Data Science Project

Charl Schoeman - 21625247

Data Science 871 Project

24 June 2023

Can we predict Academy Awards Best Picture Winners?

Introduction

In this project, I will be using supervised learning machine learning methods to build a prediction model that predicts if films that were nominated for the Academy Award (Oscar) for Best Picture will win the award or not. Two different methods will be used to create two alternative models:

1) Linear Discriminant Analysis (LDA)
2) Random Forests (RF)

The results of the two models will be compared using a few metrics, including sensitivity, specificity, an ROC curve and kappa (a reliability measure). Based on these metrics, the most appropriate model will be identified.

This model will then be used to predict the 2023 Best Picture Winners, and then compare it to the actual winner of that year.

Basic Model Setup

The machine learning models will both have the same basic structure. A film either winning the Oscar for Best Picture (or not) will be used as the predicted variable, and a series of attributes of the film will be used as predictor variables. These attributes include the IMDB and Metacritic scores of the film, the amount of Oscars that the film was nominated for (and which specific Oscars), as well as if it won the Golden Globe or BAFTA award for Best Picture. To improve the skill of the models, the amount of Golden Globe awards (across categories) that the film won is also included.

IMDB and Metacritic scores are used as a measure of the critical and audience reeption of the films, which is believed to have a large influence on the film's chances of winning. These two scores were used instead of Rotten Tomatoes scores or those of other websites, due to the availability of data for them.

A film's chance of winning Best Picture is usually correlated with how many other Oscars it was nominated for. It is highly unlikely that a film that was only nominated for a single Oscar would be seen as the "best", so the amount of nominations is taken as a score of the Academy's view on the film's quality. Even though historical data is used, and thus data is available on whether the film ended up winning Oscars in other categories, only nomination data is used to allow predictions before the announcement of winners. Certain Oscar nominations are usually associated with Best Picture winners, for example almost all Best Picture winners in the past 20 years have also won or been nominated for Best Director. Thus it is also necessary

to include a measure of which specific Oscar categories the film is nominated for, in the form of binary indicators for each applicable category. Obviously this means that categories like Animation, Short Films and Documentaries are excluded as such films are not considered for Best Picture.

The Oscars is only one of many film awards ceremonies, and is held at the end of what is called "Awards Season". This means that, for each film, there is also data available for how it performed at previous awards ceremonies in that season. This is useful, as there is often an overlap between winners of different awards ceremonies. For this project, data from two high profile awards ceremonies that take place before the Oscars are used: The BAFTAs and the Golden Globes. Due to data availability, for the BAFTAs only an indicator for whether the film won Best Picture is used. Golden Globe data is more readily available, so measures of whether the film won Best Picture, as well as how many Golden Globes it won overall, are used.

Data

The data used come primarily from two sources. The Oscar nomination and winner data, as well as the Golden Globes winner data was found on Kaggle, an open source online data repository. IMDB scores, Metacritic scores and BAFTA winner data were obtained by webscraping IMDB pages using Octoparse. Only IMDB was used for webscraping as public use of its data is allowed.

The data was thoroughly cleaned, combined and processed through a combination of work in r and some work in Microsoft Excel, simply to easily combine the separate datasets. The code used to process the data is included in the README of this repository: https://github.com/CharlHS/Data_Science_871_21625247_ML_Project

Only films that were nominated for the Oscar for Best Picture were included in the final dataset, and all films that had some null values (like no available Metacritic or IMDB scores, which is quite common for very old movies) were removed. This should not be an issue for the predictive ability of the project, as this mainly affects films that were nominated early in the Oscars' history, and thus weren't covered by the Golden Globes or the BAFTAs.

