remind us of the past or to reflect the present, but other times, it can help guide us through our individual futures or perhaps even through our collective future as a society. It is rare to find a film that achieves the latter effect, but it is certainly possible. I have found a couple of gems that, after watching, continue to influence the choices I make and the path I take in life.

Cinema is also important in that it can help us learn about things with which we may have little to no experience, or it can allow us to look at something we know a bit about, but through an entirely new perspective. The brilliant 2016 film, *Moonlight*, winner of the Academy Award for Best Picture, is a perfect example of a film that explores a human experience not many are familiar with, that of a young, gay, black man growing up in a poor neighborhood in

Miami. The film provides insight into the intersecting oppressions of race, class, and sexuality. Films like *Moonlight* can broaden awareness and understanding of differences among individuals, communities, and cultures. Greater exposure to these differences can, in turn, have real and lasting positive effects on society. Often times, discrimination and violence are motivated by a failure or refusal to acknowledge or appreciate individual differences, whether it be from a lack of a quality education, or beliefs passed down through families, or a general inexperience interacting with diverse groups of people.

Movies have the ability to spark conversations, which can inspire change, but for some, they are simply a source of entertainment, a temporary escape from day-to-day life. Movies can make us laugh, excite us, make us cry, or scare us, and there is enjoyment to be found in each one of those emotional responses. No matter what draws people to movies or how movies affect people, cinema holds an important place within culture.

A Brief History of The Academy Awards

In 1927, the International Academy of Motion Picture Arts and Sciences was founded, following a proposal presented at the Los Angeles Ambassador Hotel to 36 professionals from across all branches of the film industry ("Academy Story"). The idea for this organization came from Metro-Goldwyn-Mayer executive, Louis B. Mayer. The first Academy Awards ceremony took place in May 1929, recognizing films released in 1927 and 1928. Recipients of the very first Academy Awards were announced months before the actual ceremony, though this changed the very next year when the results were kept secret. After the ceremony in 1940, the sealed-envelope setup used today was implemented. Throughout the 1930s and 1940s, additional

awards were added to the lineup, including awards for best editing, song, acting in a supporting role, special effects, and documentary, among others. The Academy Awards ceremony was first televised, in black and white, to U.S. and Canadian audiences in 1953, and in color in 1966. In 1968, the Academy Awards ceremony was postponed a few days due to the assassination of Dr. Martin Luther King Jr. Another postponement occurred again in 1981, after the assassination attempt on President Ronald Reagan. In 1969, the ceremony was broadcast to an international audience for the first time. Over the following decades, additional awards continued to be added, including awards for best makeup and animated film. Figure 1 below shows the Oscar statuettes that are handed to the winners of Academy Awards each year.

Figure 1. Oscar Statuettes, Academy of Motion Picture Arts and Sciences

In 2012, plans were announced to build a museum dedicated to the film industry; the museum, now named The Academy Museum of Motion Pictures, is set to open on September 30, 2021 (Academy Museum). In recent years, the Academy has been rightfully criticized for the lack of representation in both the nominees and in its membership/voting body. The hashtag #OscarsSoWhite, along with backlash for failing to recognize the incredible achievements of female directors throughout history, and most recently Greta Gerwig for her work on *Little*

Women, has caused quite a shift in recent years. There was a successful push to add more diversity to the Academy's membership, and nominations since these criticisms were voiced have improved in terms of racial and gender diversity. The amount of diversity seen in this year's nominees is very exciting and hopefully something that will continue moving forward. The Academy Awards have long recognized the importance of cinema and the incredible achievements that have been made in film. The goal of this project is to use historical data from the Academy Awards and statistical methods to develop models which can be used to predict Best Picture winners.

Chapter 2

Methods

This chapter describes the decisions that go into determining which data to collect, the process of collecting the data, and the statistical methods that are used to analyze the data. The latter section explores the following: building models using logistic regression and decision tree classification, evaluating and comparing model performance upon running leave-one-out cross validation, and finally, performing model selection and discussing why this is not very useful for this project.

Deciding Which Data to Collect

The first stage of this project is determining which data to collect. These decisions are naturally influenced by the intent of the project, namely to develop and compare models that can be used to predict the winner of the Best Picture Academy Award. Given the project objective, it

reasons that Oscars data should be considered. More specifically, only nominations for the various Academy Awards will be included because the models are intended to generate predictions before the Oscars ceremony takes place and thus, before any winners are announced. The categories excluded from this analysis are Best Animated Feature Film, Best Animated Short Film, Best Documentary Feature Film, Best Documentary Short Subject, Best International Film, Best Live Action Short Film, Best Original Score, Best Sound Editing, and Best Sound Mixing. The first six of these were excluded because they are not applicable to every film; in other words, unlike with a category such as Best Directing, in which every film could be nominated, only select films can be considered for Best Animated Feature Film. The award for Original Score, while one of the bigger awards, was excluded from this analysis because of its complicated history. There are multiple awards under the umbrella of Best Original Score and those awards have not been consistent over time. The sound categories were excluded because they are technical awards and were not expected to be influential on the outcome of the race for Best Picture. However, a future extension of this project might be to include a variable that considers whether a film was nominated for a technical sound award.

In addition to Academy Awards data, the data from other major awards ceremonies that celebrate achievements in cinema will be collected. More specifically, the British Academy of Film and Television Arts (BAFTA) award for Best Film, the Golden Globes award for Best Film (accounting for the division between Drama and Musical or Comedy), and various Guild awards distributed by bodies such as the Directors Guild of America (DGA), the Screen Actors Guild (SAG), the Writers Guild of America (WGA), and the Producers Guild of America (PGA), will be collected. With these external awards, variables will be included for both nominations and wins (since these awards are handed out prior to the night of the Oscars). Of course, there are

many other awards given out, particularly at film festivals like the Cannes Film Festival, the Sundance Film Festival, and the Toronto International Film Festival; however, the ones that have been included in the models seem, from a qualitative analysis, to be the most relevant. Appendix A includes all explanatory variables that will be used in this analysis, along with descriptions of the variables, the years for which data was collected, and the models in which the variables were used.

Other seemingly valuable data, such as box office numbers and film reviews (both from critics and the audience), have been considered but will not be included in this project. Box office data are difficult to compare across time, considering the effects of inflation, competition in terms of the number of films being shown as well as alternative options for entertainment, the number of screens showing movies, etc. Also, since this project is interested in predicting the Best Picture winner before the Academy Awards ceremony takes place each year, the portion of a film's box office gross before the Academy Awards ceremonies each year would need to be determined if that variable were to be included, and that data is not easily attainable. As for critic and audience reviews, there would need to be some way to take into consideration only those reviews released before the Academy Awards take place so as to remove any potential bias. This is simply not feasible. Also, by including reviews, there will arise an issue with missing values for films that may not have any reviews listed on review amalgamation sites such as Rotten Tomatoes and Metacritic. Reviews are subjective as well, and the scales on which films are rated are variable. Thus, only awards data will be considered in this project.

Data Collection

This next stage of the project will examine how the R programming language can be used to effectively scrape relevant data from online sources (R Core Team). Though other sources exist, data will be scraped from Wikipedia, a source highly conducive to the data scraping process because it organizes the data into readable tables.

To perform the data scraping, several functions from the "rvest" package in R were used (Wickham, "rvest"). The specific purpose of the "rvest" package is to facilitate the scraping of webpages. It contains a number of useful functions. The first is "read_html()," which reads in a webpage provided the webpage's URL is given as an input. The next is the "html_node()" function, which is used to select specific elements or "nodes" from an HTML source. An inspection of the appropriate Wikipedia webpages is necessary to identify the useful nodes. Then, the "xpath" for each node is determined and fed it into the "html_node()" function. Another useful function from this package is "html_table()," which parses tables available on an HTML source into a data frame. These three functions facilitate the data collection process by allowing tables from the appropriate webpages to be scraped and to be accessible in the R environment. A sample of the code used to scrape the data for the Academy Award for Best Picture from Wikipedia is included in Appendix B.

