Defining Offensive and Defensive Positions in College Basketball to Build Optimal Rosters for Maximum Tournament Success

By: Kushal Shah and Colin Krantz

1. Abstract

While basketball is considered a "positionless" sport, we attempt to create more precise position labels that help teams understand the value and role of each player. We chose to do this for the NCAA rather than the NBA because the amount of research regarding position definitions is limited for the NCAA in comparison to the NBA. We defined the traditional G, F, and C positions using a random forest model that incorporated players' physical attributes and shot selections. We also clustered players separately on offense and defense using a variety of different metrics along with their physical attributes and shot selections. We defined players separately on offense and defense to more accurately group players together. The random forest and clusters allowed us to use a probabilistic approach that outputs the likelihood of each player being labeled in a specific position or cluster grouping. This approach allowed us to also create hybrid positions for players who had similar probability values for different positions or clusters. We believe a combination of our predicted traditional position with offensive and defensive cluster groups provides more insight into the true nature of a player's role on a team. For example, Mike Daum is labelled as F-4-7 which indicates that our model identified him as a primary forward in offensive cluster 4 and defensive cluster 7. With ten years of NCAA data, we were able to create the ideal roster by relating team lineups to tournament success. We looked at the rosters of teams who made it to the Elite 8 or further during March Madness to understand which specific positions commonly appeared together on these teams. We believe that this will provide insights to coaching staffs when targeting high school and transfer recruits.

2. Data and Methodology

2.1 Data Acquiring and Cleaning

To create our clusters, we scraped and cleaned 10 years of college basketball player data from Bart Torvik. We joined this data with sports reference data to accumulate over 50 variables that describe the player and their college career. After applying a minutes restriction (Minute Percentage > 24.99) on the dataset, we had a total 26,250 observations with 50 variables. These 50 variables were a combination of their physical statistics (Height, Weight, etc.), advanced metrics (Usage%, etc.), and shot distributions (Close2Att, Far2Att, etc.). In our data set, each row corresponds to a player during that specific season, so a player could appear 3 times if he played for 3 years in college.

2.2 Offensive and Defensive Clustering

We decided to cluster players based on 3 major aspects. Their physical build, style of play, and shot selection. We did not use any singular performance measures such as box plus minus or offensive rating as the clusters would essentially club players based on their skill allowing good players to be in the same cluster. This would not define new positions or provide insight for a team when recruiting or understanding how to build the most effective March Madness roster. We also created 2 sets of clusters (offensive and defensive). This allows us to have different position definitions based on a player's offensive and defensive capabilities. We believe this is important because the style of play for a player is different on both sides of the ball and these clusters will help define a player's position in a far more effective manner. Below is a table that summarizes the variables we incorporated for our offensive and defensive clusters:

Variable	Cluster(s) Used For	Description
Height	Offense and Defense	The height of a player
Weight	Offense and Defense	The weight of a player
Minute Percentage	Offense and Defense	The percentage of minutes a player was on the court
Experience	Offense and Defense	The number of years the player has already played in college for
USG%	Offense	The percentage of team plays used by a player while he was on the floor.
AST%	Offense	The percentage of teammate field goals a player assisted while he was on the floor.
ORB%	Offense	The percentage of available offensive rebounds a player grabbed while he was on the floor.
TO%	Offense	Turnover percentage is an estimate of turnovers per 100 plays.
Three Pointers Attempted	Offense	Number of 3 pointers attempted
Three Point Percentage	Offense	Percentage of shots made from 3 by a player
Free Throw Percentage	Offense	Percentage of free throws made by a player
Free Throw Rate	Offense	Free throws made per field goal attempted
Close 2 Attempted	Offense	Number of close 2 pointers attempted
Close 2 Percentage	Offense	Percentage of close 2 pointers made
Far 2 Attempted	Offense	Number of far 2 pointers attempted
Far 2 Percentage	Offense	Percentage of far 2 pointers made
Dunks Attempted	Offense	Number of dunks attempted
Dunk Percentage	Offense	Percentage of dunks made
DRB %	Defense	The percentage of available defensive rebounds a player grabbed while he was on the floor.
Personal Foul Rate	Defense	Proportion of team fouls committed by a player adjusted for their minutes.
STL%	Defense	Percentage of opponent possessions that end with a steal by the player while he was on the floor.
BLK%	Defense	Percentage of opponent possessions that end with a block by the player while he was on the floor.
Stops	Defense	Number of defensive stops by a player

