CMSE 381 Honors Project

Josh Wolf

Professor Munch

April 28, 2022

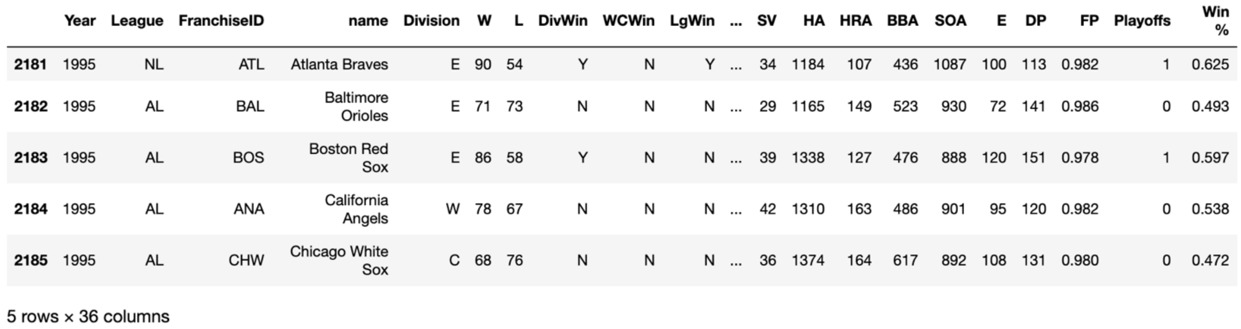
# Analyzing Team Success in Major League Baseball

## INTRODUCTION

Major League Baseball is a league centered around statistics. Since the 1800s, teams and scorekeepers have been tracking meaningful and relevant data. With all this data available, many questions are sparked and answered. The question I decided to focus on centers around winning and team success. In this project, I analyzed what affects winning the most in the MLB. More specifically, I looked at win totals while also focusing on making the playoffs as quantifying "winning".

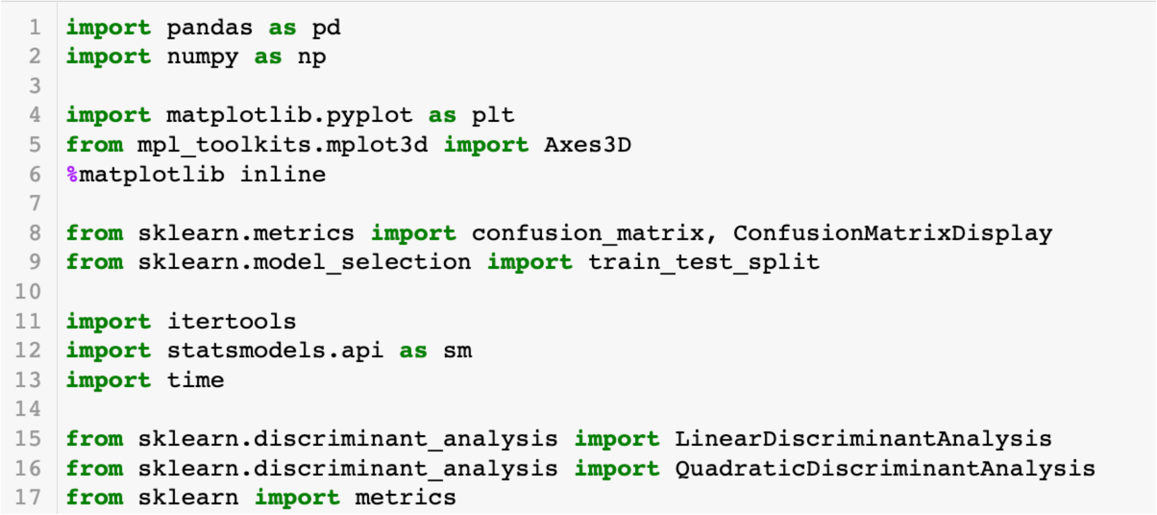
## DATA and PACKAGES

The entirety of this project was completed using Python through Jupyter Notebooks. The first step before any analyzing was making sure I imported all necessary packages. These imports included Pandas, Numpy, Matplotlib, various tools from Scikit Learn, and more. The next task was to properly load in the data and mask the dataframe as necessary. I dropped all the categorical variables that did not make sense to use, such as team name, ballpark name, and attendance. I filtered the data to only include years from 1995 and on. This is because the playoff format has dramatically changed since the early years of the MLB. I also had to drop all team stats from 2020 since it was a shortened season (teams played 60 games instead of the typical 162-game schedule). Some other data cleaning included creating a binary variable for playoff appearance and a win percentage variable. Shown below are the first 5 rows of the dataframe after the cleaning and filtering, and all the imports used in the project.