Statistical Methods

Once the data has been collected, the logical next step would be to explore and understand the data. Upon gaining a clearer understanding of the data, statistical methods can be applied to develop meaningful results.

Since this project involves using several variables with known values to gain information on a variable whose value is unknown, it reasons that regression analysis would be useful. There are many different forms of regression that apply under unique circumstances. Arguably the simplest form of regression is called simple linear regression (SLR), which relates a continuous, quantitative response variable to a quantitative predictor variable. If, however, the response variable is not continuous but instead dichotomous, or rather, binary, then another form of regression, called logistic regression, should be applied.

Logistic regression, unlike SLR, produces probabilities, p, which depend on the value of a predictor variable, x. This relationship between p and x is given by the logit function (Devore 557):

$$p = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

With some algebra, the equation becomes the following:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x$$

The expression on the left side of the equation is called the log odds ratio. By exponentiating the log odds ratio, you can determine the odds, which is a measure of how likely a "success" is relative to a "failure." In this project, a film winning the Academy Award for Best Picture is considered a "success" and is assigned a value of 1. It follows that a film which does not win Best Picture is classified as a "failure" and is assigned a value of 0. These are the only two options for the response variable, hence why logistic regression applies. Furthermore, the parameters β_0 and β_1 in the equation above are estimated using the maximum likelihood technique, though R does this process automatically (Devore 558). Just as the ideas behind simple linear regression can be extended to consider multiple predictor variables, thus leading to

multiple linear regression, logistic regression can also be extended to include multiple predictor variables. In this case, if there are n predictor variables, the relevant equation becomes:

$$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

In this project, each predictor variable, x_i, is a binary variable with a value of either 0 or 1. For example, one variable present in this project's main data set is "NominatedBE," which keeps track of whether a film was nominated for Best Editing at the Academy Awards. There are only two options: the film was nominated (assigned a value of 1) or it was not nominated (assigned a value of 0).

Once the method by which to relate the variables has been identified, another consideration can be made, namely which variables are worth including in the model. The process of identifying specific combinations of predictor variables that achieve some statistical criterium is referred to as model selection. There are many criteria that could be applied, from p-value to AIC, as well as many model selection methods to choose from, such as forward selection, backward selection, and best subsets. Though backward model selection using three different criteria (p-value, AIC, and BIC) was performed, it was not at all valuable to what this project is meant to accomplish. It confirmed that directing, editing, and screenplay nominations are influential to a film's chances of winning Best Picture. However, limiting the number of predictor variables to just those three created many problems. Quite often, multiple films nominated for Best Picture in any given year are nominated in all three of these categories as well, which gives them equal probabilities of winning according to the model. Thus, rather than trying to identify which variables are most significant, this project will ignore the model selection results and will instead examine how different models compare to each other when all applicable variables are included.

In addition to logistic regression, the other modeling technique used in this project is decision tree classification. A decision tree is a type of flowchart that contains elements called nodes and branches (Le). Nodes can be thought of as the checkpoints at which observations are classified before being sent along the appropriate path. The root node is the node in which all observations begin. The root node will then branch off and lead to the next level of nodes, which can further branch off. Eventually, the branching will end and the observations will be contained in a leaf node, a node that does not have any more branches stemming from it. It is at the leaf node that the observation receives the classification that is of interest. In this project, the classification of interest is whether a film will win the Academy Award for Best Picture. A branch is a particular outcome or quality that an observation is evaluated on and which carries observations from one level of nodes to the next. Branches coming from the same node are mutually exclusive; otherwise, some observations would not have a clearly defined path to follow.

Two common parameters that will be used to build the decision trees in this project are the minimum split value and the complexity parameter. Both parameters help determine whether another branch is produced at a particular node. The minimum split value is the minimum number of observations that must exist in a node for that node to branch off. For every decision tree classification model in this project, the minimum split value is set to be 10. That way, if fewer than 10 films end up in a particular node, that node is prohibited from producing anymore branches. This prevents the trees from becoming so complex and specific that they lose their value. The complexity parameter is another way to control the branching of a decision tree. It is the minimum improvement in the model required to generate a branch. R will be used to identify the ideal complexity parameter for each model. If, at a particular node, the improvement in the

model that comes along with creating a branch does not meet or exceed the value of the complexity parameter, the branch will not be produced.

In addition to the different methods used to generate the predictive models, there are a few methods that will be used to evaluate the performance of the models. These methods must be consistent and comparable across the various models. The technique that will be used to achieve this is leave-one-out cross validation (LOOCV). In leave-one-out cross validation, a model is trained using all but one observation (or subset of related observations) and is then evaluated on how well it does when classifying the observation that was left out. This process is continued until each observation in the data set has been left out of the model-building process once. For simplicity, consider a data set with three observations: A, B, and C. If LOOCV is performed on this data set, the model will be built three times. First, the model will be built using a training set consisting of observations A and B, and will be evaluated based on whether it correctly classifies observation C. Then, the model will be built using a training set consisting of observations B and C, and evaluated on whether it correctly classifies observation A. The final time will see the model being built using a training set with observations A and C. This time, the model will be used to predict observation B and evaluated on whether it made the correct prediction or not. One measure that will be used to compare across different models is the overall model accuracy, which is simply calculated by taking the number of correct classifications divided by the number of years included in the particular model (which is equivalent to the number of Best Picture winners included).

Since the data set for this project contains observations that are related by year, LOOCV will be performed based on the year. In other words, every model generated will be trained using all but one year's worth of observations (5-10 depending on the number of films nominated each

year). Then, that model will be used to predict which film wins in the year that was left out. If the model correctly predicts the winner, it is given a score of 1. As mentioned previously, once all years have been cycled through, the model accuracy is determined by summing the number of correct predictions and dividing by the total number of films to win Best Picture within that specific time frame.

The other measure that will be calculated with LOOCV and used to compare models is the root mean square error, or RMSE. The RMSE creates a slightly different picture than does the model accuracy measure just described. It is a measure of how far off the models are when the film they predicted to win did not actually win. The models in this project will generate the probabilities of each film winning Best Picture. An important distinction to make here is that the probability included in the model output for each film is telling of that film's probability of winning Best Picture on its own, without thinking of it in the context of the other films nominated. Thus, an additional step was taken to normalize these probabilities, which simply amounts to summing up the probabilities in a given year and dividing each of those individual probabilities by the year's sum. By doing this, the probabilities within a year will sum to one. Throughout the remainder of this project, any time probabilities are discussed, it will be strictly in reference to these normalized probabilities, not the probabilities produced directly by the models.

Each year, the maximum probability generated will be recorded, along with the film with which that probability is associated. If that film predicted by the models to win did not actually win the award, the square error for that given year will be nonzero and is computed by taking the difference between the maximum probability for that year's nominees and the probability of the film that actually won the award, and squaring the difference. Then, once winners have been

predicted for all years included in the data set, the squared errors will be totaled and that total will be divided by the number of films that won Best Picture in the timeframe spanned by the data. Finally, taking the square root of that resulting value produces the RMSE for that model.

Chapter 3

Data Analysis and Results

This chapter begins with an exploration of the data that has been collected before it provides a discussion of the results from building and comparing the predictive models.

Data Preparation

With the data collected, the natural next step of the project is to prepare the data for analysis. At this point, data specific to each of the predictor variables are contained in their own R scripts. For example, in its original form, the Best Production Design data frame contains columns for the Film, Year, Art Directors, Interior Decorators, Set Decorators, and Production Designers. This project is only interested in the films that were nominated for Best Picture throughout history and whether or not they received nominations for these predictor variables (and in the case of some other awards show categories, whether or not they won). Thus, the only relevant data are the name of the film and the year. To prepare the data in each of these separate R scripts to be joined with the Best Picture historical data, all columns that are not relevant to the analysis will be removed (leaving only the name of the film and the year).

