Predicting the Best Director at the Oscars with Bayesian Logistic Regression

Glen Morgenstern and Jason Gerber

4/24/2021

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Name = col_character(),
##   Movie = col_character()
## )

## See spec(...) for full column specifications.
```

Introduction: A few paragraphs which (i) motivate problem importance & relevance

Every year, billions of people around the globe watch movies, and many tune in to watch the Academy Awards. This awards show is heavily publicized, promoted, and talked about every year, where millions of fans watch the ceremony and tens of millions(perhaps more) read about it later. Receiving an Oscar is the highest honor that one can procure in the business of movie-making, and since movies are such a large industry and many of our country's most influential celebrities are movie stars, the Oscars is one of the biggest events there is each year in the United States. All this attention inevitably leads to general curiosity and predictions about who will take home the most prized awards in all of filmmaking, which is where the reason for this project arises.

The goal of this project is to develop a model which will predict with as much accuracy as possible the winner of the "Best Director" category at the Oscars. Being able to predict Oscar winners is an analysis which many across the globe will have an interest in, whether they be data-lovers, typical cinephiles, or even gamblers who feel that they need an edge.

To achieve our goal, we will take several steps. First, we will perform exploratory data analysis on the data set we have in order to discover which variables have the potential to be explanatory variables for the model. Second, we will establish a clear prior for the model. Third, we will determine which explanatory variables should be used for the model by combining them with the prior to form a posterior, by creating several Bayesian logistical regression models and testing until we reach a model with optimal fit. Lastly, we will interpret the model coefficients and discuss its applications.