| | Year | League | FranchiseID | name | Division | W | L | DivWin | WCWin | LgWin | ... | SV | HA | HRA | BBA | SOA | E | DP | FP | Playoffs | Win % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2181** | 1995 | NL | ATL | Atlanta Braves | E | 90 | 54 | Y | N | Y | ... | 34 | 1184 | 107 | 436 | 1087 | 100 | 113 | 0.982 | 1 | 0.625 |
| **2182** | 1995 | AL | BAL | Baltimore Orioles | E | 71 | 73 | N | N | N | ... | 29 | 1165 | 149 | 523 | 930 | 72 | 141 | 0.986 | 0 | 0.493 |
| **2183** | 1995 | AL | BOS | Boston Red Sox | E | 86 | 58 | Y | N | N | ... | 39 | 1338 | 127 | 476 | 888 | 120 | 151 | 0.978 | 1 | 0.597 |
| **2184** | 1995 | AL | ANA | California Angels | W | 78 | 67 | N | N | N | ... | 42 | 1310 | 163 | 486 | 901 | 95 | 120 | 0.982 | 0 | 0.538 |
| **2185** | 1995 | AL | CHW | Chicago White Sox | C | 68 | 76 | N | N | N | ... | 36 | 1374 | 164 | 617 | 892 | 108 | 131 | 0.980 | 0 | 0.472 |

5 rows × 36 columns



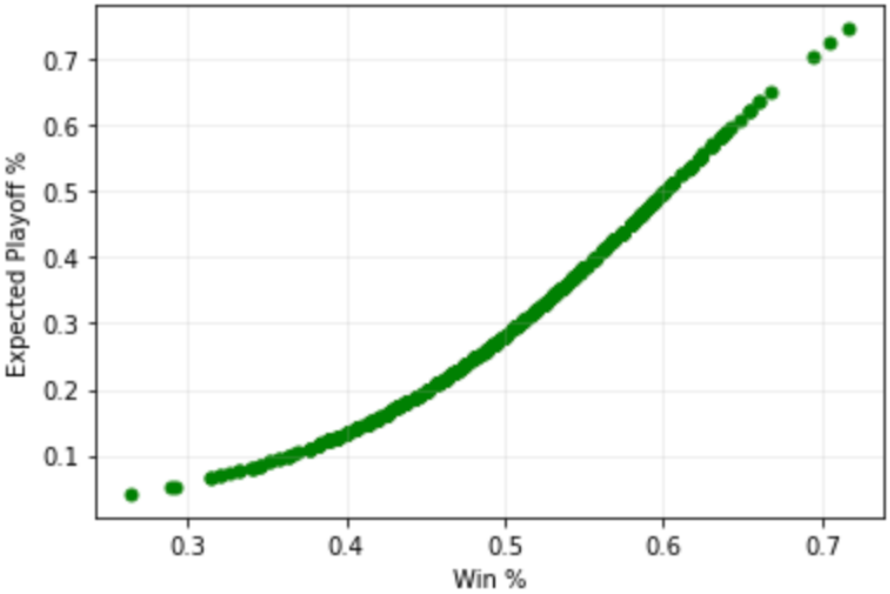
```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
%matplotlib inline

from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.model_selection import train_test_split

import itertools
import statsmodels.api as sm
import time

from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn import metrics
```