In addition to removing columns that are irrelevant to this analysis, columns which keep track of whether a film was nominated in a particular category or for an external award, as well

as if it won, will be created for each R script outside of the main Best Picture script. Using Best Production Design again as an example, a table listing all films ever to be nominated in this category was scraped from the appropriate Wikipedia page into R. From there, all irrelevant columns were removed and two additional columns were added. One column keeps track of whether the film listed was nominated for Best Production Design (0 if no, 1 if yes). Given that the data frame in this script is a list of all films nominated in this category, the column will be filled with all 1s. This is important when joining the Best Production Design data frame to the main Best Picture data frame using the "left_join()" function in R. When the left join is performed, only the data from the Best Production Design data frame associated with a film that also appears in the Best Picture data frame will be carried over. This is how the Best Picture nominees which received nominations in other categories are identified. The second column created keeps track of the winners (though this is only relevant to the external award shows data). This is a bit more complicated, but since the tables in Wikipedia always list the winner first each year, a loop is used to cycle through all of the years, assigning the first film in a given year a value of 1, indicating it won the award. There are some special cases which involve multiple awards being given out in the same year, such as ties or, for cinematography, awards being distributed for both color and black-and-white films. These require some further adjustments.

Issues also arose in the acting categories when there were multiple nominations for the same film. For example, in the Best Actor in a Supporting Role category at the 2019 Oscars, both Al Pacino and Joe Pesci were nominated for their respective roles in *The Irishman*. In the Wikipedia table, part of which is shown in Figure 2, both actors have their own row, but the film for which they were nominated, because it is the same, spans the two actors' rows. When the

data is scraped into R, this creates an issue where the film entry for Joe Pesci is filled with his character's name rather than the film's title. This requires some manual adjustments to the data frame, replacing the character's name with the film title in the Film column. This happens in all of the acting categories, but fortunately, a list of multiple nominations for the same film is provided on the appropriate Wikipedia pages. Hence, it is easy to identify where the issues are.

2019 (92nd)	**Brad Pitt ‡**	**Cliff Booth**	***Once Upon a Time in Hollywood***
	Tom Hanks	Fred Rogers	*A Beautiful Day in the Neighborhood*
	Anthony Hopkins	Pope Benedict XVI	*The Two Popes*
	Al Pacino	Jimmy Hoffa	*The Irishman*
	Joe Pesci	Russell Bufalino	

Figure 2. 2019 Academy Award for Best Supporting Actor, Wikipedia

In the Best Picture data frame, another adjustment that is made to prepare for data analysis is combining certain categories which are mutually exclusive. For instance, a film cannot be nominated for both Best Adapted Screenplay and Best Original Screenplay; it must be one or the other. However, both are screenplay awards. Thus, these two columns are added together to form a column which considers simply whether the film was nominated for a screenplay award at the Oscars. Another example of this is seen with the Golden Globes awards for Best Drama and Best Musical or Comedy. A film cannot be nominated for both awards, so the two variables are combined into one that keeps track of which films were nominated for Best Film at the Golden Globes.

Data Visualization

The data collection process involved collecting data for 33 relevant explanatory variables (listed in Appendix A) for 563 films in total (that is, the number of films to be nominated for the Academy Award for Best Picture from 1927 through 2019). While the size of this data frame is certainly manageable, it is also a bit overwhelming, unless there is specific information one is looking for. In this project, all but two models are built using historical data starting in 1967 or later; the two that include earlier data capture years between 1940 and 2019. Thus, when visualizing the data, our focus will be on the Best Picture nominees from 1967 onward. Figures 3-6 below have been created to better visualize the data that has been collected. Along the x-axis are the titles of the films that won Best Picture in the time frame described in the caption, arranged in order by year. For example, the first film labeled in Figure 3 is *In the Heat of the Night*, which won Best Picture in 1967. *Oliver!* is the next film listed along the axis, so it won in 1968. The final film, *Ordinary People*, won in 1980.

Along the y-axis is the count of the number of Oscar nominations, aside from Best Picture, for which the films were nominated. It is important to note that the counts shown in these Figures only consider the categories that were included in this project. There are a few additional Oscar categories that were excluded and, therefore, do not contribute to the counts. Also, there is one award, namely the Academy Award for Best Makeup and Hairstyling that was not added until 1981. This influenced the way the years were split up: 1967-1980, 1981-1993, 1994-2006, and 2007-2019. The total number of nominations possible (again, from those included in this project) for films between 1967 and 1980 was 12, while it was 13 for all films starting in 1981 (due to the addition of the Best Makeup and Hairstyling award).

Bars were included for all films nominated each year. The blue bars are associated with the films that won the Academy Award for Best Picture, while the gray bars are for the films that were nominated but did not win (whose titles are not listed).

