March Madness Classification

Saul Lozano, Jackson Garro, Cody Timarong

Table of Contents

What is Our Data?

Our dataset “March Madness Historical DataSet (2002 to 2025)” contains information on all NCAA teams, whether or not they made it to a post-season tournament, March Madness. Alongside team information, it includes various offensive and defensive stats, seed rankings and where teams exited during the tournament.

What we used

Our data has over 100 variables but based on our knowledge we narrowed it down to 7 variables that we observed to have the most significant impact on our Seed Variable. Our Seed Variable was also modified depending on the model used, either into binary or even groups.

- Adjusted Offensive Efficiency
- Adjusted Defensive Efficiency
- RankeFGPct - Rank of Effective Field Goal Percentage
- eFGPct - Effective Field Goal Percentage
- TOPct - Turnover Percentage
- ORPct - Offensive Rebounding Percentage
- FTRate - Free Throw Rate

Goals and Guidelines

Can we predict March Madness seed groupings?

What team performance metrics best predict tournament seeding?

Which classification methods work best for predicting tournament seeding?

Data Wrangling

Data Wrangling

Seed Grouping:

- We grouped teams into two different categories
- Low seed: 1-8
- High seed: 9-16

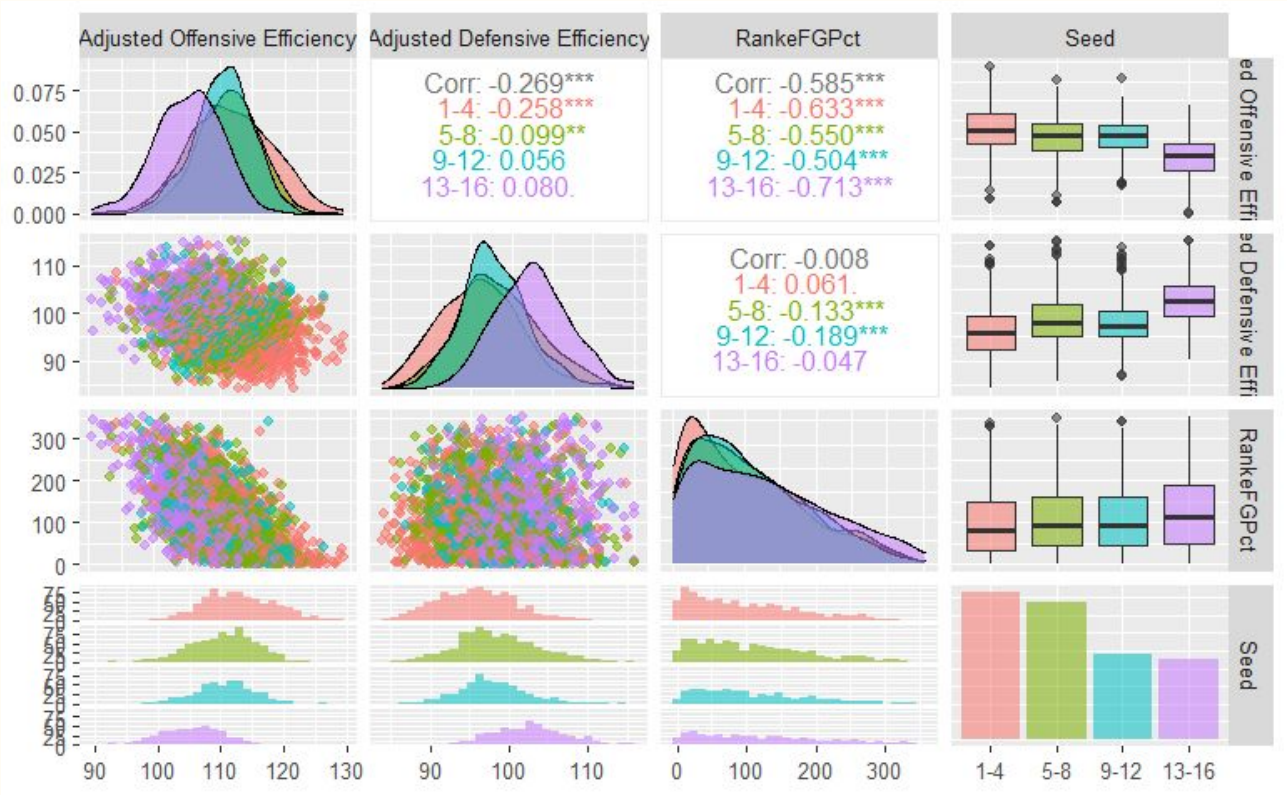
Conference:

- We created a power_5 variable that identifies teams in a power 5 conference
- ACC, BIG10, BIG12, PAC 12, SEC

Filtering:

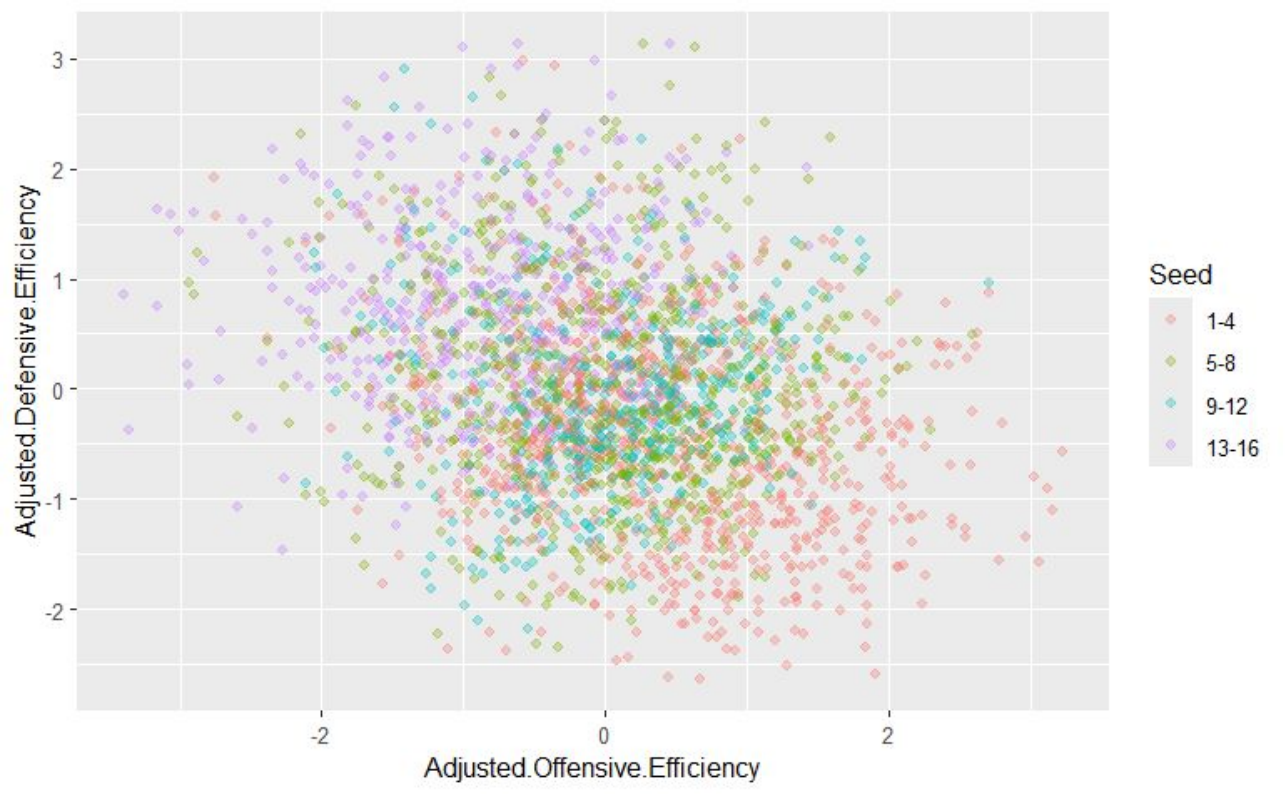
- We removed all teams that didn’t make the tournament “DNF”

Pair Plots - Correlations



K - Nearest Neighbor

Do Groups Exist?