## METHODOLOGY

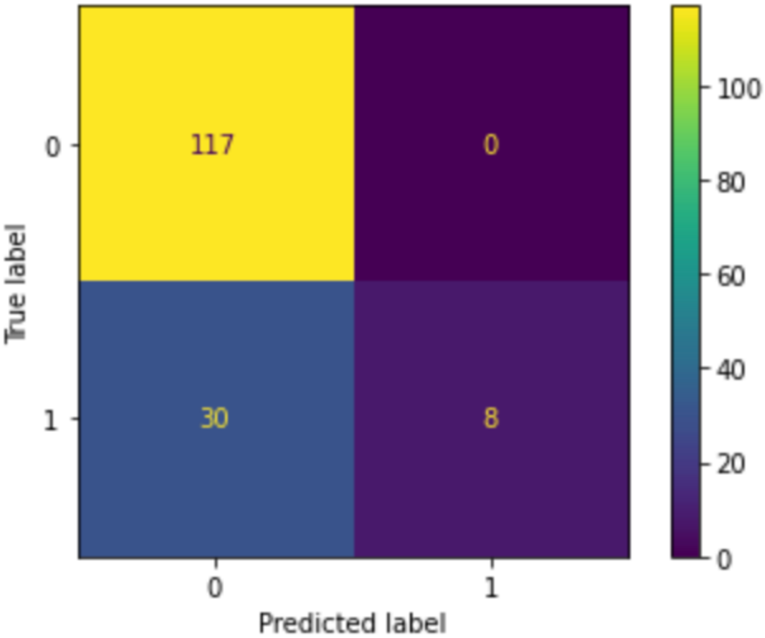
The first method I looked at was logistic regression, using win percentage to predict if a team made the playoffs. I felt this would be a good method to look at, as the playoff variable was binary which satisfies the main assumption of logistic regression. I used a train/test split to get the most out of the model. The next part of my project was finding a *subset* of variables to analyze winning. To do this, I utilized best subset selection and forward selection. These processes are both very applicable because my dataset had many predictors, and it would have been computationally expensive, exhausting, and nearly impossible to manually test all the different subsets. Once I found the ideal subset to predict winning, I created models using the subset of predictors for linear discriminant analysis and quadratic discriminant analysis. For both models, I am under the assumption that the response is gaussian. The only difference in assumptions is that for LDA, the covariance matrix must be the same for both playoff and non-playoff teams, whereas for QDA the covariance matrix between the response classes is different.

## RESULTS

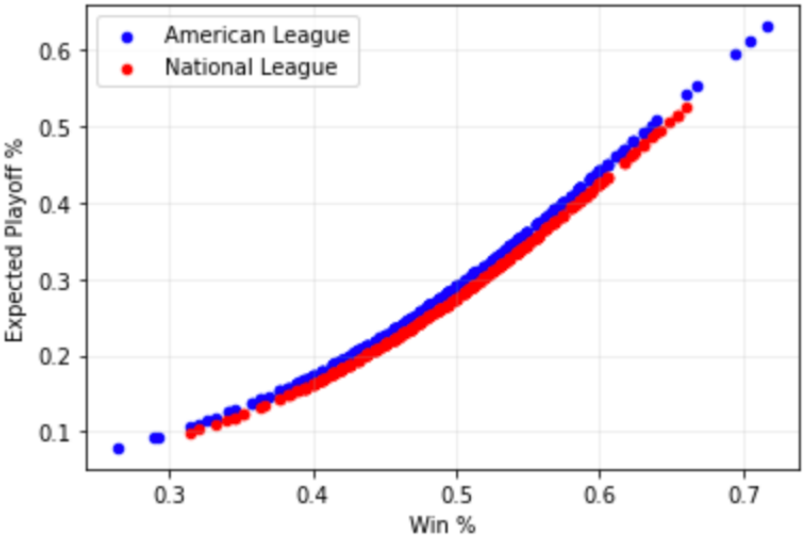
Starting with the logistic regression, the model performed fairly well considering I only used one predictor. The test score of the model was .806. The one thing I was a little disappointed in was the shape of the plot using the coefficient (9.31) and intercept (-5.59). I was hoping for more of a distinct logarithmic shape to separate the two responses.