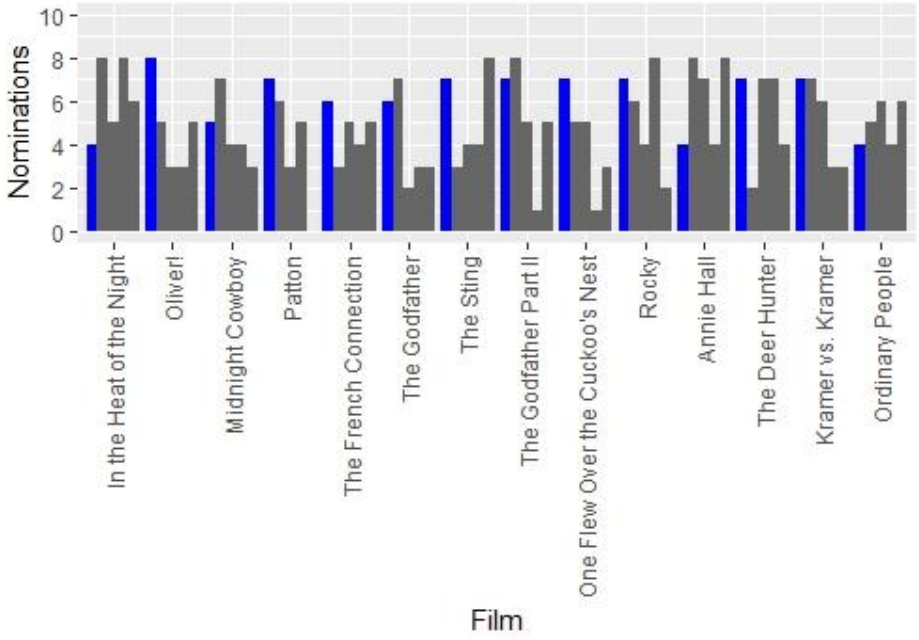

Figure 3. Number of Nominations for 1967-1980 Best Picture Winners

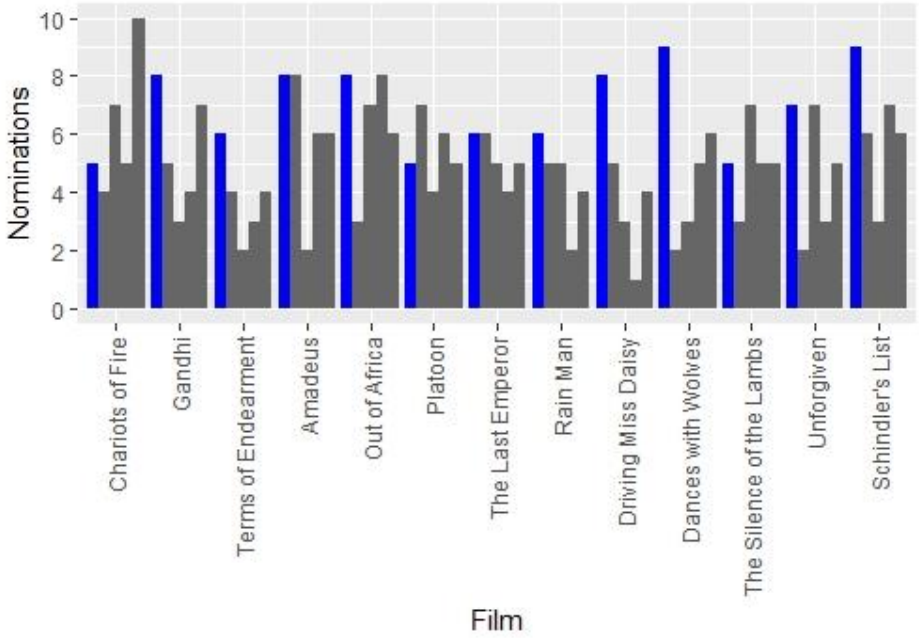

Figure 4. Number of Nominations for 1981-1993 Best Picture Winners

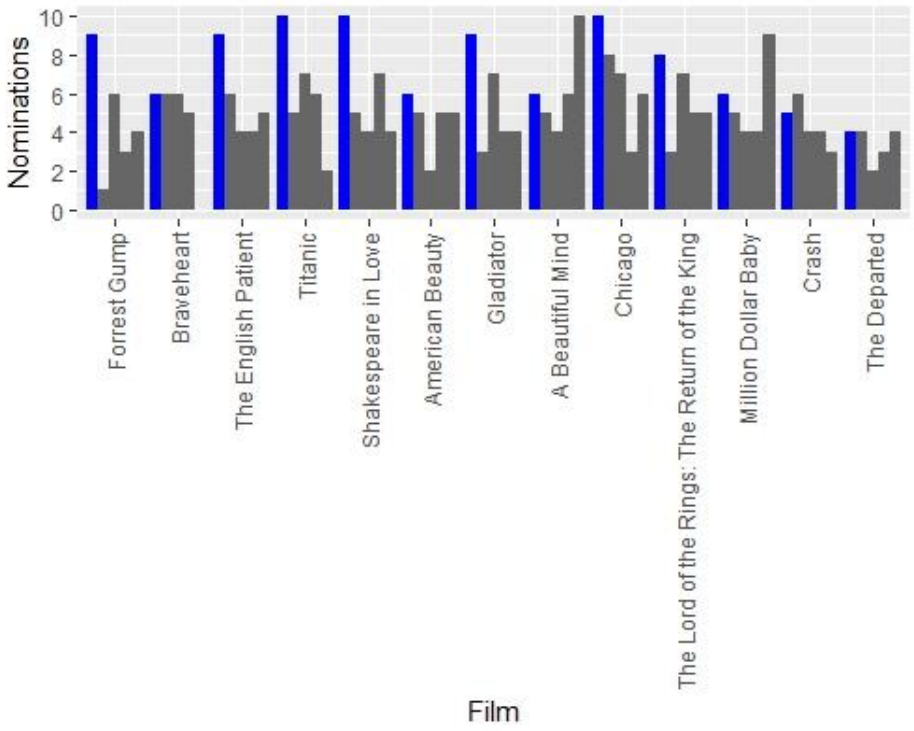

Figure 5. Number of Nominations for 1994-2006 Best Picture Winners

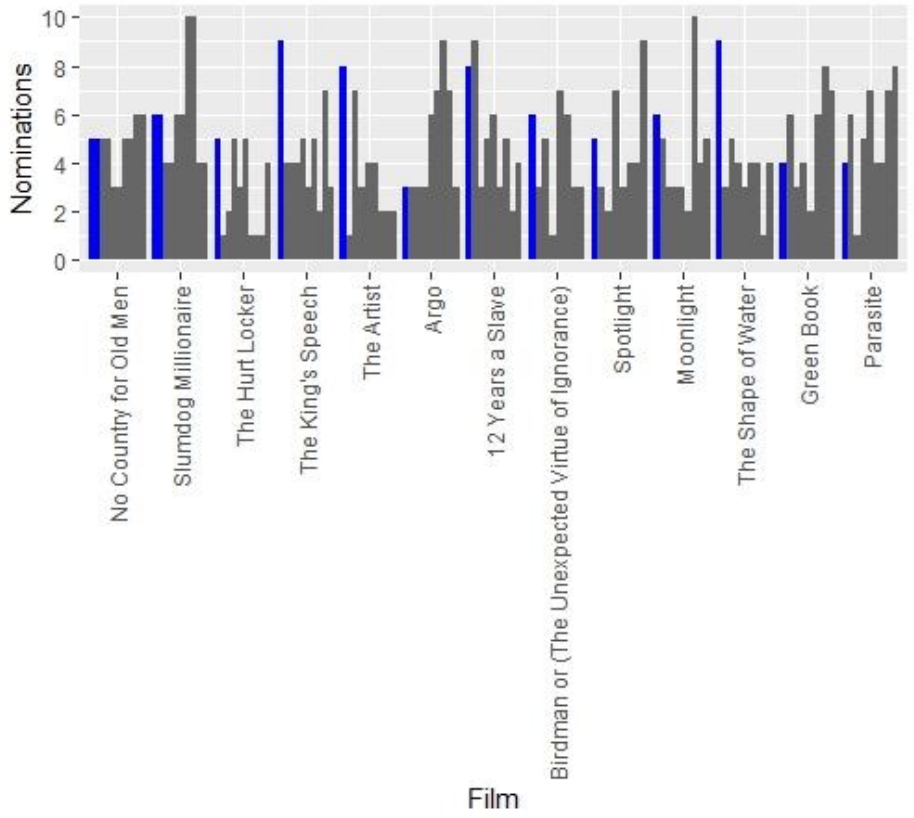

Figure 6. Number of Nominations for 2007-2019 Best Picture Winners

What these bar plots show is that there is tremendous variation in the number of Oscar nominations received by the Best Picture winners over time, ranging from as little as 3 nominations to as many as 10. In the past, it appears that the films that won Best Picture received quite a few nominations. However, in more recent years, that appears to be less valuable. Between 2007 and 2019, only 3 Best Picture winners had the most nominations among their respective competitors. While these plots help to visualize the data that has been collected, they do not lead to any significant conclusions being drawn. This shows that additional analysis via modeling is necessary.

Results

The first modeling technique used in this project was logistic regression. Several logistic regression models were constructed, incorporating different sets of predictor variables and utilizing historical data from varying time frames. In general, for each unique time frame, there was a model built using only available Oscar nominations data as well as a model that accounts for all applicable explanatory variables (this includes Oscar nominations data and other awards data). The time frames of interest, chosen based on when certain data become available, are the following: 1940-2019 (Oscars data only), 1967-2019, 1969-2019, and 1995-2019. Table 1 shows the logistic regression models that were produced in addition to their corresponding prediction accuracies and root mean square errors. Recall that the prediction accuracy is determined by dividing the total number of correct predictions by the number of years included in the model. The square error is the squared difference between the maximum probability of each year's Best Picture nominees and the probability associated with the film that won the award. These two

probabilities only differ when the film predicted by the model to win Best Picture did not. In these cases, the square error for the year is nonzero. The root mean square error, as the name implies, is then determined by taking the square root of the average of all these squared errors.

Table 1. Logistic Regression (LR) Model Performance

Model ID	Model Description	Prediction Accuracy	Root Mean Square Error (RMSE)
LR.1	Oscar Nomination Variables, 1940-2019	0.450	0.211
LR.2	Oscar Nomination Variables, 1967-2019	0.434	0.215
LR.3	All Possible Variables, 1967-2019	0.642	0.326
LR.4	Oscar Nomination Variables, 1969-2019	0.412	0.215
LR.5	All Possible Variables, 1969-2019	0.667	0.367
LR.6	Oscar Nomination Variables, 1995-2019	0.320	0.267
LR.7	All Possible Variables, 1995-2019	0.560	0.564

There are a few interesting results from Table 1. First, looking at the models consisting only of Oscar nominations data (LR.1, LR.2, LR.4, and LR.6), the models that include more years' worth of data have a greater prediction accuracy. In other words, LR.1 includes Oscar nominations data from 1940 to 2019 and has a greater prediction accuracy (0.45) than does LR.6, which goes back to 1995 and has a prediction accuracy of 0.32. Another notable result is that in each time frame of interest, when all applicable predictor variables are added into the model, the prediction accuracy improves substantially. This is to be expected because, in this case, more data means more information at the disposal of the computer software used to generate these models. However, there is a tradeoff that comes with improved model prediction accuracy, which can be seen in the root mean square errors in both Table 1 and Table 2. With more data

available to the models, the models become more confident in their predictions. When they are right, they are right, but when they are wrong, they are very wrong. As an example, for 2005, LR.5 predicted *Brokeback Mountain* to win, with a probability of 84%. That is a very confident prediction, but it was also wrong. The film that won that year, *Crash*, was given a probability of winning of 9.5%. Because of the huge difference between these two probabilities, the square error that year (0.555) was among the highest for the model and contributed significantly to the model's overall root means square error. While these results are interesting, they are not the most satisfying. Most of the prediction accuracies for the logistic regression models shown in Table 1 are around or below 50%. Ideally, those numbers would be a bit larger.