Data: A couple of paragraphs describing the data to be used. You may wish to

The data set that we are using for this analysis is one that contains information about all Oscar nominations from 1928-2006. The data has been taken from the personal website of Iain Pardoe, an online educator in mathematics and statistics, in the form of a csv, which we read into R and made into a data frame.

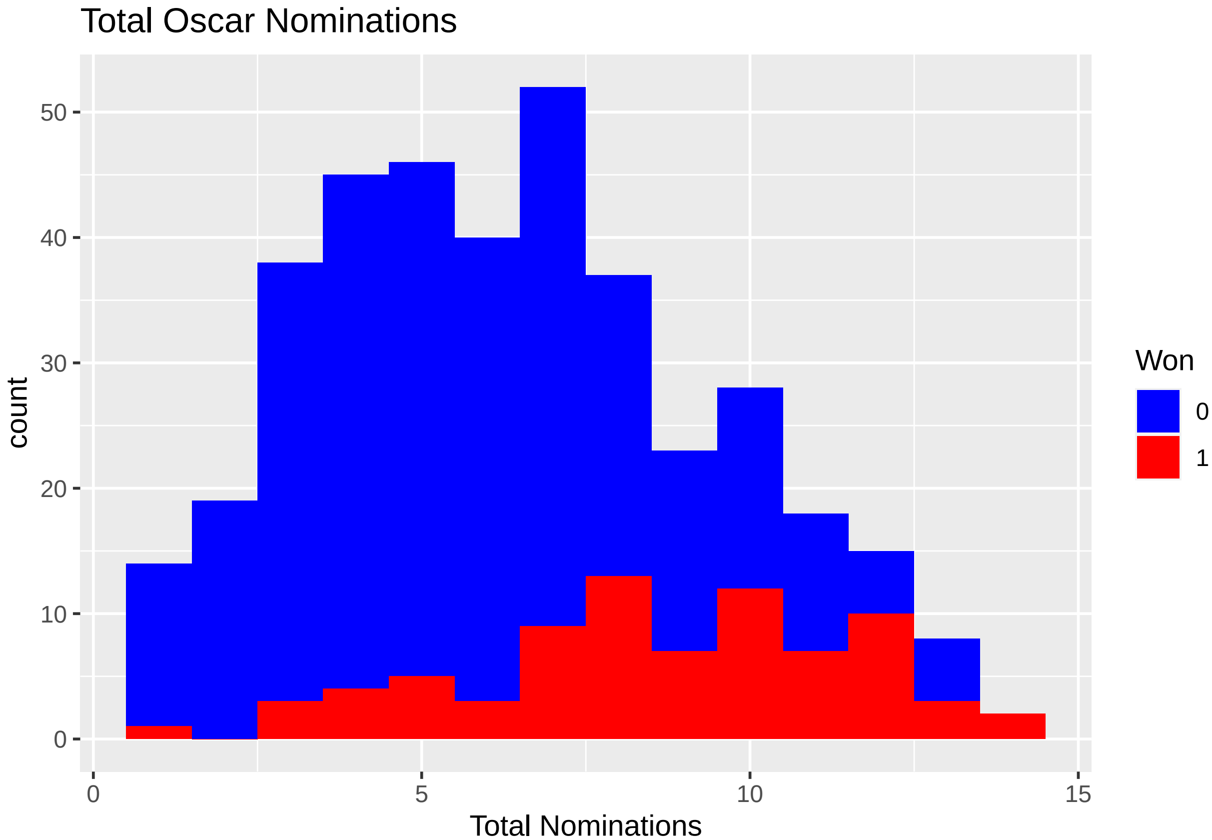
In our model, we will attempt to use all relevant variables that may be significant predictors of whether a director wins best director or not. The data set given has 62 variables, of which one is an indicator variable specifying if the nominee is up for Best Director, and one of which is an indicator variable specifying if the nominee won their award or not. For example, if Martin Scorsese is nominated for best director, the variable DD(director nominee indicator) will equal 1. We will create our data set using only the instances where DD is equal to 1, as we are only focusing on predicting winners among nominated directors.

To decide which variables to use, we performed EDA on many variables, but since we have limited space, we are only discussing the variables which might be relevant to our model. In other words, we did not perform EDA on variables like AD (Whether the Assistant Director was nominated) because it simply did not have any relevance or correlation with the director winning Best Director. This was true for all built-in interaction variables as well.

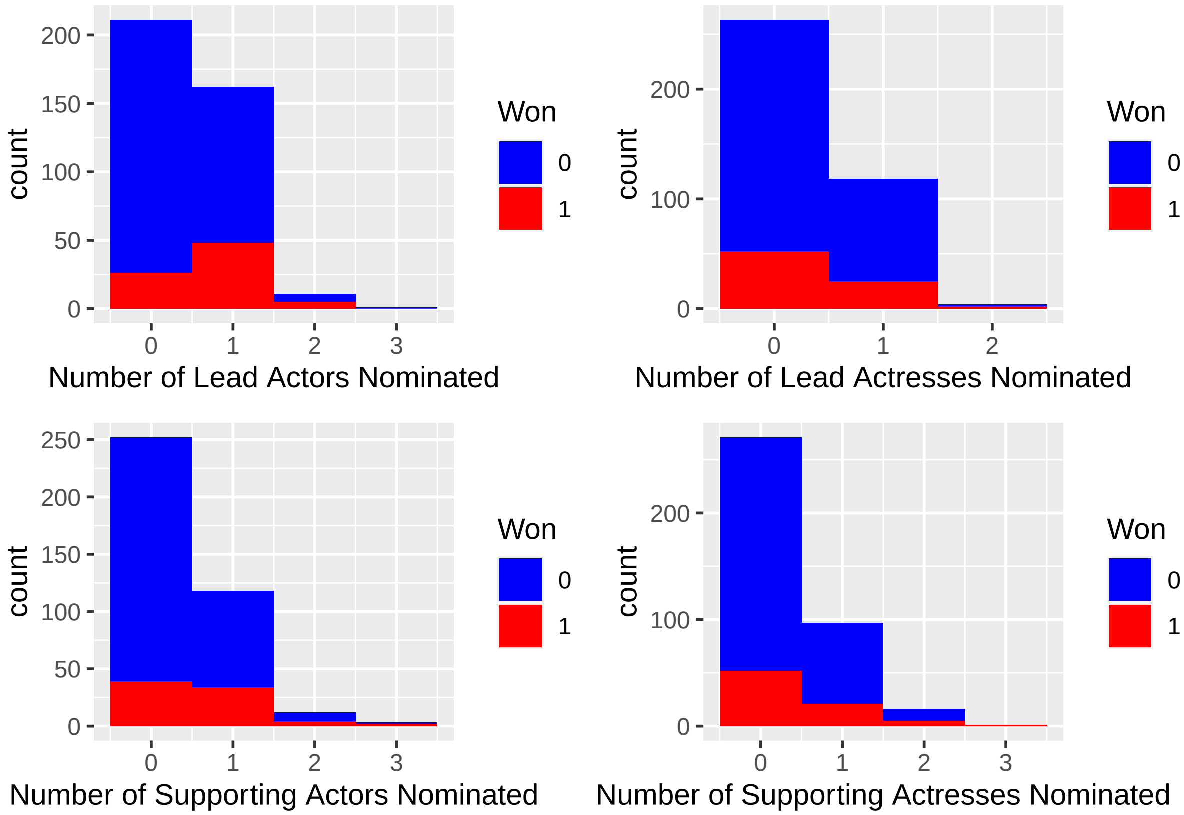
The variables that we performed EDA on are described below:

```
Nom: The total number of oscar nominations that the director's movie received.

Pic: An indicator variable that specifies if the director's movie was also nominated for best picture.

Aml and Afl: Variables that indicate how many lead actors or actresses in the director's movie were nomi

Ams and Afs: Variables that indicate how many lead actors or actresses in the director's movie were nomi

Scr: An indicator variable that specifies if the director's movie was also nominated for best screenpla

Cin: An indicator variable that specifies if the director's movie was also nominated for best cinematog

Art: An indicator variable that specifies if the director's movie was also nominated for best art direc

Cos: An indicator variable that specifies if the director's movie was also nominated for best costumes.

Sco: An indicator variable that specifies if the director's movie was also nominated for best muscial s

Edi: An indicator variable that specifies if the director's movie was also nominated for best editing.

Sou: An indicator variable that specifies if the director's movie was also nominated for best sound mix

Eff: An variable that specifies if the director's movie was nominated for sound editing, visual effects

PrN: The total number of oscar nominations that the director and actors for the movie had received prev

PrW: The total number of oscars that the director and actors for the movie had won previously.

Gd: An indicator variable that specifies if the director's movie won a Golden Globe for Directing earli

DGA: An indicator variable that specifies if the director won an award from the Director's Guild of Ame

Year: The Year that the movie was nominated for an Oscar.
```

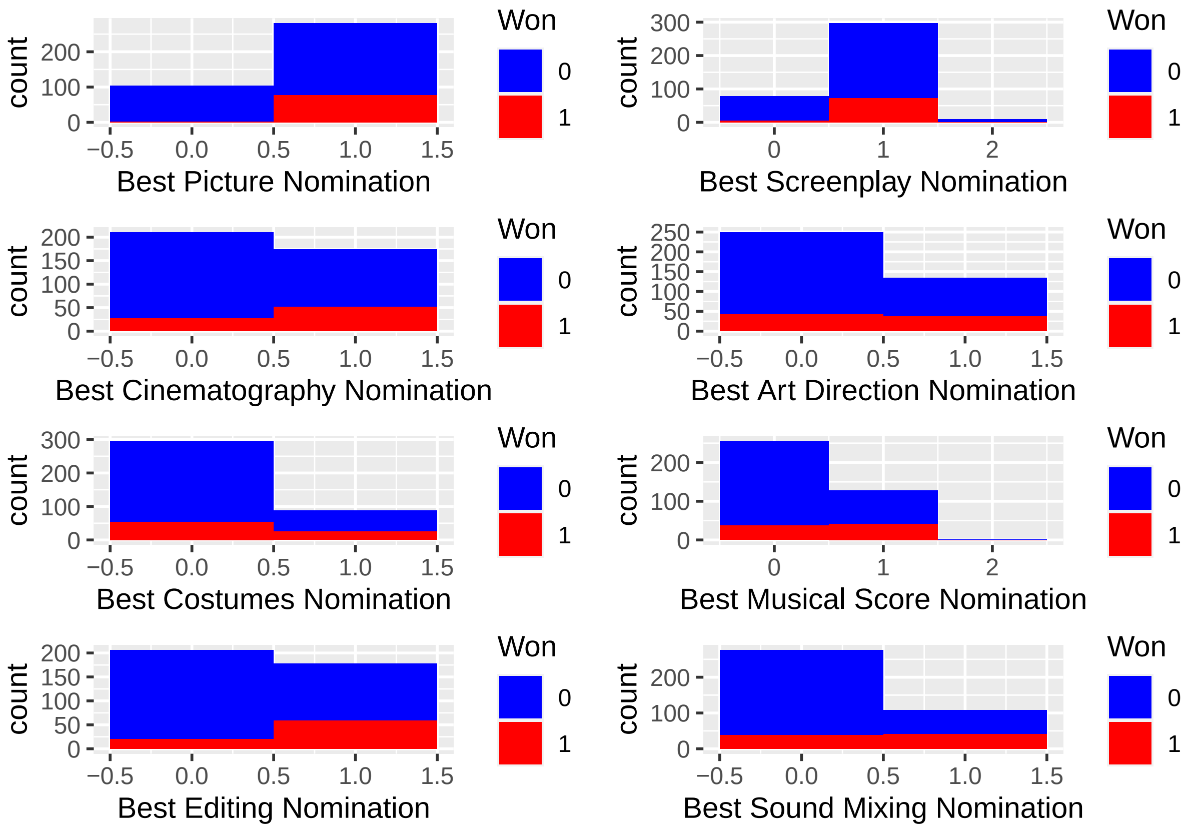
Exploratory Data Analysis



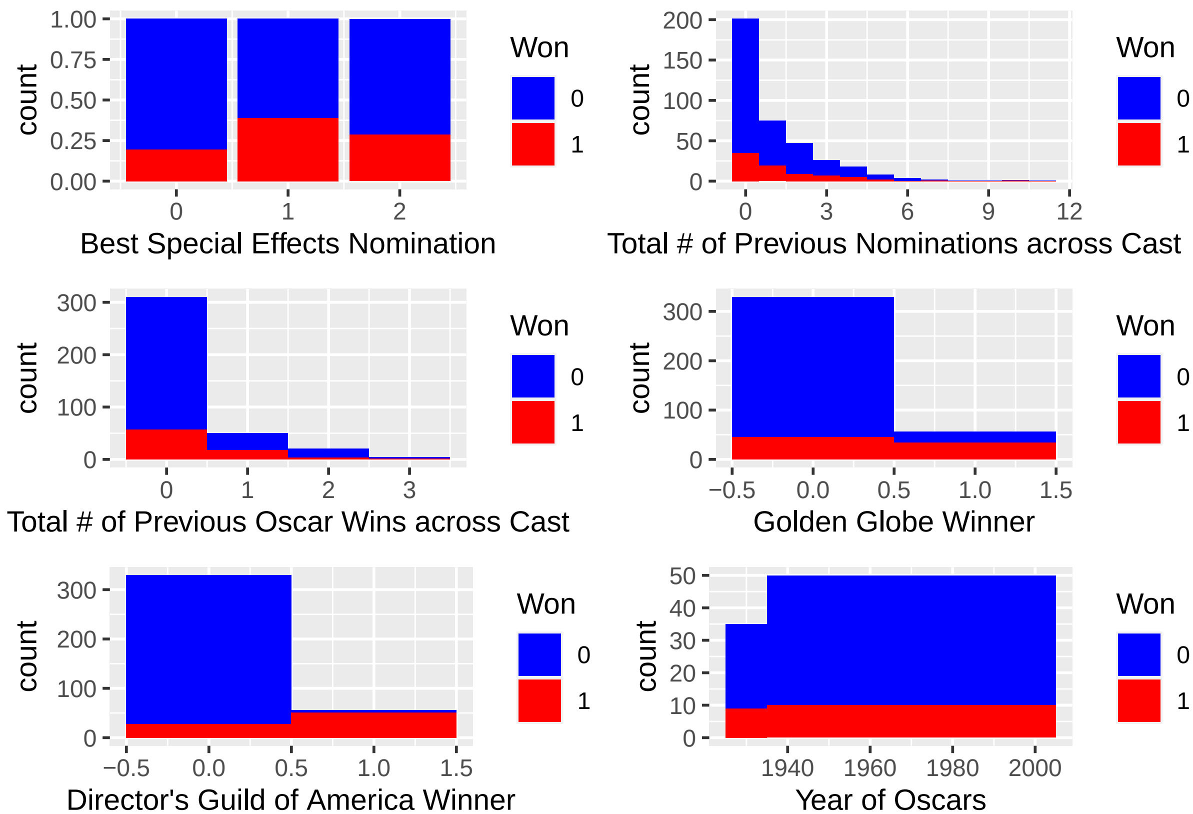
The above analysis makes it appear that the more nominations a movie has, the more likely the Director is to win an Oscar. This is enough correlation to warrant adding it to our initial model.



The only variable out of the above that may have some correlation is number of lead actors nominated. However, when added to the model, this was insignificant and increased looic, so we did not keep it in any model. The other three others(number of supporting actors, number of supporting actresses, and number of lead actors) do not have any obvious correlation with Director wins, so we shall not add them to the model.



Of the above, Picture, Cinematogrpahy, Editing, Sound Mixing, Musical Score and Screenplay all appear to have a much higher proportion of wins if they are nominated in any of those categories. Therefore, we shall add them to the initial model. Costumes and Art Direction seem rather irrelevant in terms of predicting the winner of best director, so we shall not add them.



Previous Oscar Wins, Previous Oscar Nominations, and Special Effects do not appear to be clear predictors of a Director win, so we will not add them to the model. However, Golden Globe Winner, and Director's Guild of America Winner, and Year do, so we shall add them.