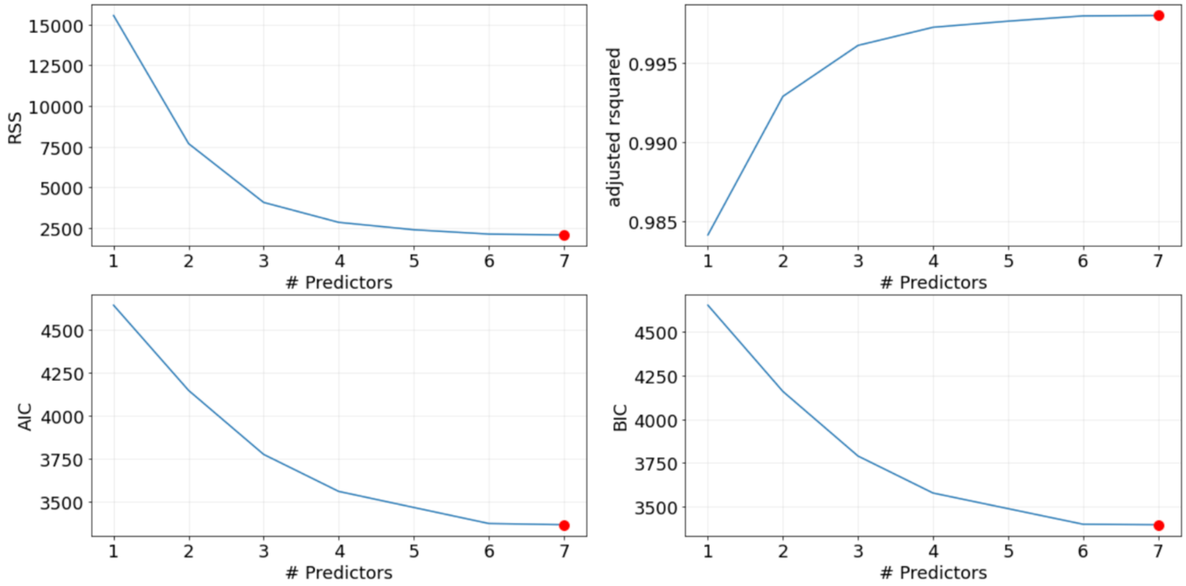
In addition to computing the score of the model, I also created a confusion matrix to show the predictions vs. actual. The model was heavy to predict non-playoffs, which resulted in quite a few false negatives. There were no false positives.

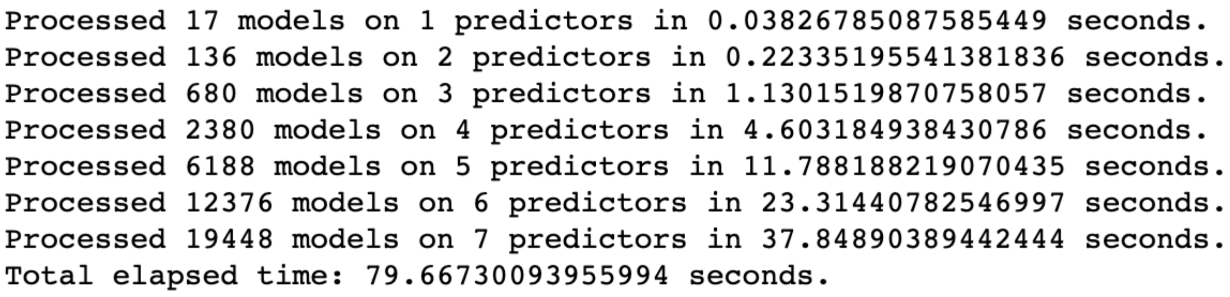


Out of my own curiosity, I also decided to see if there was a noticeable difference between the two leagues within the MLB. The plot below shows the difference between the American League and National League. There was not much of a difference, however the AL does require a higher win percentage, on average, to make the playoffs.