We used the mclust package in R to employ a model-based clustering technique. This clustering technique selects the number of clusters and generates probability values for a player to be in each specific cluster. Model-based clustering uses an expectation-maximization (EM) algorithm to fit Gaussian finite mixture models. A distribution of our offensive clusters and defensive clusters can be seen below:

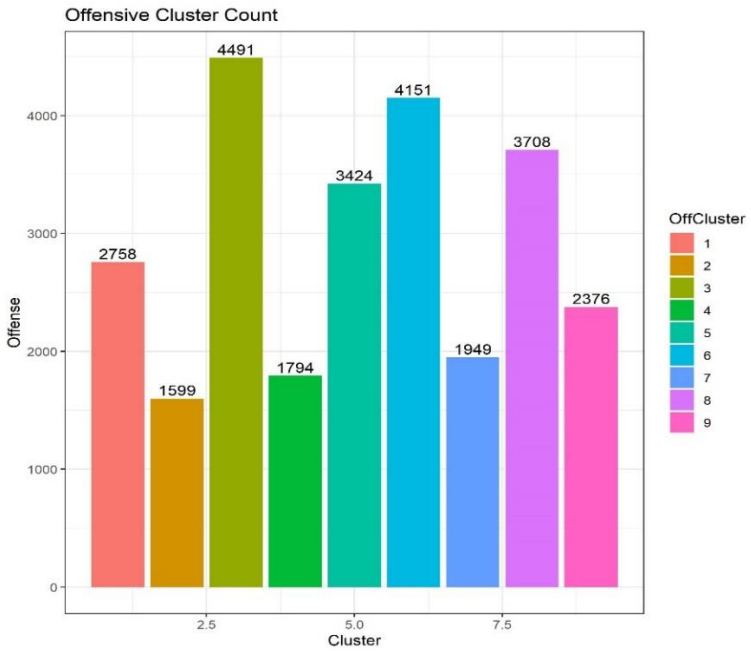

Figure 1: Offensive Cluster Counts

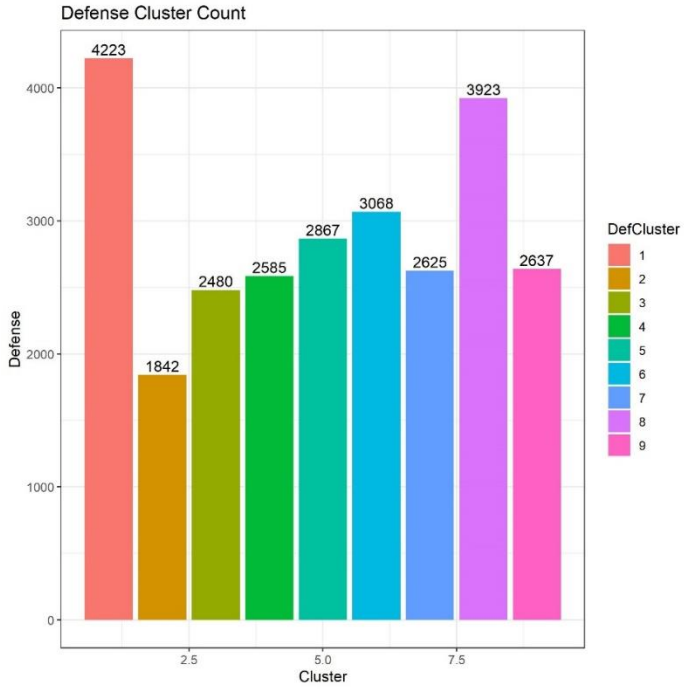

Figure 2: Defensive Cluster Counts

Below is an in depth investigation into our clusters and some notable players within these clusters:

Cluster	Description	Prototype Player
OffenseCluster1	Young players with low minutes and a high TO%	Freshman Elfrid Payton (2012)
OffenseCluster2	Big guys with a high eFG% and a high USG%	Freshman Montrezl Harrel (2013)
OffenseCluster3	Average guards with a low TO% and high number of 3 pointers attempted	Freshman Kyle Anderson (2015)
OffenseCluster4	Players with low usage and minutes and a very high TO%	Senior Tereke Eckwood (2019)
OffenseCluster5	Big guys with great inside scoring but very poor outside scoring	Freshman Alex Len (2012)
OffenseCluster6	Small players with poor shooting and low usage	Freshman Devon Reed (2014)
OffenseCluster7	Experienced players with high usage, great shooting all around, and a low TO%	Junior Kemba Walker (2011)
OffenseCluster8	Small players with high minutes played, high three pointers attempted, and high AST%	Junior Jimmer Fredette (2010)
OffenseCluster9	Three point shooters	Sophomore Joe Harris (2012)
DefenseCluster1	Slightly below average defenders who are guards	Junior Reggie Jackson (2011)
DefenseCluster2	Smaller players with high STL% and above average BLK%	Freshman Marcus Smart (2013)
DefenseCluster3	Small Players who are poor defenders with low DRB% and low BLK%	Freshman Colin Sexton (2018)
DefenseCluster4	Big guys who are good defenders with high DRB% and BLK%	Freshman Anthony Davis (2012)
DefenseCluster5	Mostly guards with high minutes, an above average STL%, and a high personal foul rate	Senior Grayson Allen (2018)
DefenseCluster6	Very average defenders who are typically guards	Senior C.J. McCollum (2013)
DefenseCluster7	Younger players with very low minutes who play average defense	Freshman Tony Snell (2011)
DefenseCluster8	Forwards with high foul rates, high STL% but also low minutes played	Sophomore OG Anunoby (2017)
DefenseCluster9	Experienced and bigger players with a high DRB% and BLK%, spread out across Guards, Forwards, Centers.	Senior Matisse Thybulle (2019)

2.3 Position Predictions using a Random Forest

In addition to our clustering, we used a random forest model to predict the position of a player. The positions we predicted were if a player were to be a Forward, Guard, or a Center. The random forest made use of only their physical attributes and shot selections. We believe the traditional labels are not an accurate representation of players today and hence predicting their position using this model can lead to more accurate representations of their play style. The random forest generated a probability value for each player being a Forward, Guard, or Center.

This also allowed us to label player as hybrid positions where the probability of a player being labelled in 2 positions are close to 0.5. For example, is a player is to have 0.49 chance to be a forward and 0.51 chance to be a guard, he was labelled as Guard-Forward hybrid.

2.4 Final 3-Part-Position Definitions

This eventually led us to our final position definition. The definition is made up of 3 components:

a. Offensive Cluster
b. Defensive Cluster
c. Random Forest Predicted Position

Hence, a 4-7-G player is a player in offensive cluster 4, defensive cluster 7, and is predicted to be a Guard by our model. The hybrid players have their second position added to this and would be represented like this: 4-7-G-F (Guard-Forward Hybrid).

3. Modeling Success

3.1 Logistic Regression Model

After gathering every player's probability values to fall within each possible cluster, we created a cluster value rating for each offensive and defensive cluster for every team on a season by season basis. This value was made by adding the probability values for each player with qualified minutes for a team value (If a team had two players with a 0.6 probability to fall in offensive cluster 1, the team offensive cluster 1 value would be 1.2). We also found the total number of each 3-part-position that was on a team's roster. Eighteen different logistic regressions were created from this data for predicting probability values for a team to move onto each subsequent round in March Madness. The regressions have three different sets of independent variables and six different dependent variables each season. The three sets of independent variables are:

- Team offensive cluster rating for all 9 offensive clusters
- Team defensive cluster rating for all 9 defensive clusters
- Total number of players identified for all 3-part-positions

The six dependent variables are binomial variables representing if a team appears in each subsequent round of the NCAA tournament (Round of 32, Sweet 16, Elite 8, Final 4, Final, Champion), 1 if the team appears in the round in question, 0 if they do not. These binomial variables are only assigned to teams that made the NCAA tournament, so all teams who did not qualify for the tournament are assigned NA values.

