Mary Buczynski

4/27/2022

Predicting Best Picture Oscar Winning Movies Based on Attributes and Ratings

The data set was found on Kaggle in the form of a CSV named “Oscar Best Picture Movies”. The data was downloaded from the site and read into the file by the function read.csv(). When I began working with the dataframe, I realized that for almost every film there were multiple genres and it would be difficult to use for analysis because there were 128 different combinations. I then opened the csv file in google sheets and added my own column for genres and put what I thought made the most sense per movie. I also noticed how there were missing values for 132 films in columns like content rating and audience status so I made a second data frame excluding those movies. Lastly, I deleted 16 columns as they were difficult to use in my analysis and had too many unique values.

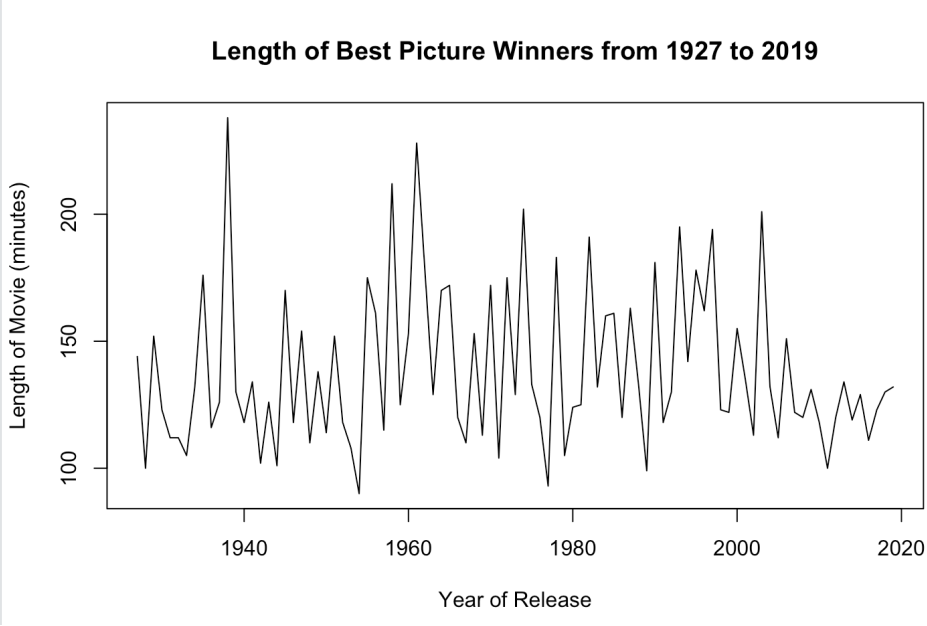
Movie Genre Categories given

Drama,Romance 62	Drama 50	Biography,Drama,History 35
Comedy,Drama,Romance 33	Comedy,Drama 21	Biography,Drama 16
Crime,Drama 13	Drama,Romance,War 13	Drama,War 13
Action,Adventure,Drama 10	Biography,Crime,Drama 10	Crime,Drama,Thriller 10
Adventure,Drama,History 9	Comedy,Romance 9	Crime,Drama,Mystery 9
Biography,Comedy,Drama 8	Biography,Drama,Music 8	Biography,Drama,Romance 7
Crime,Drama,Film-Noir 7	Comedy,Musical,Romance 6	Drama,Western 6
Biography,Drama,Sport 5	Drama,History,Thriller 5	Drama,Musical,Romance 5
Drama,Thriller 5	Action,Drama,History 4	Adventure,Drama,Fantasy 4
Adventure,Drama,Romance 4	Adventure,Drama,Western 4	Comedy,Crime,Drama 4
Comedy,Drama,Family 4	Comedy,Drama,Music 4	Comedy,Drama,War 4
Comedy,Fantasy,Romance 4	Drama,Music 4	Drama,Music,Romance 4
Drama,Sport 4	Action,Adventure,Fantasy 3	Action,Adventure,Sci-Fi 3
Action,Crime,Drama 3	Adventure,Biography,Drama 3	Adventure,Drama 3
Adventure,Drama,War 3	Biography,Drama,Family 3	Biography,Drama,Thriller 3
Crime,Drama,Romance 3	Drama,Family,Fantasy 3	Drama,Family,Musical 3
Drama,Film-Noir 3	Drama,History 3	Drama,History,Romance 3
Drama,Music,Musical 3	Drama,Mystery 3	Drama,Mystery,Romance 3
Drama,Mystery,Thriller 3	Action,Adventure,Romance 2	Action,Biography,Drama 2
Action,Drama,Thriller 2	Action,Drama,War 2	Adventure,Comedy,Family 2
Adventure,Drama,Family 2	Adventure,Drama,Mystery 2	Animation,Adventure,Comedy 2
Biography,Drama,Musical 2	Biography,Drama,War 2	Comedy,Drama,Fantasy 2
Comedy,Drama,History	Comedy,Drama,Musical	Crime,Drama,History