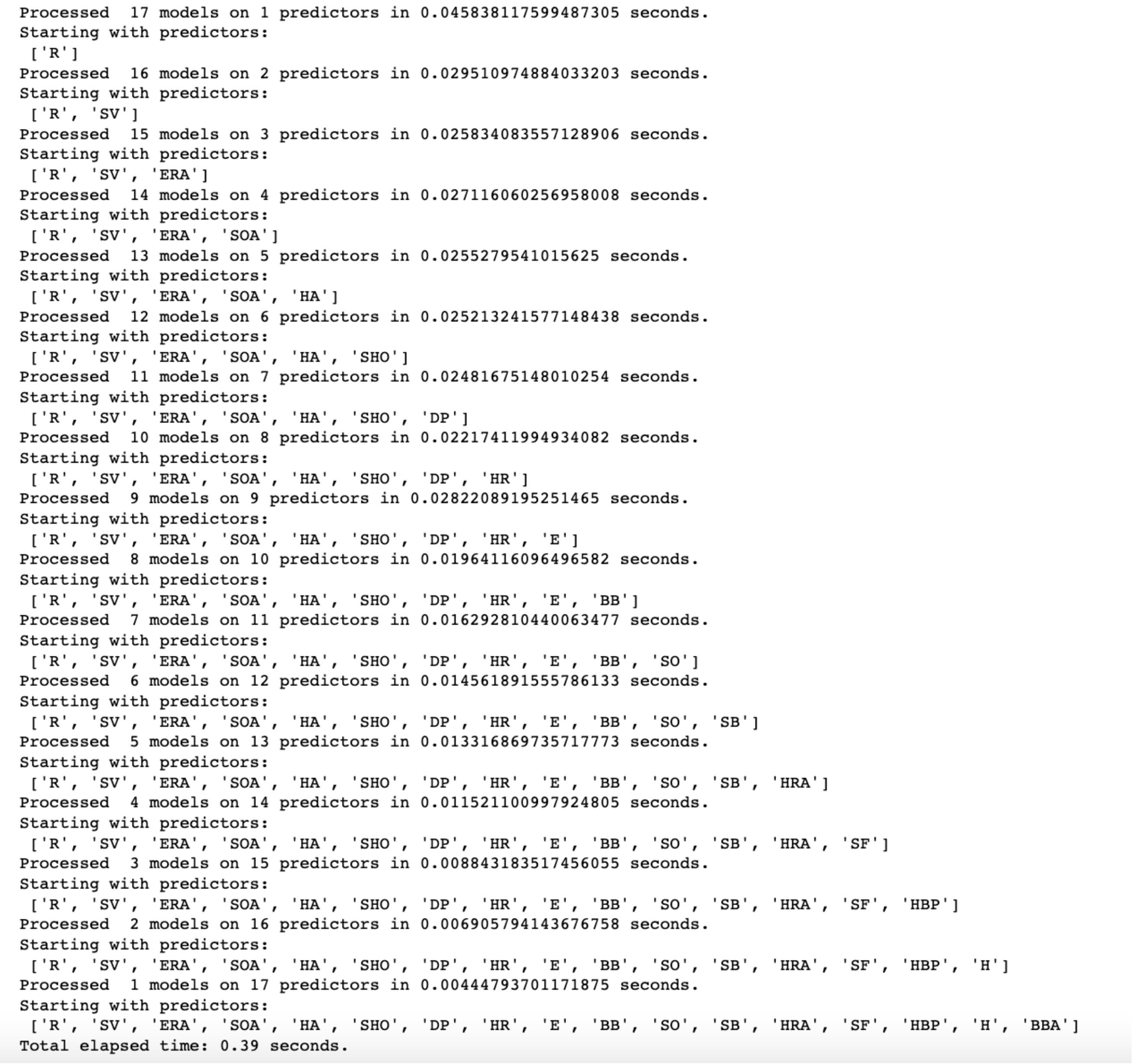
The results of my best subset selection and forward selection turned out great. For best subset, I ran over 40,000 models. After looking at the $R^2$, RSS, AIC, and BIC, I determined that it was most efficient to use the model with 4 predictors in the subset. The variables determined were Runs, Earned Run Average (ERA), Saves, and Strikeouts (pitching). As expected, each of these variables had p-values of less than .05, and they combined for an $R^2$ value of .997.