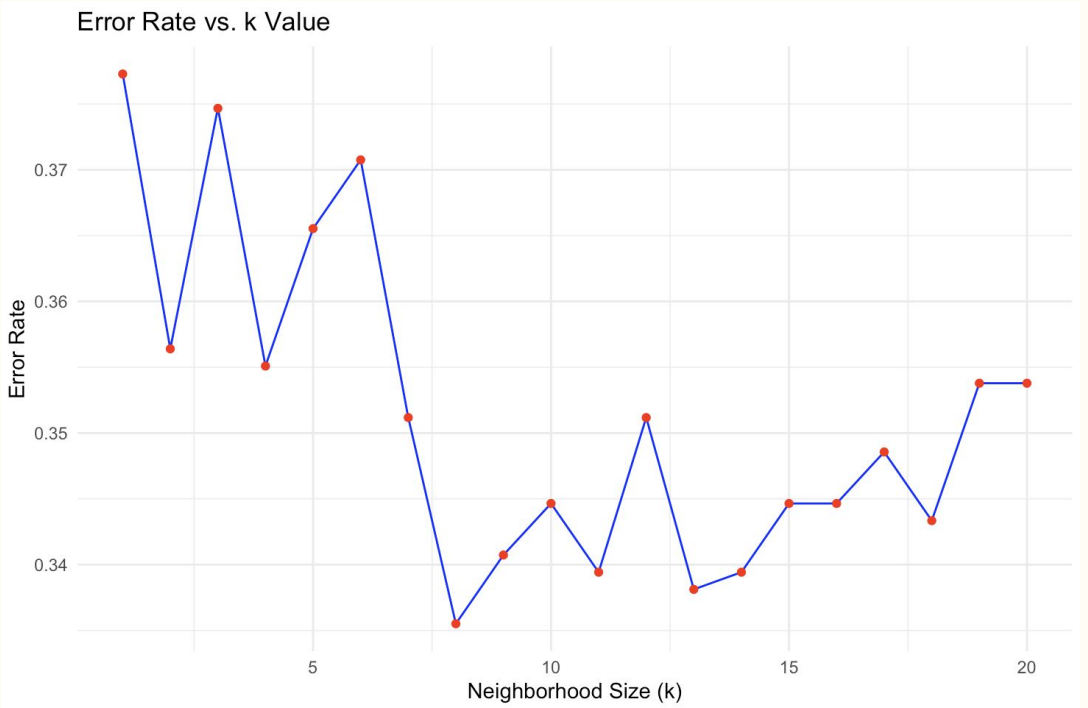
Based two of on most important variables, groups do exist.

The best teams are closer to the bottom right while the worst are close to the top left.

KNN Model

What K value is the most optimal?