The other modeling technique used in this project was decision tree classification. Decision tree models were developed in correspondence with the logistic regression models displayed in Table 1. In other words, a decision tree model was generated for each of the time frames and including the same set of variables as the logistic regression models. However, a few additional considerations needed to be made regarding the construction of these decision tree classification models. The minimum split value was set to be 10, which means that the minimum number of observations needed to attempt a split at a given node is 10. Whether or not a split is made is determined by the complexity parameter, which can be set by the user or which can be generated by the R software. In this project, R was used to identify the ideal complexity parameter for each model.

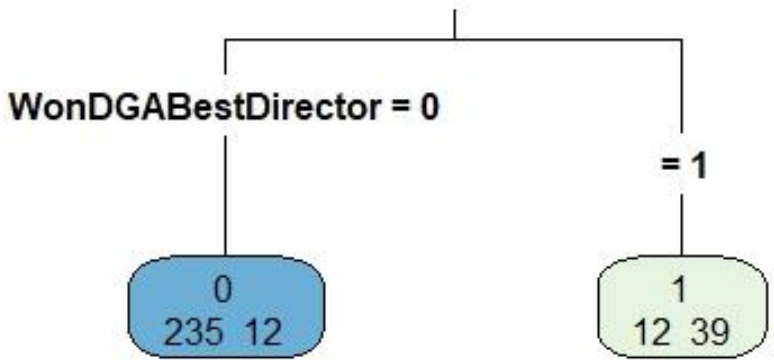

Figure 7. Decision Tree (1969-2019, All Possible Variables Included)

Figure 7 shows the decision tree associated with the "best" decision tree classification model (the one with the highest prediction accuracy) that was developed, namely DTC.5, which included all possible variables and was built using historical data from 1969 to 2019. This is the most simplistic tree of all the trees generated as there is only one branch. Other trees contain more than a single branch; however, the added complexity in those trees did not make their overall accuracy greater than that of this simplistic model. From Figure 7, we can see in the two terminal nodes that additional splits could have been made because the terminal nodes each have more than the minimum of 10 observations. However, additional splits were not made because the cost of adding additional variables to the model was greater than the model's complexity parameter.

Each terminal node consists of three numbers. The top number in each terminal node is the classification of all observations that belong to that node. The top number in the left terminal node in Figure 7 is a 0, which indicates that the 247 total observations in that node were all predicted not to win the award for Best Picture. Of the 247 observations in that node, 235 did not

win Best Picture (i.e., they were correctly classified) while 12 were misclassified and did win the top award. The top number in the right terminal node is 1, which means all the films belonging to that node (i.e., those that won the DGA award for Best Director) were predicted to win Best Picture. Of the 51 films in this node, 39 did win Best Picture and were therefore correctly classified. The other 12 films in this node were misclassified; they did not win the award for Best Picture. As discussed previously, this model is very confident in its predictions, but when its predictions are incorrect, the root mean square error is much higher than it is with some of the other models.

Figure 8 below shows a slightly more complex decision tree, though other models are much more complex than even this one. Exploring this tree, five films nominated for Best Picture since 1995 met all of the following conditions: won Best Director at the DGA Awards (WonDGABestDirector), received an Oscar nomination for Best Production Design (NominatedBPD), and were nominated for Best Supporting Actress at the SAG Awards (SAGBestSupportingActressNom). Every one of those five films won Best Picture at the Academy Awards, evident through the third terminal node from the left. Similarly, ten films since 1995 simultaneously won Best Director at the DGA and were not nominated for Production Design. Of those films, nine went on to win Best Picture while one did not.

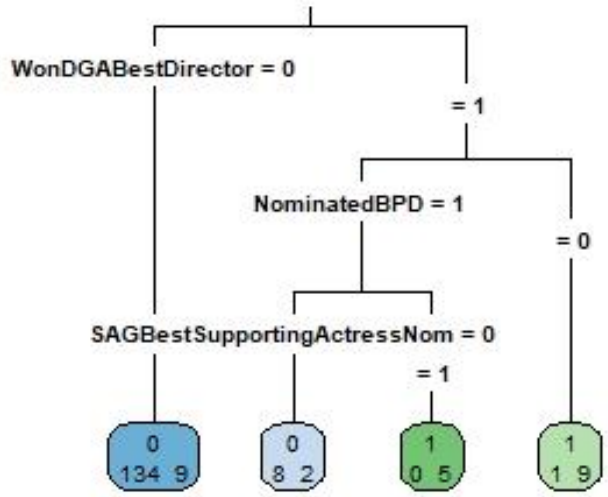

Figure 8. Decision Tree (1995-2019, All Possible Variables Included)

These decision tree classification models, together with their overall prediction accuracies and root mean square errors, are shown in Table 2. Compared to the logistic regression models, the corresponding decision tree classification models had higher prediction accuracies in every case. Furthermore, all but one of the decision tree classification models had lower root mean square errors relative to their logistic regression counterparts. The one exception was the model built using all possible variables from 1967 to 2019. One explanation for these higher prediction accuracies is the presence of ties in any given year. For instance, some models were such that multiple films in a particular year met all the conditions that would classify them as a winner. Normally, if the film predicted by the model to win was the film that won, it would be given a 1 in the "Correct" column. If multiple films were predicted to win Best Picture, and one of those films did indeed win, then it would be assigned a value equal to 1 divided by the number of films predicted to win in the "Correct" column. The adjusted prediction accuracy was computed by dividing the sum of the "Correct" column by the total number of winners in the

years included in the model. Adjusting for ties did reduce the prediction accuracy for some models, but not significantly. For a full list of explanatory variables and the models in which they were included, see Appendix A.

Table 2. Decision Tree Classification (DTC) Model Performance

Model ID	Model Description	Prediction Accuracy * adjusted for ties	Root Mean Square Error (RMSE)
DTC.1	Oscar Nomination Variables, 1940-2019	0.675 * 0.604	0.184
DTC.2	Oscar Nomination Variables, 1967-2019	0.679 * 0.629	0.198
DTC.3	All Possible Variables, 1967-2019	0.736 * no ties	0.345
DTC.4	Oscar Nomination Variables, 1969-2019	0.667 * 0.585	0.200
DTC.5	All Possible Variables, 1969-2019	0.765 * no ties	0.340
DTC.6	Oscar Nomination Variables, 1995-2019	0.640 * 0.495	0.118
DTC.7	All Possible Variables, 1995-2019	0.640 * no ties	0.201

Tables 3 and 4 below present the nine predictions for each of the "best" models (again, those with the highest prediction accuracies) generated using logistic regression (Table 3) and decision tree classification (Table 4) that were most incorrect, meaning the predictions that contributed the most to the root mean square error. Both models (LR.5 and DTC.5) were constructed using all possible variables from 1969 to 2019. Table 3 organizes the films by which had the greatest difference in probabilities between the predicted winner and the actual winner. Table 4 is organized the same way, though when the predicted films have the same probability, they are arranged in ascending order by year.