Consolidated Movie Genres

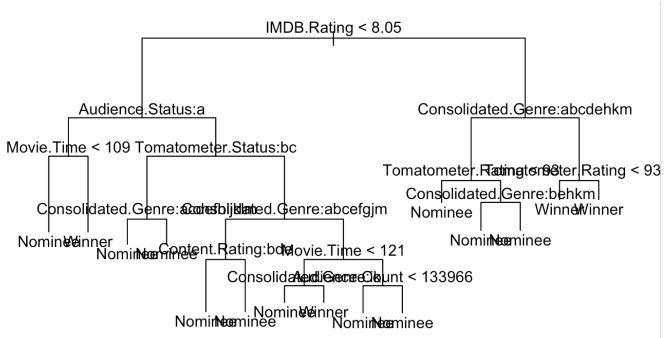
Action 25	Comedy 50	Crime 72	Drama 100	Fantasy 33
History 62	Horror 3	Romance 91	Romantic Comedy 49	Sport 10
Thriller 19	War 42	Western 15		

After making a new dataframe that included only the Oscar winning movies, I created a time series that compared the length of the film over time since the first Oscar's Award Show. The time series is additive, with a slight downward sloping trend. It is relatively stationary, however until 2000 the trend was fairly random.

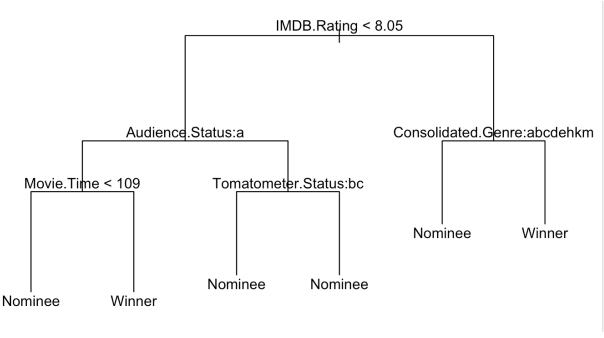


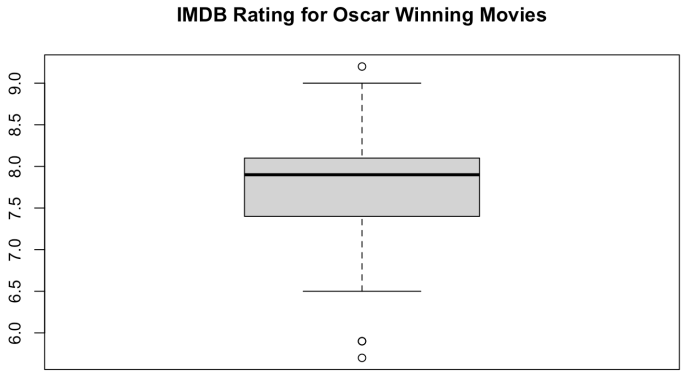
I made a single tree model using nine variables with the "Award" variable as the predictor. Because the tree was too busy and had so many variables that it was unreadable, I made a pruned tree to only show the most influential determinants. In both trees, IMDB rating, audience status (upright = good, spilled = bad), and the genre were most important in predicting.