Ultimately, we have chosen variables DGA, Year, Gd, Sou, Edi, Sco, Scr, Cin, Pic, and Nom as our possible predictors.

Model and parameters

We set a relatively weakly informative prior model of Normal (0, 10) for model coefficients. We do this because we have rather little practical prior information on the effects of each regressor. We also wanted to leave open the possibility that a coefficient could have a negative effect on the Oscar chances. In addition, Pardoe found that using more informative priors based on results from each previous year produced equivalent, but not better results.

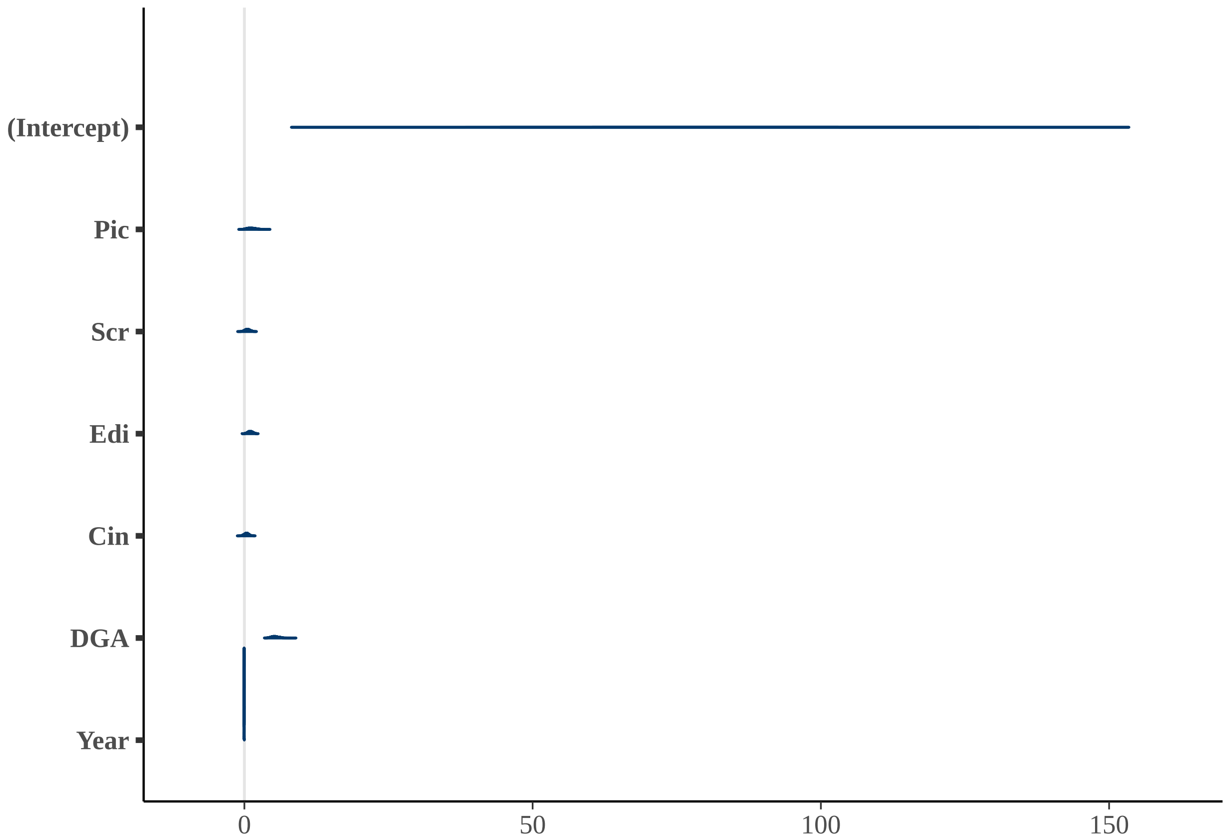
We assumed independent priors for the purpose of this problem because while we do not know much about the Academy itself, we assumed that no film companies bribed members to vote for a certain movie in all categories. The intercept prior was Normal(0,1) because it led to lower values of looic.

Our sampling model used a logit link rather than a probit link for ease of interpretation. The likelihood for a single observation under this model is:

We used the stan_glm function in the `rstanarm` package. This function performs full Bayesian estimation via the Markov Chain Monte Carlo method. Below is our final model:

$$\binom{n}{y}\left(\frac{e^{\eta}}{1+e^{\eta}}\right)^{y}\left(\frac{1}{1+e^{\eta}}\right)^{n-y}$$

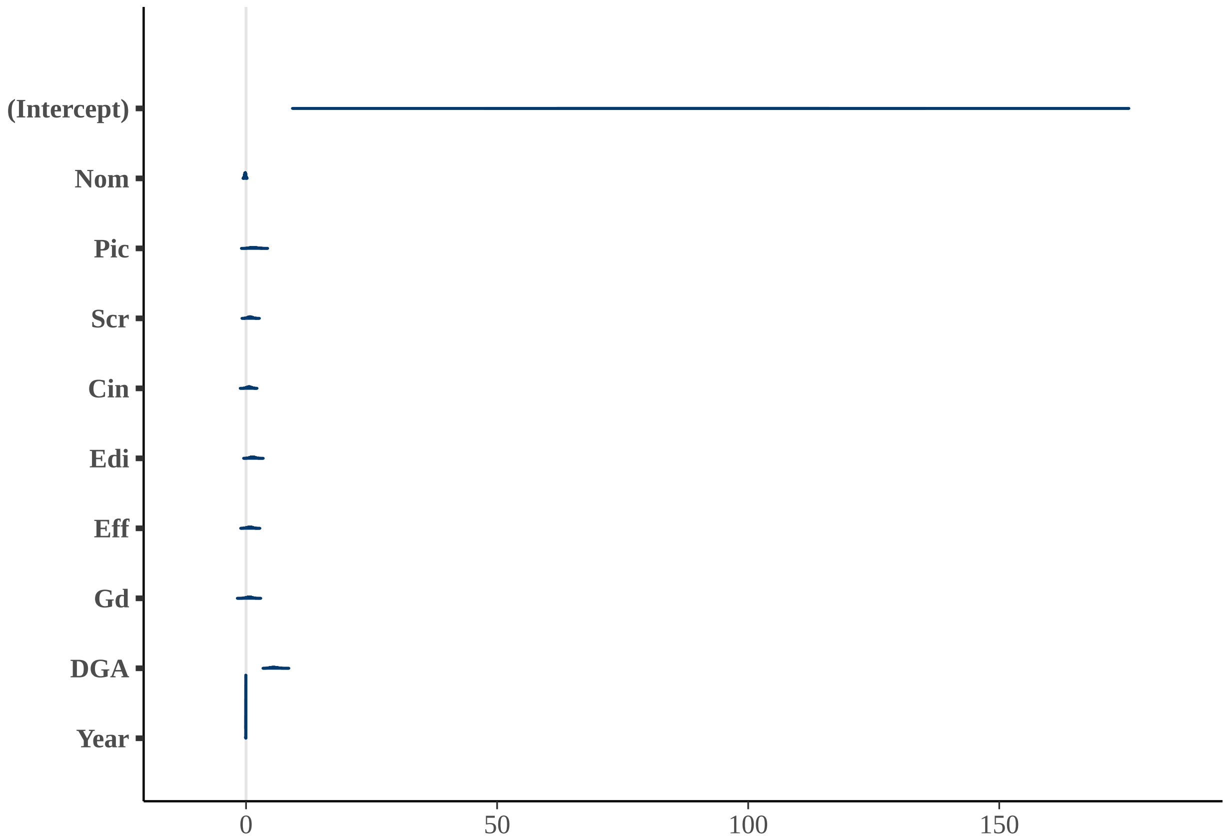
Figure 1: Caption for the picture.