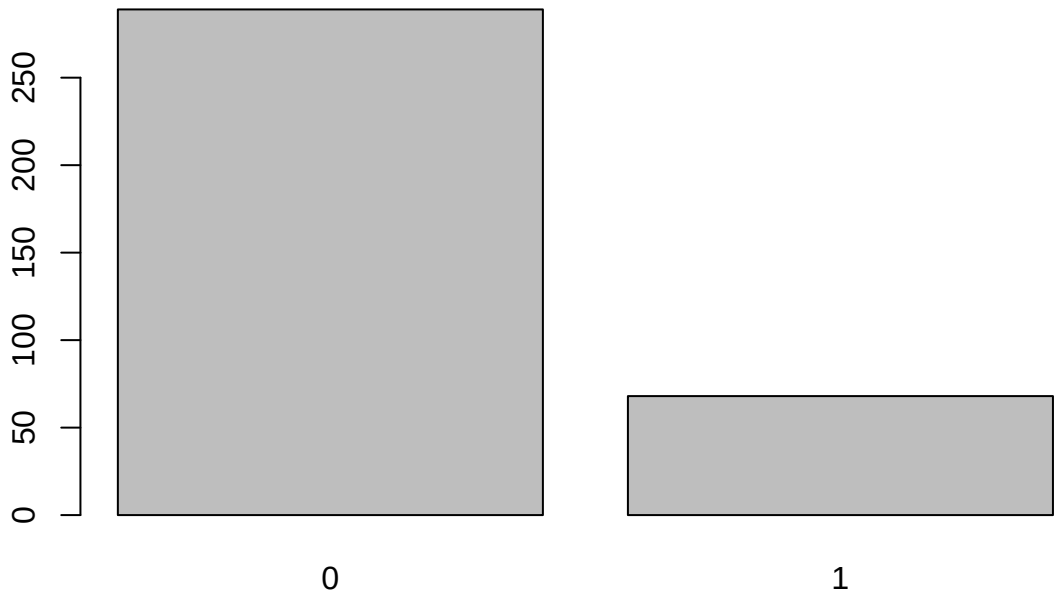
Preliminaries for Machine Learning training

As stated before, LDA and RF methods are used to train two possible prediction models. All training is done in R, using the caret package. Only the outputs of the machine learning steps and an overview of the coding is given in this document, but all steps and code are available in the README of this repository: https://github.com/CharlHS/Data_Science_871_21625247_ML_Project

The data is partitioned into training and validation data. 90% of the dataset is used as the training data, on which the two prediction models are trained, and the remaining 10% is used as the verification dataset, which is used to test the predictive ability of the models.

Below is a frequency table and a bar plot of the predicted variable won_bp within the training data, where 0 indicated that the film didn't win Best Picture and 1 indicates that it did:

```
##   freq percentage
## 0  289   80.95238
## 1   68   19.04762
```



This shows that the training data includes 357 films, of which 81% did not win Best Picture at the Oscars and 19% did. This large different in frequency between the two possible outcomes could be problematic, but unfortunately a lot more films are nominated than win, by design.

Models

Next, the models can be trained. For parameter tuning, 10-fold Cross Validation methods are used as controls.

```
# Build model:
# Run algorithm with 10-fold cross validation
control = trainControl(method="cv", number=10, search = "random",savePredictions = TRUE)
metric = "Accuracy"
```

First, the Linear Discriminant Analysis (LDA) model is trained. It is the simpler of the two models, and more computationally efficient.

```
# a) linear algorithms
set.seed(7)
fit.lda = train(won_bp~., data=dataset, method="lda", metric = metric, preProc=c("center", "scale") , t
```

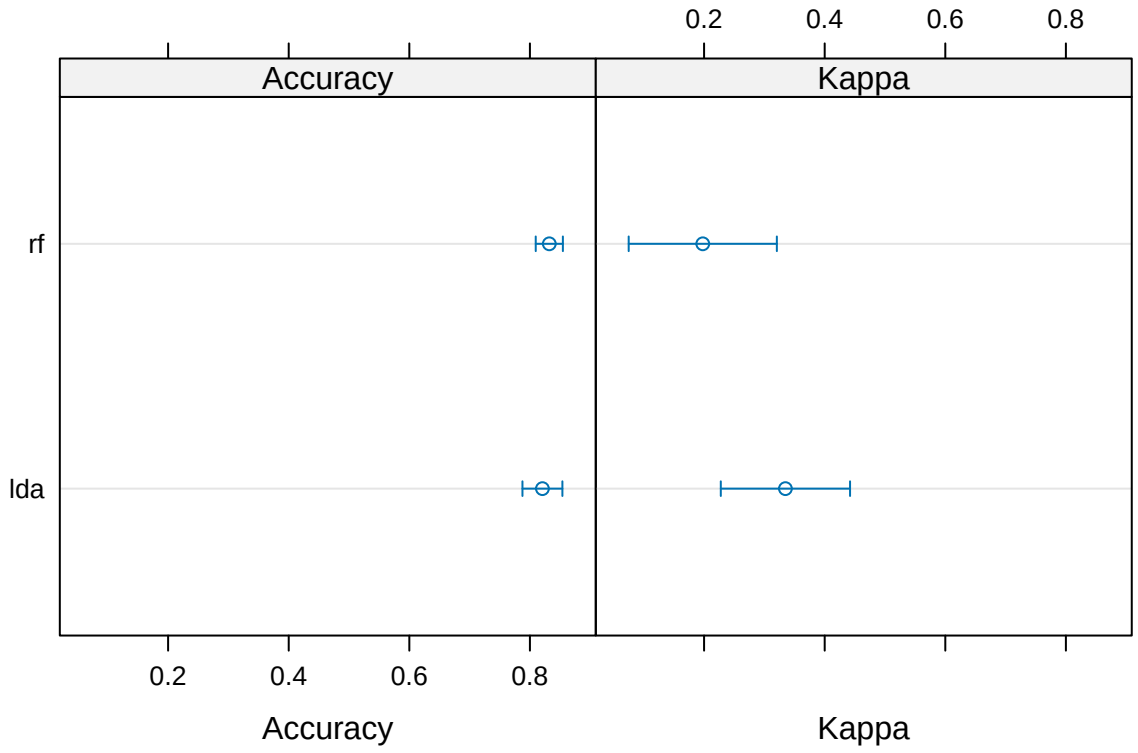
Next, the Random Forest (RF) model is trained. While it is a lot more computationally complex, it can in some cases be better at predicting values.

```r
# Random Forest
set.seed(7)
fit.rf = train(won_bp~., data=dataset, method="rf", metric=metric, preProc=c("center", "scale"), trCont
```