Table 3. Most Incorrect Predictions for Model LR.5

Year	Predicted Winner	Predicted Winner's Probability of Winning	Actual Winner	Actual Winner's Probability of Winning
1989	Born on the Fourth of July	0.991	Driving Miss Daisy	<0.001
1997	As Good as It Gets	0.908	Titanic	0.035
1995	Apollo 13	0.889	Braveheart	0.019
2016	La La Land	0.846	Moonlight	0.059
1985	Prizzi's Honor	0.830	Out of Africa	0.073
2015	The Revenant	0.787	Spotlight	0.041
2005	Brokeback Mountain	0.840	Crash	0.095
2000	Crouching Tiger, Hidden Dragon	0.735	Gladiator	0.009
2018	Roma	0.732	Green Book	0.029

Table 4. Most Incorrect Predictions for Model DTC.5

Year	Predicted Winner	Predicted Winner's Probability of Winning	Actual Winner	Actual Winner's Probability of Winning
1981	Reds	0.797	Chariots of Fire	0.051
1985	The Color Purple	0.797	Out of Africa	0.051
1989	Born on the Fourth of July	0.797	Driving Miss Daisy	0.051
1995	Apollo 13	0.797	Braveheart	0.051
1998	Saving Private Ryan	0.797	Shakespeare in Love	0.051
2000	Crouching Tiger, Hidden Dragon	0.797	Gladiator	0.051
2005	Brokeback Mountain	0.797	Crash	0.051
2015	The Revenant	0.692	Spotlight	0.044
2018	Roma	0.692	Green Book	0.044

Furthermore, six of the top nine incorrect predictions shown in Table 3 also appear in Table 4. In other words, the models are consistent with one another in terms of the predictions they got most incorrect. This seemingly suggests that the films that did win in those years were perhaps considered "surprises" at the time or even "mistakes" today. *Driving Miss Daisy* (1989), *Shakespeare in Love* (1998), and *Crash* (2005) are films that, in retrospect, may not have been the most obvious or deserving films to win the Academy Award for Best Picture, especially over films like *Born on the Fourth of July*, *Saving Private Ryan*, and *Brokeback Mountain*, respectively. Interestingly, the top seven films listed in Table 4 have equal probabilities. This is because the decision tree classification model (DTC.5) only assigns probabilities based on one factor, namely winning the DGA award for Best Director. Since all the films in Table 4 are ones that won this award, the model sees no difference between them. Thus, they are given the same probability. The only reason the last two films listed in the table are different is that there were more films in competition for Best Picture those years.

Chapter 4

Conclusions

This chapter begins with some of my thoughts on the models generated through the completion of this project. There will then be a discussion of the project's limitations and the ways in which the project could be extended in the future. Lastly, the developed models will be applied to predict the winner of this year's Academy Award for Best Picture.

My Thoughts on Model Performance

Upon generating and evaluating the logistic regression models, I was a tad disappointed to see prediction accuracies that were around or below 50%, with a few exceptions. I was expecting the models to have greater predictive capabilities. However, I needed to realize that this project essentially attempts to predict human behavior, which is often quite unpredictable. It is decisions made by voters, each with their own inconsistencies and opinions, that determine which films will win at the Academy Awards. After considering this, I was happier with the performance results from the logistic regression models. Then, once the decision tree classification models were generated, I was thrilled to see such an improvement in model performance. These models now had prediction accuracies hovering around 65% and a couple that were around 75%. I think those are some strong results. Furthermore, the predictive power of these models becomes stronger when used in conjunction with one another to make predictions, as will be demonstrated shortly.

One of the most valuable takeaways from the results of this project is the importance of winning the award for Best Director from the Directors Guild of America. The top two decision tree classification models, in terms of their prediction accuracies, were DTC.5 and DTC.3, which were built using all possible variables from 1969-2019 and 1967-2019, respectively. The model building process for each of these two models determined that including just one variable, WonDGABestDirector, in the tree was the best option of all possible combinations. Not only that, but the logistic regression models also showed improvement in their prediction accuracies once the WonDGABestDirector variable was included in the model building process. To demonstrate this point further, in Figure 5, *Argo* is shown to have received Oscar nominations in only three of the categories included in this project. One of the biggest awards at the Oscars for

which it failed to be recognized was Best Directing. However, it still went on to win Best Picture. By looking at the list of nominees for the DGA award for Best Director, *Argo* was not only on that list, but it won the award. This explains why, despite not receiving too many Oscar nominations, the film still won the top prize at the Oscars, influenced significantly by its win at the DGA Awards.

Limitations and Future Extensions

While this project has some neat applications, it also has a few notable limitations. For one, these models generate the probability of a film winning Best Picture, which is quantitative and based on a number of binary variables. Quantitative analysis alone will likely not yield the greatest results, as there are plenty of qualitative factors that should also be considered. For instance, in the interest of making accurate predictions for the Academy Awards, paying attention to the conversation surrounding a film, by fans, critics, and awards voters, can be quite informative. What do others think of this film's chances of winning a particular award and what is their reasoning? Sources like Gold Derby and various YouTube film pundits can be invaluable when thinking about these types of questions.

In addition to paying attention to what others are saying, another valuable qualitative variable that one should consider when making Oscar predictions, is their own enjoyment of the films nominated. Of course, trying to watch all the nominated films may not be realistic for some people, as some films are extremely limited in their release and because of the time commitment that would be necessary. Also, there is the element of personal bias. A certain genre of film that you do not like or a particular director whose work you really enjoy can be an influential factor.

Another qualitative variable not considered by these models is the specific talent associated with a film. For instance, a film directed by Steven Spielberg or Martin Scorsese may have its probability of winning increased because those filmmakers are involved. However, to determine if this is true, additional analysis would be required.

Furthermore, these models excluded a few quantitative variables which might be of interest. For reasons discussed earlier in the paper, factors like a film's box office numbers or its critic score on a platform such as Rotten Tomatoes were not included in these models. That is not to say, however, that these variables are insignificant. I would reason that the critical reviews a film receives would be influential on its chances of winning Best Picture. Thus, these are items that should be considered in addendum to the results generated by this project's models and the other qualitative variables discussed.

One of the most exciting aspects of this project is how it could be extended in the future. For one, it would be fascinating to take what was done in this project and to develop new models for predicting the winners of other awards given out at the Oscars or even at other awards shows entirely. Are certain awards "easier" to predict than others? Which variables are most influential in predicting each of these other awards? Since winning the DGA award for Best Director was shown to be so influential, it would be interesting to look at the factors most significant to predicting the winners of that award. Another obvious extension of this project would be to use and to compare more predictive modeling techniques. For instance, would the random forest technique produce more accurate models, or how about neural networks? All in all, this project has its limitation and ways in which it could be extended in the future.

An Application

Probably the most apparent application of this project is using it to make predictions for future winners of the Academy Award for Best Picture. The next Academy Awards ceremony, at the time this thesis was written, will take place on Sunday, April 25, 2021. Nominations were announced on Monday, March 15, 2021. The films nominated for Best Picture include *Nomadland*, *Minari*, *Promising Young Woman*, *The Trial of the Chicago 7*, *Mank*, *Sound of Metal*, *Judas and the Black Messiah*, and *The Father*. Since the Screen Actors Guild Awards, the Directors Guild of America Awards, and the British Academy Film Awards do not take place until April 4th, 10th, and 11th, respectively, dates near or after the submission of this thesis, a few variables were excluded from the models when making predictions for this year's Best Picture winner. In particular, the variables keeping track of which films won at each of the aforementioned award shows have been removed from the relevant models. However, nominations for these awards ceremonies have been announced and were included in the predictive models. Table 5 shows the films with the greatest probabilities of winning the Academy Award for Best Picture, according to each of the developed models. Since DTC.3 and DTC.5 only classify a movie based on a single predictor variable, namely whether the film won the DGA award for Best Director, and that variable has been removed from the models, these two models predict an eight-way tie, which is not a meaningful prediction.