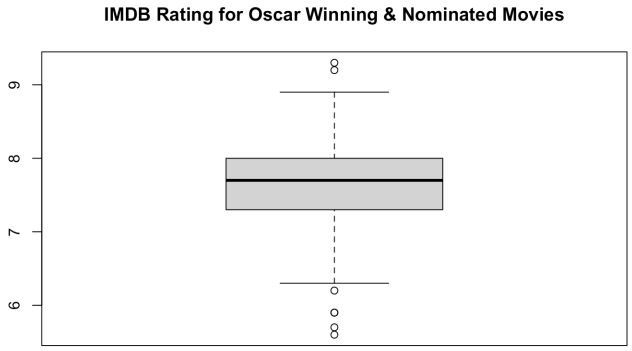
Single Tree



Pruned Single Tree

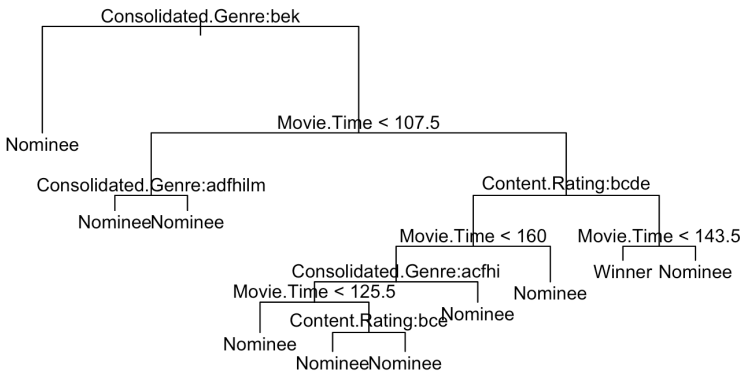


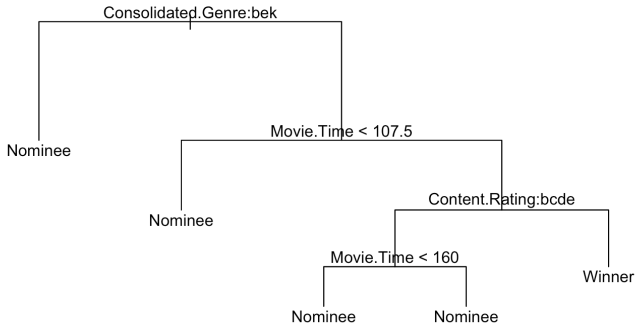




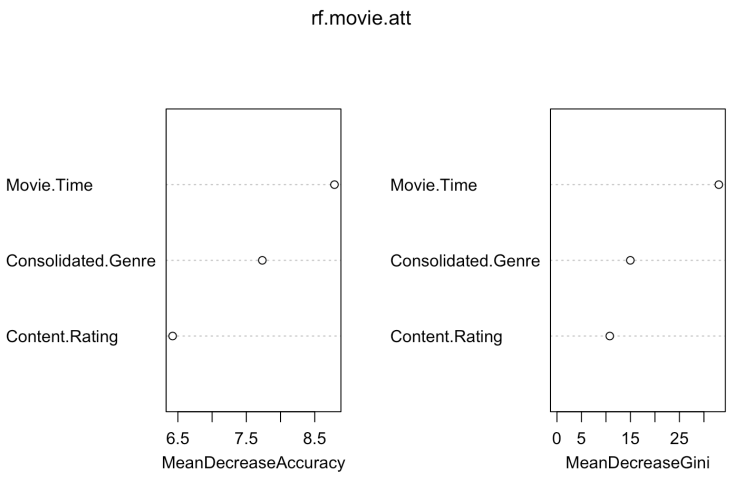
I made boxplots to compare the IMDB ratings for Oscar winning movies and IMDB ratings for Oscar winning as well as nominated movies. Although the winning movies have a slightly higher median rating and overall ratings, the difference is not as stark as one would expect. This is likely due to the fact that the people who pick the Oscar winners are professional film critics and any person can make an IMDB account.

After making a new data frame for only variables related to attributes of a movie (award, length of movie, consolidated genre column that I made, and content rating(G, NR, PG, etc)), I decided to make a single tree and then a pruned tree. I then made a random forest, which showed that the most influential variable was the length of the movie, with longer movies being more likely to win. This was further substantiated when I did a logistic regression as well as a step model which showed the only significant variable was the length of the movie with a very small p-value(0.000163).

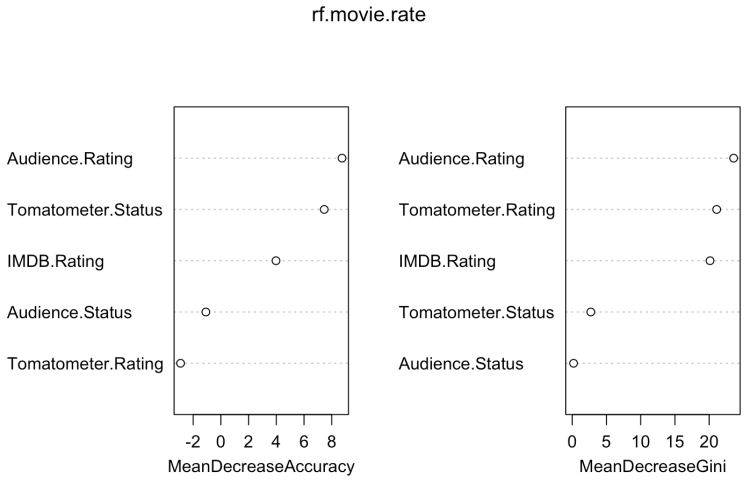




Random Forest for Movie Attributes



Random Forest for Movie Ratings



I made a new data frame that included variables only related to the ratings of a movie, consisting of award, IMDB rating, tomatometer status(certified fresh, fresh, and rotten), tomatometer rating, audience rating, and audience status. Audience rating had both a high mean decrease accuracy and high mean decrease gini, meaning that it was an important variable in the model. When doing a logistic regression and step model, the step model contained many significant variables, such as IMDB rating, tomatometer status fresh, and audience status upright.

Comparing the null deviance of the logistic regression model of movie attributes(422.8) to movie ratings (421.5), it is evident that both models have similar effectiveness at predicting an Oscar winning movie.

In summary, I believe that it is very difficult to predict Best Picture winners for the Oscars. Using random forests, logistic regression, and trees, the most important determinant that I found was the length of the movie as I did not find genre or rating to be as predictable. Audience ratings can sometimes be good predictors, however there is the possibility of bias that the viewer rated the movie after it received critical acclaim or that they judge movies differently than another viewer. I did not use directors, box office earnings, or production companies which

would help predict the likelihood of winning. From my findings, I deduce that the attributes of a movie are more likely to decide if a film will win the Best Picture Award compared to audience ratings.