```
## (Intercept)         Pic         Scr         Edi         Cin         DGA
##      82.384       1.159       0.522       1.024       0.379       5.281
##        Year
##      -0.044
```

```
##               2.5%   97.5%
## (Intercept) 44.380 127.487
## Pic         -0.078   2.759
## Scr         -0.294   1.372
## Edi          0.184   1.862
## Cin         -0.403   1.164
## DGA          4.121   6.681
## Year        -0.067  -0.025
```

This can be compared to a model with total nominations and whether the candidate won Golden Globe for Best Director.




```
## (Intercept)         Nom         Pic         Scr         Cin         Edi
##      89.223      -0.171       1.471       0.807       0.591       1.276
##         Eff          Gd         DGA        Year
##       0.734       0.635       5.502      -0.048
```

```
##               2.5%   97.5%
## (Intercept) 47.494 135.212
## Nom         -0.421   0.078
## Pic          0.060   3.189
## Scr         -0.089   1.776
## Cin         -0.305   1.521
## Edi          0.195   2.328
## Eff         -0.411   1.764
## Gd          -0.519   1.729
## DGA          4.218   7.011
## Year        -0.071  -0.027
```

Results

Posterior Predictive Diagnostics

```
##
## Computed from 4000 by 385 log-likelihood matrix
##
```

```
##          Estimate   SE
## elpd_loo   -101.5 12.4
## p_loo         7.0  1.3
## looic       202.9 24.7
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

In the code chunk above, we assessed the strength of our model via its posterior preditive LOOCV. However as we know, this accuracy rate is quite meaningless unless we have something to compare it to. So let's create a baseline model with no predictors to compare to our final model:

```
##
## Computed from 4000 by 385 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo   -367.5 0.5
## p_loo       105.1 0.2
## looic       735.1 1.1
## ------
## Monte Carlo SE of elpd_loo is 0.4.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
## Warning: 'rstanarm::compare_models' is deprecated.
## Use 'loo_compare' instead.
## See help("Deprecated")
```

```
## Warning: 'loo::compare' is deprecated.
## Use 'loo_compare' instead.
## See help("Deprecated")
```

```
## Model formulas:
##  :  NULL
##  :  NULLelpd_diff        se
##    -266.1      12.2
```

Loo2 (LOOCV for our final model) is much lower than the intercept-only model. And now we compare our final model with the model with the extra coefficients for total nominations and the Golden Globe.

```
##
## Computed from 4000 by 385 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -102.6 12.8
## p_loo        10.0  1.7
## looic       205.3 25.7
## ------
## Monte Carlo SE of elpd_loo is 0.1.
```

```
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.

## Warning: 'rstanarm::compare_models' is deprecated.
## Use 'loo_compare' instead.
## See help("Deprecated")

## Warning: 'loo::compare' is deprecated.
## Use 'loo_compare' instead.
## See help("Deprecated")

## Model formulas:
##  :  NULL
##  :  NULLelpd_diff        se
##      -1.2       2.2
```

Below, we compute posterior predictive probabilities of the linear predictor via the posterior_linpred() function provided in the rstanarm package. We calculate these posterior predictive probabilities in order to determine the classification accuracy of our model. This method gives us 91% accuracy.

```
## [1] 0.914
```

We could also evaluate the classification accuracy of our model on unseen data. This can be done via a LOOCV approach, as shown below. The results are similar to above, with about a 91% accuracy rate.

```
## [1] 0.914
```

Convergence diagnostics

Below are some traceplots of the variables in the final model. The model clearly mixes rather well and appears stationary.