The neighborhood size of 8 has the lowest error rate



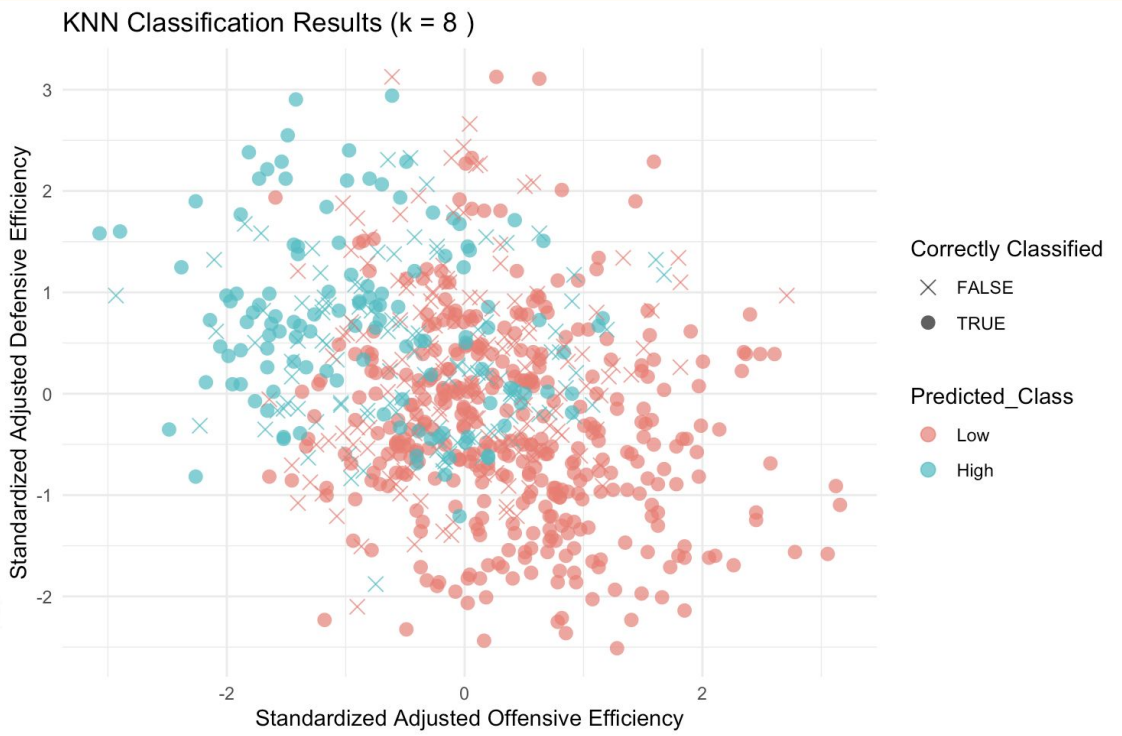
KNN Model

Used Adjusted Offensive and Defensive Efficiency for this model

Final KNN Accuracy: 0.655

Final KNN Error Rate: 0.344

	Low	High
Low	386	97
High	151	132



Trees + Bagging

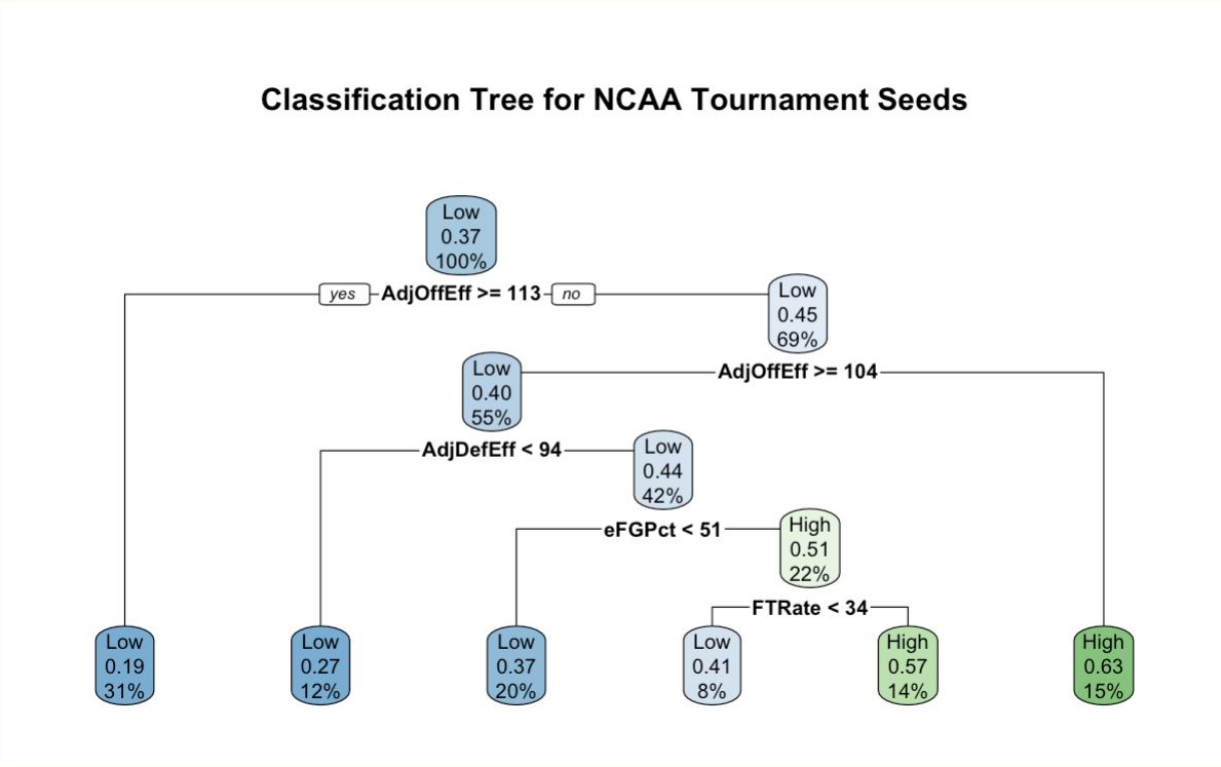
Pruned Classification Tree

Predicted

Actual		Low	High