Table 5. Predictions for the 93rd Academy Awards

Model ID	Film with Highest Probability of Winning	Probability	Film with Second Highest Probability of Winning	Probability
LR.1	Nomadland	0.283	Promising Young Woman	0.219
LR.2	Nomadland	0.280	Promising Young Woman	0.226
LR.3	Nomadland	0.528	The Trial of the Chicago 7	0.178
LR.4	Nomadland	0.281	Promising Young Woman	0.223
LR.5	Nomadland	0.461	Promising Young Woman	0.258
LR.6	The Trial of the Chicago 7	0.390	Nomadland	0.258
LR.7	Nomadland	0.700	Promising Young Woman	0.235
DTC.1	Nomadland	0.389	Promising Young Woman	0.324
DTC.2	Promising Young Woman	0.566	Nomadland	0.179
DTC.3	8-way tie 0.125			
DTC.4	Nomadland	0.314	Promising Young Woman	0.233
DTC.5	8-way tie 0.125			
DTC.6	The Father	0.203	4-way tie 0.188	
DTC.7	Nomadland	0.606	7-way tie 0.056	

To synthesize what Table 5 is suggesting, it is clear that *Nomadland* is the favorite to win the Academy Award for Best Picture this year, having been given the highest probability of winning by nine models. Of the five models that did not explicitly predict *Nomadland* to win, two generated an eight-way tie. Thus, only three models predicted a film other than *Nomadland* to win. Further, each of those films predicted a different winner from one another. *Promising*

Young Woman, *The Trial of the Chicago 7*, and *The Father* were each predicted to win once. Just as *Nomadland* had the highest probability of winning in most models, *Promising Young Woman* typically had the second highest probability of winning. Based on these models, other sources like Gold Derby, and my own enjoyment of the films nominated, I suspect *Nomadland* will indeed take home the Oscar for Best Picture this year.

Appendix A

Explanatory Variables

Variable	Description	Available Data	Used in Models x (LR.x and DTC.x include the same variables)
NominatedBDirector	Academy Award for Best Director (Nomination)	1927-2019	1, 2, 3, 4, 5, 6, 7
NominatedBE	Academy Award for Best Film Editing (Nomination)	1934-2019	1, 2, 3, 4, 5, 6, 7
NominatedBC	Academy Award for Best Cinematography (Nomination)	1927-2019	1, 2, 3, 4, 5, 6, 7
NominatedBPD	Academy Award for Best Production Design (Nomination)	1927-2019	1, 2, 3, 4, 5, 6, 7
NominatedBCD	Academy Award for Best Costume Design (Nomination)	1948-2019	2, 3, 4, 5, 6, 7
NominatedBScreenplay	Academy Award for Best Adapted Screenplay (Nomination)	1927-2019	1, 2, 3, 4, 5, 6, 7
	Academy Award for Best Original Screenplay (Nomination)	1940-2019	
BestActorNom	Academy Award for Best Actor (Nomination)	1927-2019	1, 2, 3, 4, 5, 6, 7
BestActressNom	Academy Award for Best Actress (Nomination)	1927-2019	1, 2, 3, 4, 5, 6, 7

BestSupportingActorNom	Academy Award for Best Supporting Actor (Nomination)	1936-2019	1, 2, 3, 4, 5, 6, 7
BestSupportingActressNom	Academy Award for Best Supporting Actress (Nomination)	1936-2019	1, 2, 3, 4, 5, 6, 7
NominatedBSong	Academy Award for Best Original Song (Nomination)	1934-2019	1, 2, 3, 4, 5, 6, 7
NominatedBestVFX	Academy Award for Best Visual Effects (Nomination)	1938-2019	1, 2, 3, 4, 5, 6, 7
NominatedBestMakeupHair	Academy Award for Best Makeup and Hairstyling (Nomination)	1981-2019	6, 7
NominatedBBF	BAFTA Award for Best Film (Nomination)	1947-2019	3, 5, 7
WonBBF	BAFTA Award for Best Film (Win)	1947-2019	3, 5, 7
NominatedBGoldenGlobes	Golden Globe Award for Best Motion Picture – Drama (Nomination)	1943-2019	3, 5, 7
	Golden Globe Award for Best Motion Picture – Musical or Comedy (Nomination)	1951-2019	
WonBGoldenGlobes	Golden Globe Award for Best Motion Picture – Drama (Win)	1943-2019	3, 5, 7
	Golden Globe Award for Best Motion Picture – Musical or Comedy (Win)	1951-2019	
NominatedSAGBestCast	Screen Actors Guild Award for Outstanding	1995-2019	7

	Performance by a Cast in a Motion Picture (Nomination)		
WonSAGBestCast	Screen Actors Guild Award for Outstanding Performance by a Cast in a Motion Picture (Win)	1995-2019	7
SAGBestActorNom	Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Leading Role (Nomination)	1994-2019	7
WonSAGBActor	Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Leading Role (Win)	1994-2019	7
SAGBestActressNom	Screen Actors Guild Award for Outstanding Performance by a Female Actor in a Leading Role (Nomination)	1994-2019	7
WonSAGBestActress	Screen Actors Guild Award for Outstanding Performance by a Female Actor in a Leading Role (Win)	1994-2019	7
SAGBestSupportingActorNom	Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Supporting Role (Nomination)	1994-2019	7

WonSAGBSuppActor	Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Supporting Role (Win)	1994-2019	7
SAGBestSupportingActressNom	Screen Actors Guild Award for Outstanding Performance by a Female Actor in a Supporting Role (Nomination)	1994-2019	7
WonSAGBSuppActress	Screen Actors Guild Award for Outstanding Performance by a Female Actor in a Supporting Role (Win)	1994-2019	7
NominatedPGABestPicture	Producers Guild of America Award for Best Theatrical Motion Picture (Nomination)	1989-2019	7
WonPGABestPicture	Producers Guild of America Award for Best Theatrical Motion Picture (Win)	1989-2019	7
NominatedDGABestDirector	Directors Guild of America Award for Outstanding Directing – Feature Film (Nomination)	1948-2019	3, 5, 7
WonDGABestDirector	Directors Guild of America Award for Outstanding Directing – Feature Film (Win)	1948-2019	3, 5, 7
WGAScreenplayNom	Writers Guild of America Award for Best Adapted	1969-2019	5, 7

	Screenplay (Nomination)		
	Writers Guild of America Award for Best Original Screenplay (Nomination)		
WGAScreenplayWin	Writers Guild of America Award for Best Adapted Screenplay (Win)	1969-2019	5, 7
	Writers Guild of America Award for Best Original Screenplay (Win)		

Appendix B

Sample Data Scraping Code

```
url <- "https://en.wikipedia.org/wiki/Academy_Award_for_Best_Picture"

best_picture_nominees_1920s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[3]') %>%
  html_table(fill=TRUE)
bpn1920s <- data.frame(best_picture_nominees_1920s)

best_picture_nominees_1930s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[4]') %>%
  html_table(fill=TRUE)
bpn1930s <- data.frame(best_picture_nominees_1930s)

best_picture_nominees_1940s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[5]') %>%
  html_table(fill=TRUE)
bpn1940s <- data.frame(best_picture_nominees_1940s)

best_picture_nominees_1950s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[6]') %>%
  html_table(fill=TRUE) %>% rename(Producer.s. = "Film Studio/Producer(s)")
bpn1950s <- data.frame(best_picture_nominees_1950s)

best_picture_nominees_1960s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[7]') %>%
  html_table(fill=TRUE)
bpn1960s <- data.frame(best_picture_nominees_1960s)

best_picture_nominees_1970s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[8]') %>%
  html_table(fill=TRUE) %>% rename(Film = Films)
bpn1970s <- data.frame(best_picture_nominees_1970s)

best_picture_nominees_1980s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[9]') %>%
  html_table(fill=TRUE)
bpn1980s <- data.frame(best_picture_nominees_1980s)

best_picture_nominees_1990s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[10]') %>%
  html_table(fill=TRUE)
bpn1990s <- data.frame(best_picture_nominees_1990s)
```

```
best_picture_nominees_2000s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[11]') %>%
  html_table(fill=TRUE)
bpn2000s <- data.frame(best_picture_nominees_2000s)

best_picture_nominees_2010s <- url %>% read_html() %>%
  html_node(xpath='/html/body/div[3]/div[3]/div[5]/div[1]/table[12]') %>%
  html_table(fill=TRUE)
bpn2010s <- data.frame(best_picture_nominees_2010s)

best_picture_nominees <-
bind_rows(bpn1920s,bpn1930s,bpn1940s,bpn1950s,bpn1960s,bpn1970s,bpn1980s,bpn1990s,bp
n2000s,bpn2010s)
```

BIBLIOGRAPHY

"Academy Award for Best Actor." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Actor.