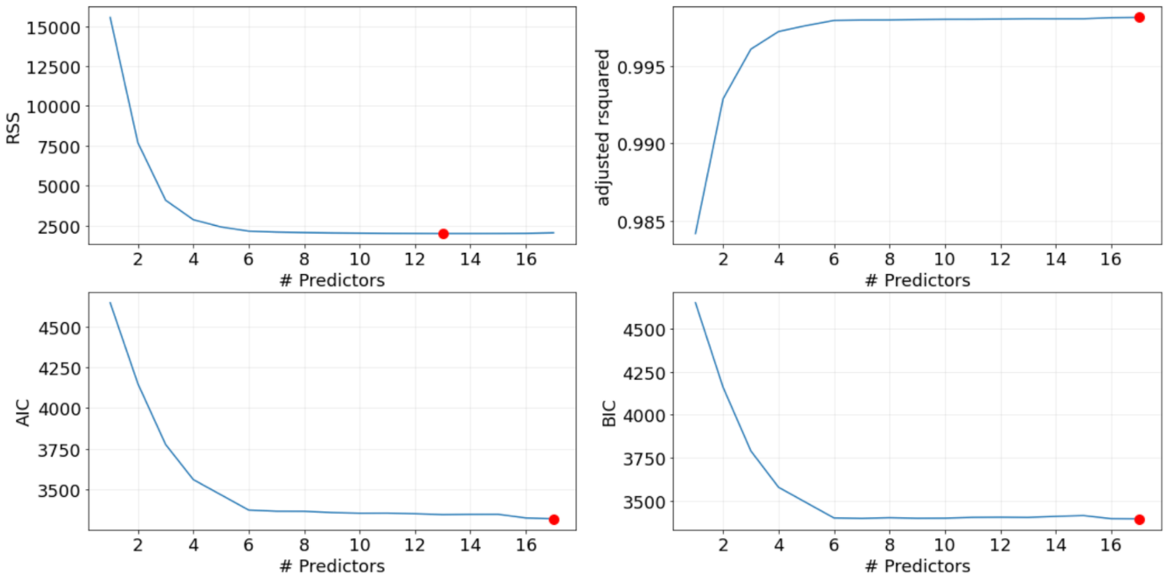


```
Processed 17 models on 1 predictors in 0.03826785087585449 seconds.
Processed 136 models on 2 predictors in 0.22335195541381836 seconds.
Processed 680 models on 3 predictors in 1.1301519870758057 seconds.
Processed 2380 models on 4 predictors in 4.603184938430786 seconds.
Processed 6188 models on 5 predictors in 11.788188219070435 seconds.
Processed 12376 models on 6 predictors in 23.31440782546997 seconds.
Processed 19448 models on 7 predictors in 37.84890389442444 seconds.
Total elapsed time: 79.66730093955994 seconds.
```