	Low	396	87
	High	158	125

Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Effective Field Goal Percentage, and Free Throw Rate for our splits

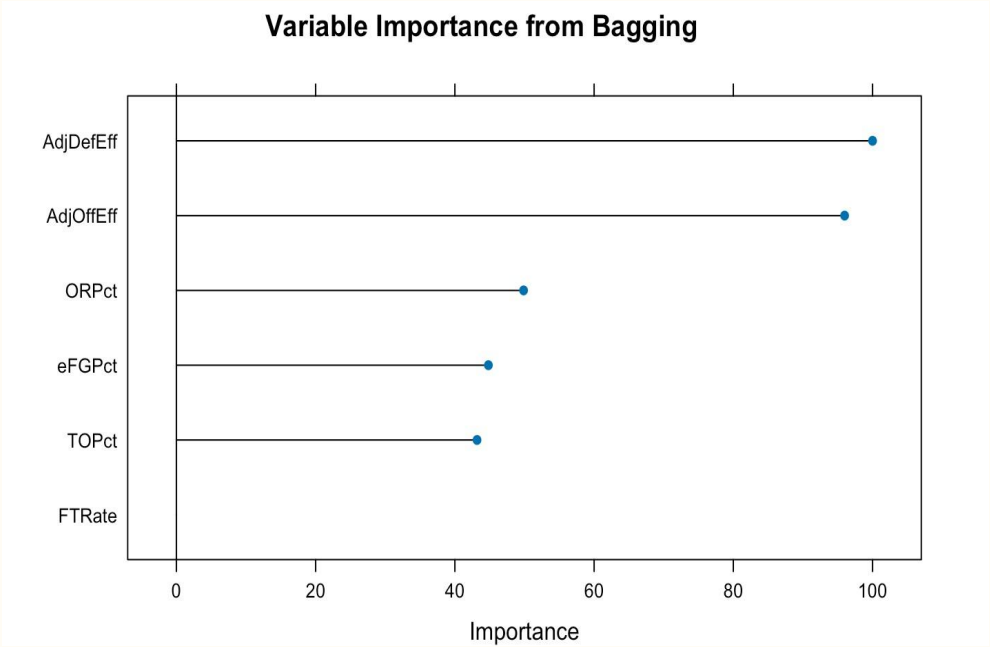
- 68% accurate



Bagging

AdjDefEff and AdjOffEff are influential predictors

FTRate had the least contribution.



Bagging

Accuracy of 65.9%, a moderate performance for predicting.