Interpreting Model Coefficients

If all explanatory variables are equal to zero, the log-odds are equal to 82.384. The intercept from the model with no predictor variables is equal to 82.384 as an estimation for the log odds of winning Best Director.

Changing Pic from 0 to 1 and holding all else equal results in a change of 1.159 in the log-odds. This means that the film being nominated for best picture multiplies the log-odds of winning Best Director by 1.159 times what it would be if film was not nominated for best picture.

Increasing Scr from by 1 and holding all else equal results in a change of 0.522 in the log-odds. This means that the film being nominated for best screenplay multiplies the log-odds of winning Best Director by 0.522 times what it would be if film was not nominated for best screenplay.

Changing Edi from 0 to 1 and holding all else equal results in a change of 1.024 in the log-odds.This means that the film being nominated for best editing multiplies the log-odds of winning Best Director by 1.024 times what it would be if film was not nominated for best editing.

Changing Cin from 0 to 1 and holding all else equal results in a change of 0.379 in the log-odds.This means that the film being nominated for best cinematography multiplies the log-odds of winning Best Director by 0.379 times what it would be if film was not nominated for best cinematography.

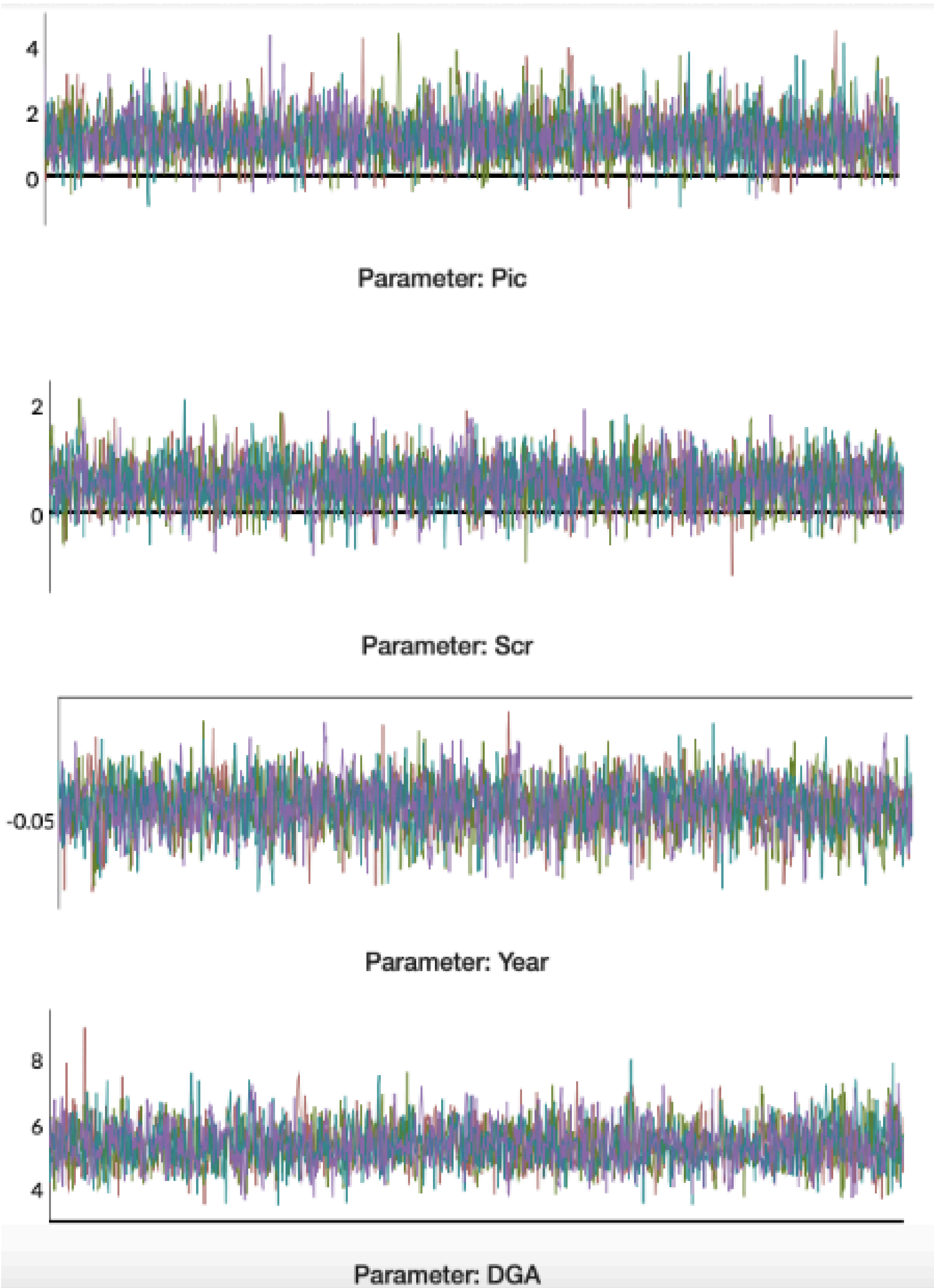