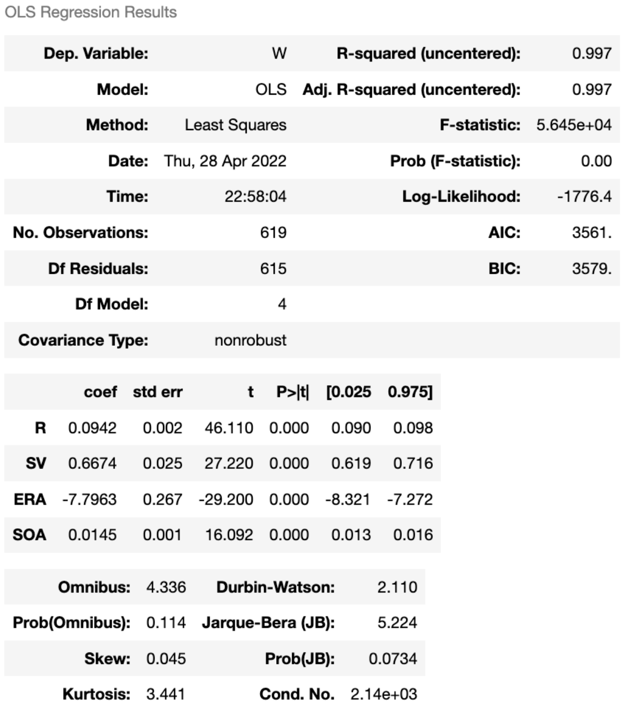
Following best subset, I was curious to see if forward selection would produce similar results. The results were nearly identical, which I was happy to see. There was a little bit more variation with the adjusted R$^2$ and RSS. Below are the steps for the forward selection. Notice the model with 4 predictors is the exact same model chosen by the best subset selection!



```
Processed  17 models on 1 predictors in 0.045838117599487305 seconds.
Starting with predictors:
 ['R']
Processed  16 models on 2 predictors in 0.029510974884033203 seconds.
Starting with predictors:
 ['R', 'SV']
Processed  15 models on 3 predictors in 0.025834083557128906 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA']
Processed  14 models on 4 predictors in 0.027116060256958008 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA']
Processed  13 models on 5 predictors in 0.0255279541015625 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA']
Processed  12 models on 6 predictors in 0.025213241577148438 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO']
Processed  11 models on 7 predictors in 0.02481675148010254 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP']
Processed  10 models on 8 predictors in 0.02217411994934082 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR']
Processed  9 models on 9 predictors in 0.02822089195251465 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E']
Processed  8 models on 10 predictors in 0.01964116096496582 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E', 'BB']
Processed  7 models on 11 predictors in 0.016292810440063477 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E', 'BB', 'SO']
Processed  6 models on 12 predictors in 0.014561891555786133 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E', 'BB', 'SO', 'SB']
Processed  5 models on 13 predictors in 0.013316869735717773 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E', 'BB', 'SO', 'SB', 'HRA']
Processed  4 models on 14 predictors in 0.011521100997924805 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E', 'BB', 'SO', 'SB', 'HRA', 'SF']
Processed  3 models on 15 predictors in 0.008843183517456055 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E', 'BB', 'SO', 'SB', 'HRA', 'SF', 'HBP']
Processed  2 models on 16 predictors in 0.006905794143676758 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E', 'BB', 'SO', 'SB', 'HRA', 'SF', 'HBP', 'H']
Processed  1 models on 17 predictors in 0.00444793701171875 seconds.
Starting with predictors:
 ['R', 'SV', 'ERA', 'SOA', 'HA', 'SHO', 'DP', 'HR', 'E', 'BB', 'SO', 'SB', 'HRA', 'SF', 'HBP', 'H', 'BBA']
Total elapsed time: 0.39 seconds.
```

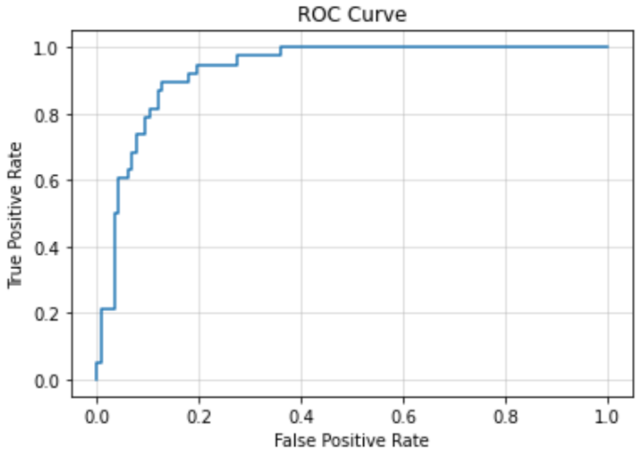


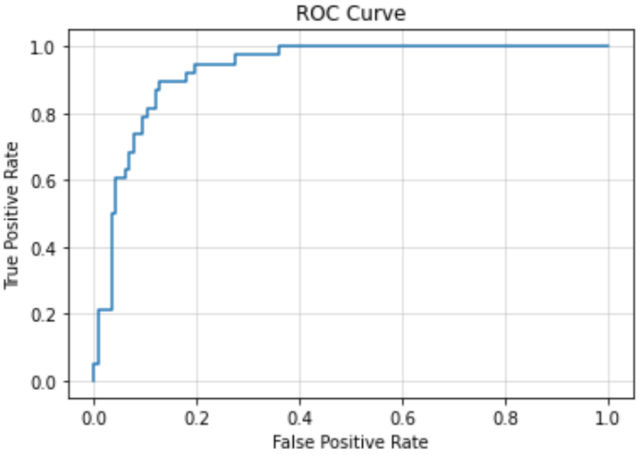
Shown below are the statistics for the model selected from both best subset and forward selection.