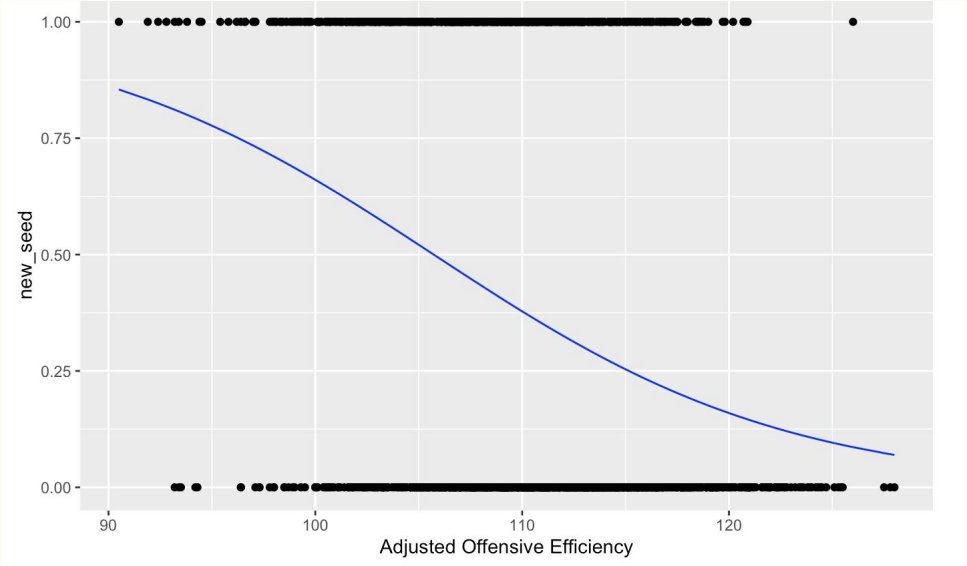
Predicted accurately for low seeds than high seeds.

```
[1] "Confusion Matrix for Caret Bagging:"
      pred_caret_bag
       Low High
  Low  383  100
  High 161  122
Accuracy for Caret Bagging: 0.6592689
```

Logistic Models

Simple Logistic Regression

```
          0   1
  FALSE 426 197
  TRUE   51  91
[1] 0.675817
```



Multiple Logistic Regression

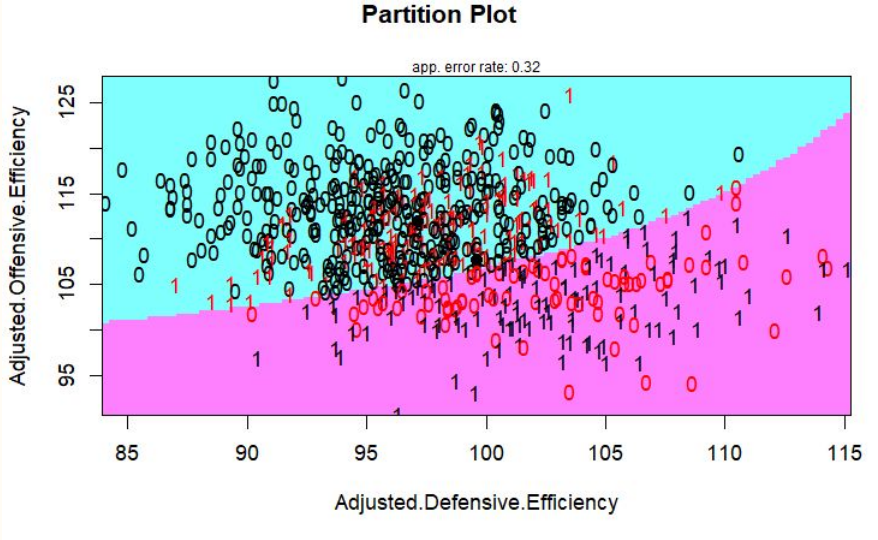
```
          0   1
  FALSE 425 150
  TRUE   80 110
[1] 0.6993464
```

```
Call:
glm(formula = new_seed ~ Adjusted.Offensive.Efficiency + Adjusted.Defensive.Efficiency +
    RankeFGPct + Adjusted.Temo, family = "binomial", data = trainCaret)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 5.350454   2.138276   2.502 0.012341 *
Adjusted.Offensive.Efficiency -0.121194   0.012987  -9.332  < 2e-16 ***
Adjusted.Defensive.Efficiency  0.084542   0.010775   7.846 4.29e-15 ***
RankeFGPct                  -0.002777   0.000820  -3.387 0.000707 ***
Adjusted.Temo               -0.009613   0.017038  -0.564 0.572605
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2346.4  on 1785  degrees of freedom
Residual deviance: 2093.6  on 1781  degrees of freedom
AIC: 2103.6

Number of Fisher Scoring iterations: 3
```