"Academy Award for Best Actress." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Actress.

"Academy Award for Best Adapted Screenplay." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Adapted_Screenplay.

"Academy Award for Best Cinematography." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Cinematography.

"Academy Award for Best Costume Design." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Costume_Design.

"Academy Award for Best Director." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Director.

"Academy Award for Best Film Editing." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Film_Editing.

"Academy Award for Best Makeup and Hairstyling." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Makeup_and_Hairstyling.

"Academy Award for Best Original Screenplay." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Original_Screenplay.

"Academy Award for Best Original Song." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Original_Song.

"Academy Award for Best Picture." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Picture.

"Academy Award for Best Production Design." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Production_Design.

"Academy Award for Best Supporting Actor." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Supporting_Actor.

"Academy Award for Best Supporting Actress." *Wikipedia*,
https://en.wikipedia.org/wiki/Academy_Award_for_Best_Supporting_Actress.

"Academy Award for Best Visual Effects." *Wikipedia*, https://en.wikipedia.org/wiki/Academy_Award_for_Best_Visual_Effects.

Academy Museum. Academy Museum of Motion Pictures, https://www.academymuseum.org/en/visit.

"Academy Story." *Academy of Motion Pictures Arts and Sciences*, https://www.oscars.org/academy-story.

"BAFTA Award for Best Film." *Wikipedia*, https://en.wikipedia.org/wiki/BAFTA_Award_for_Best_Film.

"Decision Tree in R | Classification Tree & Code in R with Example." *Guru99*, https://www.guru99.com/r-decision-trees.html.

Devore, Jay. *Probability and Statistics for Engineering and the Sciences*. CENGAGE Learning, 2015.

"Directors Guild of America Award for Outstanding Directing – Feature Film." *Wikipedia*, https://en.wikipedia.org/wiki/Directors_Guild_of_America_Award_for_Outstanding_Directing_%E2%80%93_Feature_Film.

"Golden Globe Award for Best Motion Picture – Drama." *Wikipedia*, https://en.wikipedia.org/wiki/Golden_Globe_Award_for_Best_Motion_Picture_%E2%80%93_Drama.

"Golden Globe Award for Best Motion Picture – Musical or Comedy." *Wikipedia*, https://en.wikipedia.org/wiki/Golden_Globe_Award_for_Best_Motion_Picture_%E2%80%93_Musical_or_Comedy.

Hadley Wickham (2020). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.6. https://CRAN.R-project.org/package=rvest.

Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.1. https://CRAN.R-project.org/package=dplyr.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Le, James. "Decision Trees in R." datacamp, https://www.datacamp.com/community/tutorials/decision-trees-R.

"Oscar Statuette." *Academy of Motion Picture Arts and Sciences*, https://www.oscars.org/oscars/statuette.

"Producers Guild of America Award for Best Theatrical Motion Picture." *Wikipedia*, https://en.wikipedia.org/wiki/Producers_Guild_of_America_Award_for_Best_Theatrical_Motion_Picture.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

"Screen Actors Guild Award for Outstanding Performance by a Cast in a Motion Picture." *Wikipedia*, https://en.wikipedia.org/wiki/Screen_Actors_Guild_Award_for_Outstanding_Performance_by_a_Cast_in_a_Motion_Picture.

"Screen Actors Guild Award for Outstanding Performance by a Female Actor in a Leading Role." *Wikipedia*, https://en.wikipedia.org/wiki/Screen_Actors_Guild_Award_for_Outstanding_Performance_by_a_Female_Actor_in_a_Leading_Role.

"Screen Actors Guild Award for Outstanding Performance by a Female Actor in a Supporting Role." *Wikipedia*, https://en.wikipedia.org/wiki/Screen_Actors_Guild_Award_for_Outstanding_Performance_by_a_Female_Actor_in_a_Supporting_Role.

"Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Leading Role." *Wikipedia*, https://en.wikipedia.org/wiki/Screen_Actors_Guild_Award_for_Outstanding_Performance_by_a_Male_Actor_in_a_Leading_Role.

"Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Supporting Role." *Wikipedia*, https://en.wikipedia.org/wiki/Screen_Actors_Guild_Award_for_Outstanding_Performance_by_a_Male_Actor_in_a_Supporting_Role.

Stephen Milborrow (2020). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.9. https://CRAN.R-project.org/package=rpart.plot.

Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org.package=rpart.

"The Hashtag That Changed the Oscars: An Oral History." *The New York Times*, https://www.nytimes.com/2020/02/06/movies/oscarssowhite-history.html.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686.

"Writers Guild of America Award for Best Adapted Screenplay." *Wikipedia*, https://en.wikipedia.org/wiki/Writers_Guild_of_America_Award_for_Best_Adapted_Screenplay.

"Writers Guild of America Award for Best Original Screenplay." *Wikipedia*, https://en.wikipedia.org/wiki/Writers_Guild_of_America_Award_for_Best_Original_Screenplay.

Academic Vita

Joshua J. Slayton

jms7840@psu.edu

EDUCATION

Penn State Erie, The Behrend College — Anticipated 5/21
Bachelor of Science in Mathematics (Business Option)

- Minors in Statistics, Finance, and Civic and Community Engagement
- Certificates in Actuarial Mathematics and Statistics, and Financial Risk Management

Related Coursework

- Calculus and Vector Analysis, Linear Algebra, Numerical Analysis, Real Analysis, Statistical Analysis, Experimental Methods, Probability Theory, Nonparametric Statistics, Applied Regression Analysis, Analysis of Variance, Business Forecasting Techniques, Leadership in Sustainability, Service Learning, Science and Sustainable Development, Job Preparation and Success, Foundations of Civic and Community Engagement, Leadership and Ethics, The Art of Cinema, Intro to LGBTQ+ Studies, Queer Theory

HONORS EXPERIENCE

Schreyer Honors College — 8/17 – 5/21

- Undergraduate Thesis
  - Title: "And the Oscar Goes to… (An Application and Comparison of Models Used to Predict Best Picture)"

Behrend Honors Program — 8/17 – 4/19

LEADERSHIP AND SERVICE EXPERIENCE

Anti-Racism Reading Group, Marketing Manager — 1/21 – 5/21
STEM Leaders, Mentor — 8/20 – 5/21
Gender and Sexuality Equality Club, Secretary — 1/20 – 5/21
Leadership Scholars — 8/19 – 5/21
The National Society of Leadership and Success (Sigma Alpha Pi Chapter) — 8/19 – 5/21
Science Ambassadors, Secretary — 1/19 – 5/21
Alternative Spring Break, Trip/Experience Leader — 8/18 – 5/21

- Year 1: Disaster Relief and Recovery (Puerto Rico)
- Year 2: Issues of Immigration and Access to Healthy Living (San Antonio)
- Year 3: Environmental Justice and Natural Disasters (Virtual Experience)

Reality Check, Executive Director — 8/17 – 5/21
Lambda Sigma National Honor Society (Alpha Eta Chapter), Treasurer — 4/18 – 4/19
Welcome Week, Guide — 8/18

RELATED WORK EXPERIENCE

Peer Tutor, Pennsylvania State University — 1/18 – 5/21

- Educate students in core areas of Mathematics and support their motivation to achieve their academic goals

Student Grader, Pennsylvania State University — 8/18 – 12/19

- Assisted professors within the Mathematics department with grading homework and assessments

SKILLS AND AWARDS

Software Skills

- Microsoft Excel, Access, Word, PowerPoint, Outlook; R Studio; Python and Visual Studio

Awards

- School of Science Student Marshal, Dean's List, The President's Freshman Award, The President Sparks Award, The Evan Pugh Scholar Award, Outstanding Student Organization Officer (Non-President) Award, Academic Excellence in Mathematics Award, Outstanding Mathematics Tutor Award, Smile for Sam Award