Figure 2: Caption for the picture.

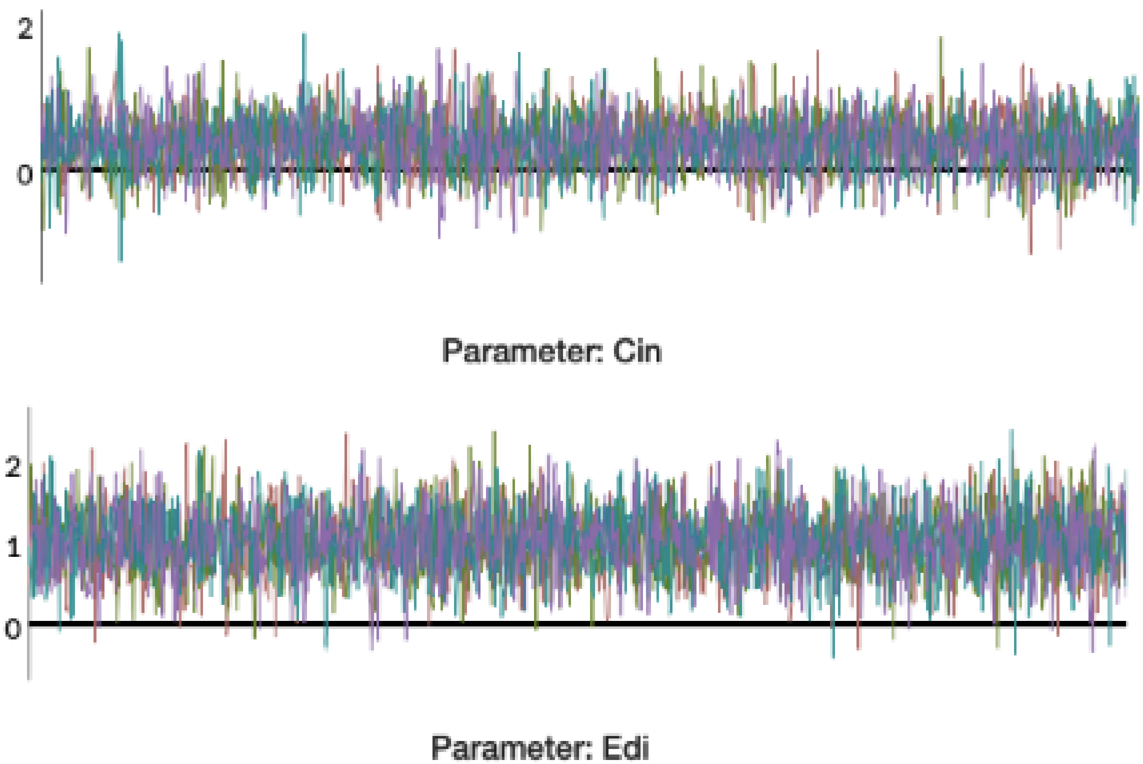


Figure 3: Caption for the picture.

Changing DGA from 0 to 1 and holding all else equal results in a change of 5.281 in the log-odds.This means that the Director winning at the Director's Guild of America multiplies the log-odds of winning Best Director by 5.281 times what it would be if the director did not win at the Director's Guild of America.

Increasing Year by 1 and holding all else equal results in a change of -0.044 in the log-odds. However, this is only because from 1932 to 1935, there were only three nominees for Best Director in each of these years.

Conclusion

The above model is highly accurate at 91% in predicting Best Director winners at the Academy Awards. This comes close to Pardoe's 93% figure. However, this is not very informative from a prediction standpoint. First of all, data range from 1928 to 2006, which leaves out more than a decade of rapid social change, including the #MeToo campaign. It is likely that this model would have changed since 2006.

The winner of the Golden Globe Best Director and the Director's Guild winner were highly multicollinear, which is why we only included Director's Guild data in our model. Even so, the winner of both of these awards is not quite the bellwether that some have thought it to be. Other characteristics of the director's movie are also very strong predictors of winning the award.