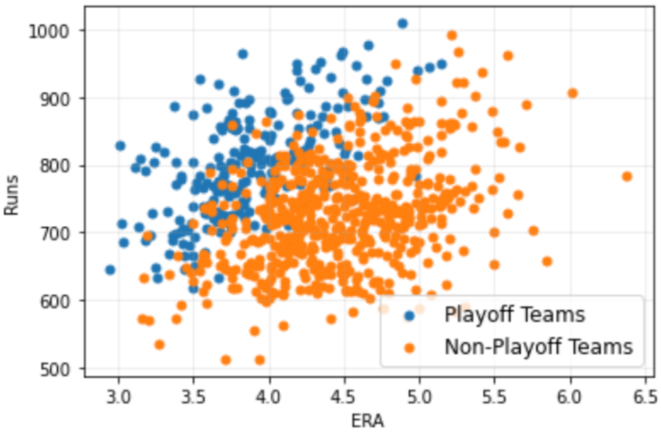
```
OLS Regression Results

    Dep. Variable:               W      R-squared (uncentered):      0.997
           Model:              OLS  Adj. R-squared (uncentered):     0.997
          Method:    Least Squares                 F-statistic:  5.645e+04
            Date: Thu, 28 Apr 2022           Prob (F-statistic):      0.00
            Time:         22:58:04               Log-Likelihood:    -1776.4
No. Observations:              619                          AIC:      3561.
    Df Residuals:              615                          BIC:      3579.
        Df Model:                4
 Covariance Type:        nonrobust

         coef  std err        t  P>|t|  [0.025  0.975]
  R    0.0942    0.002   46.110  0.000   0.090   0.098
 SV    0.6674    0.025   27.220  0.000   0.619   0.716
ERA   -7.7963    0.267  -29.200  0.000  -8.321  -7.272
SOA    0.0145    0.001   16.092  0.000   0.013   0.016

      Omnibus: 4.336    Durbin-Watson:     2.110
Prob(Omnibus): 0.114  Jarque-Bera (JB):    5.224
         Skew: 0.045          Prob(JB):   0.0734
     Kurtosis: 3.441         Cond. No. 2.14e+03
```

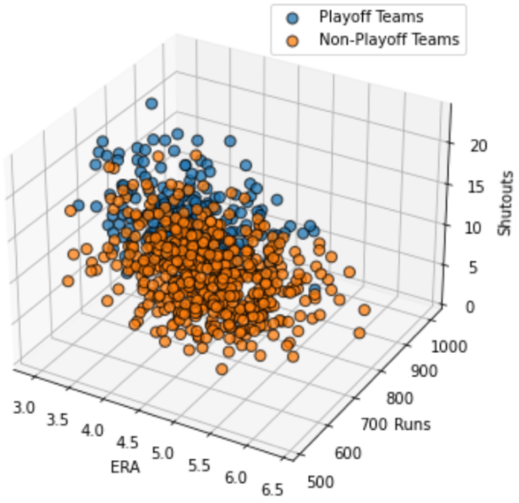
Now I had found the subset of predictors to produce the best model predicting if a team makes the playoffs or not. I used these four variables to run linear discriminant analysis and quadratic discriminant analysis with train/test splits. I was expecting the results to be similar, and they were. The test scores for both analyses were .865. The area under the ROC curves were also similar, as the LDA (left) posted a .934 area and the QDA (right) posted a .931 area.





Lastly, I wanted to visualize the subset data with a few plots. The blue datapoints are playoff teams, and the orange datapoints are non-playoff teams.





## CONCLUSION

I was very pleased with the results on my analysis. I found that runs scored, earned runs average, saves, and strikeouts are the most important variables when trying to predict if a team will make the playoffs. Runs and earned run average do not come as a surprise, as these are the two most important factors that decide games. I was a little surprised to see saves, as it is kind of an overlooked, afterthought of a statistic in modern baseball. The methods I used had a good amount of success. I think this is due to properly cleaning the data and making sure all the assumptions of the tools I used were met.