For both models, the same seed is set, to make comparison between the results of the two, as well as reproducibility.

After training the two models on the training data, it is time to compare the two models. We both summarise the results, as well as set up a plot of the accuracy and kappa measures of the two plots.

```
##
## Call:
## summary.resamples(object = results)
##
## Models: lda, rf
## Number of resamples: 10
##
## Accuracy
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda 0.7428571 0.8055556 0.8333333 0.8207143 0.8511905 0.8888889    0
## rf  0.8055556 0.8113095 0.8309524 0.8321429 0.8333333 0.9142857    0
##
## Kappa
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda 0.1509434 0.2234650 0.3322524 0.3348787 0.3956762 0.6022099    0
## rf  0.0000000 0.1486486 0.2111026 0.1978195 0.2116788 0.6236559    0
```



From this we can see that the LDA model seems to be a better fit. While the Accuracy of the LDA and RF models are quite similar and quite high, with narrow distributions, there is a large difference in the Kappa measures. While both have quite low Kappa values, the distribution of the LDA model has a better distribution, with a higher mean and median for Kappa. This seems to indicate that, between the two, the LDA might be a more appropriate model.

Comparison of the Confusion Matrices:

Next we compare the confusion matrices of the two. From this we can compare the sensitivity and specificity measure of the two models.

Estimated Skill of the LDA model:

```
## Linear Discriminant Analysis
##
## 357 samples
##  22 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (22), scaled (22)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 322, 321, 321, 322, 321, 321, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8207143  0.3348787
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 32  3
##          1  0  4
##
##                Accuracy : 0.9231
##                  95% CI : (0.7913, 0.9838)
##     No Information Rate : 0.8205
##     P-Value [Acc > NIR] : 0.06273
##
##                   Kappa : 0.6863
##
##  Mcnemar's Test P-Value : 0.24821
##
##             Sensitivity : 1.0000
##             Specificity : 0.5714
##          Pos Pred Value : 0.9143
##          Neg Pred Value : 1.0000
##              Prevalence : 0.8205
##          Detection Rate : 0.8205
##    Detection Prevalence : 0.8974
##       Balanced Accuracy : 0.7857
##
##        'Positive' Class : 0
```

```
##
```

Estimated Skill of the RF model:

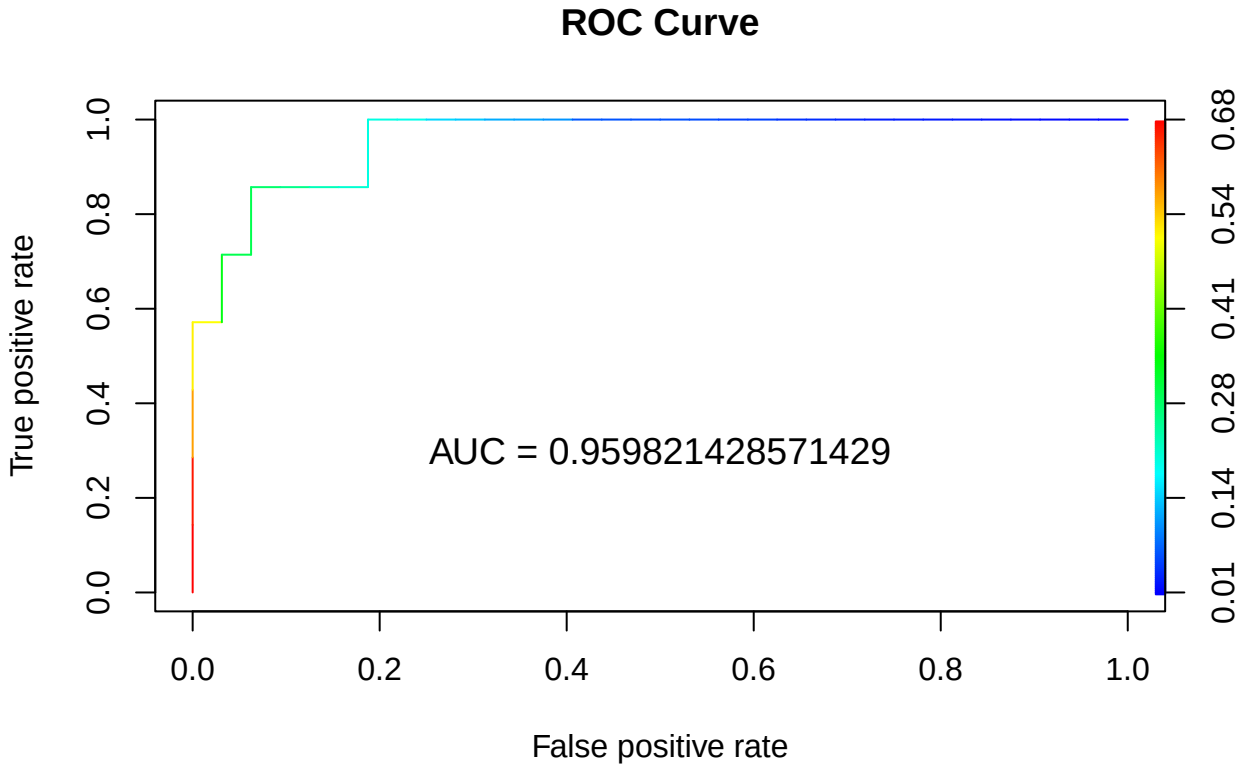
```
## Random Forest
##
## 357 samples
##  22 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (22), scaled (22)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 322, 321, 321, 322, 321, 321, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8321429  0.1978195
##    7    0.8154762  0.2198706
##   15    0.8125397  0.2587761
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 32  7
##          1  0  0
##
##                Accuracy : 0.8205
##                  95% CI : (0.6647, 0.9246)
##     No Information Rate : 0.8205
##     P-Value [Acc > NIR] : 0.59920
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 0.02334
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.8205
##          Neg Pred Value :    NaN
##              Prevalence : 0.8205
##          Detection Rate : 0.8205
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

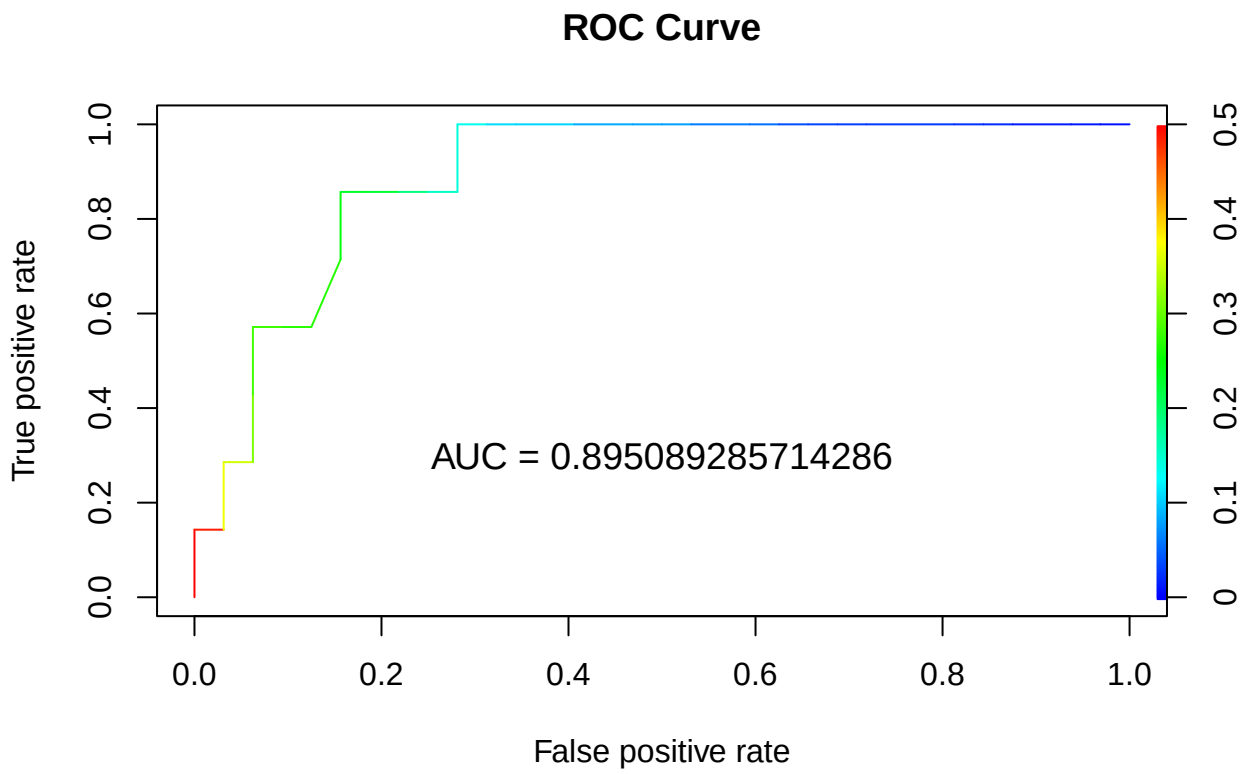
From the confusion matrices of the two, we can see that, while the RF model has higher sensitivity, it lacks adequate specificity. This means that the RF model is far more likely to give false negatives in its predictions. The LDA model is a lot more accurate in general, so it seems to be a better fit. The Kappa value is much higher for the LDA model than for the RF model, indicating better fit as well.

ROCs:

And finally, it is necessary to compare the ROC curves of the two models. The important measure here is the AUC measure (Area under curve). This indicates the tradeoff between the models' true positive rate to its false positive rate. The higher the AUC, the better.



LDA model:



RF model:

From this we can see that the LDA model is definitely more fitting. The AUC of the LDA is much larger, so it has a better trade-off of specificity and sensitivity.

Testing Predictions

Based on the results and metrics seen in the previous sections, the LDA model is most appropriate to predict actual Oscar Best Picture winners. In this section, a dataset of the 2023 Best Picture nominees (on which the model has not been trained) will be used to predict which film should have won, according to the model. Here is a table of the films' attributes:

```
##                              Film IMDB Metacritic Bafta Golden_Globe
## 1  Everything Everywhere All at Once  7.8         81    No           No
## 2     All Quiet on the Western Front  7.8         76   Yes           No
## 3          Avatar: The Way of Water   7.6         67    No           No
## 4          The Banshees of Inisherin  7.7         87    No          Yes
## 5                              Elvis  7.3         64    No           No
## 6                      The Fabelmans  7.6         84    No          Yes
## 7                                Tar  7.5         92    No           No
## 8                  Top Gun: Maverick  8.3         78    No           No
## 9                Triangle of Sadness  7.3         63    No           No
## 10                     Women Talking  6.9         79    No           No
##    Number_of_Golden_Globes Oscar_Nominations Best_Picture
## 1                        2               10          Yes
## 2                        0                8           No
## 3                        0                4           No
```

```
## 4                        3                8          No
## 5                        1                8          No
## 6                        2                7          No
## 7                        1                6          No
## 8                        0                6          No
## 9                        0                3          No
## 10                       0                2          No
```

As you can see, the actual winner was Everything Everywhere All at once. Compare this to the prediction from the model:

```
## [1] "The Banshees of Inisherin"
```

As can be seen, the model predicts that The Banshees of Inisherin should have been the winner, while this was obviously not the case. This does not necessarily mean that the model is completely inaccurate, as this films does, on paper, look quite similar to Everything Everywhere all at Once. It won more Golden Globes (3 vs 2), won the Golden Globe for Best Drama, has a similar IMDB rating and even a higher Metacritic score. Everything Everywhere all at Once did however have a lot of other attributes that aren't captured in the data, i.e. having a larger cultural impact, having a female minority lead, and being directed and written by minority individuals. This shows that there will always be some features that can't necessarily captured by numerical factors. This leaves an opportunity for further improvements to the model, namely adding some more qualitative factors beyond awards.

Conclusion

In conclusion, we can see that the LDA model is a better fit, but can still be improved. One possible way to do this would be to expand the data used in the model. One way to do this would be to add other predictor variables to the model, for example Rotten Tomatoes scores or more pre-Oscar awards ceremony data. This can help improve the model accuracy, and might even cut down its issues with false positive predictions. More sophisticated algorithms could also help improve the model.

Overall, the LDA model was not successful at predicting the 2023 Best Picture winner. This is likely due to attributes outside the model, but could also be due to flaws within the model itself. Further improvements, including adding more qualitative attributes to the model, could be beneficial to improving model quality and improving predictions.