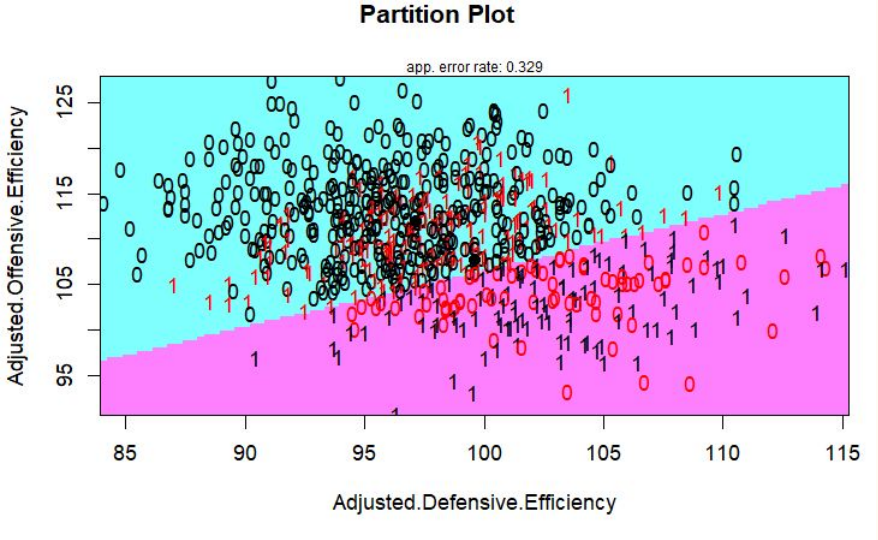
QDA + LDA

Logistic Models

QDA



LDA



QDA + LDA

```
Call:
qda(new_seed ~ ., data = trainCaret)

Prior probabilities of groups:
        0         1
0.6181411 0.3818589

Group means:
  Adjusted.Offensive.Efficiency Adjusted.Defensive.Efficiency RankeFGPct Adjusted.Temo
0                      111.6352                      97.30471   104.5154      66.29710
1                      108.0523                     100.18050   115.8680      66.37551
[1] 0.6941176
```

Assumes equal covariance across classes; calculates a linear combination of predictors to separate teams based on seeding. Coefficients indicate each variable's contribution to class separation."

```
Call:
lda(new_seed ~ ., data = trainCaret)

Prior probabilities of groups:
        0         1
0.6181411 0.3818589

Group means:
  Adjusted.Offensive.Efficiency Adjusted.Defensive.Efficiency RankeFGPct Adjusted.Temo
0                      111.6352                      97.30471   104.5154      66.29710
1                      108.0523                     100.18050   115.8680      66.37551

Coefficients of linear discriminants:
                                LD1
Adjusted.Offensive.Efficiency -0.166411001
Adjusted.Defensive.Efficiency  0.091851787
RankeFGPct                    -0.004393619
Adjusted.Temo                  0.013782769
```

Allows class-specific covariance, enabling more flexible (nonlinear) decision boundaries. Uses group means and priors to classify without linear coefficients."

Comparing Models

- KNN prediction: 65% accuracy
- Bagging prediction: 64.5% accuracy
- Tree prediction: 68% accuracy
- Simple Logistic Model: 67.5% prediction accuracy
- Multiple Logistic Model: 69.9% prediction accuracy

Conclusion

- The decision tree showed that offensive efficiency was the most important factor and had an accuracy of 68%.
- The K-Nearest Neighbors model helped us group similar teams, and the best results came when K was 8
- Bagging improved accuracy and confirmed that offense and defense were the most important stats.

Questions?