The regression in figure 3 is used to predict a team's probability to appear in the Round of 32 of the NCAA tournament based on their offensive cluster ratings

```
Call:
glm(formula = Round32 ~ . - 1 - Elite8 - Sweet16 - Final - Final4 -
    Champion - schoolID, data = Offprobdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8706  -0.3962  -0.1331   0.4309   0.9311

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
OffCluster1prob  0.01079    0.02258   0.478   0.6329
OffCluster2prob  0.11316    0.02474   4.573 5.72e-06 ***
OffCluster3prob  0.06756    0.01549   4.360 1.50e-05 ***
OffCluster4prob -0.03577    0.03370  -1.061   0.2889
OffCluster5prob  0.03951    0.02060   1.918   0.0556 .
OffCluster6prob -0.03432    0.02023  -1.697   0.0902 .
OffCluster7prob  0.19020    0.01558  12.210  < 2e-16 ***
OffCluster8prob  0.04911    0.02114   2.323   0.0205 *
OffCluster9prob  0.05318    0.02123   2.505   0.0125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Sample Regression for Round of 32 Using Offensive Clusters

The regression in figure 4 is used to predict a team’s probability to appear in the Round of 32 of the NCAA tournament based on their defensive cluster ratings

```
Call:
glm(formula = Round32 ~ . - 1 - Elite8 - Sweet16 - Final - Final4 -
    Champion - schoolID, data = Defprobdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8708  -0.4190  -0.1160   0.4386   0.9010

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
DefCluster1prob  0.053183   0.021312   2.495 0.012821 *
DefCluster2prob  0.074652   0.024697   3.023 0.002601 **
DefCluster3prob -0.048362   0.033829  -1.430 0.153296
DefCluster4prob  0.238488   0.025558   9.331  < 2e-16 ***
DefCluster5prob  0.009033   0.026655   0.339 0.734805
DefCluster6prob -0.080607   0.023606  -3.415 0.000677 ***
DefCluster7prob -0.034261   0.033617  -1.019 0.308492
DefCluster8prob  0.028963   0.023334   1.241 0.214966
DefCluster9prob  0.156882   0.027843   5.635 2.59e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: Sample Regression for Round of 32 Using Defensive Clusters

The regression in figure 5 is used to predict a team’s probability to appear in the Round of 32 of the NCAA tournament based on the total number of players on their roster identified in each 3-part-position.

```
Call:
glm(formula = Round32 ~ . - 1 - Elite8 - Sweet16 - Final - Fina
    Champion - schoolID, data = TeamPositionCount)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9801  -0.2991   0.0000   0.2925   0.8984

Coefficients: (102 not defined because of singularities)
                        Estimate Std. Error t value Pr(>|t|)
`finalposition_1-1-C`         NA         NA      NA       NA
`finalposition_1-1-F`   -0.018853   0.151101  -0.125 0.900760
`finalposition_1-1-G`   -0.029120   0.146569  -0.199 0.842601
`finalposition_1-1-G-F`  0.103107   0.520320   0.198 0.843004
`finalposition_1-2-F`   -0.198894   0.137396  -1.448 0.148392
`finalposition_1-2-G`   -0.040863   0.093011  -0.439 0.660619
`finalposition_1-2-G-C`       NA         NA      NA       NA
`finalposition_1-2-G-F`  0.100320   0.352373   0.285 0.776001
`finalposition_1-3-F`    0.114336   0.595550   0.192 0.847838
`finalposition_1-3-G`    0.026102   0.111829   0.233 0.815542
`finalposition_1-3-G-F` -0.681331   0.478374  -1.424 0.155029
```

Figure 5: Sample Regression for Round 32 Using RF Predicted Positions

3.2 Predictions

Based on these regressions we create a team's total strength rating to appear in the round of 32 which would be the sum of the results for each logistic regression. For example, we would use the offensive cluster regression pictured above (figure 3) to predict the teams rating to appear in the round of 32 based on solely their offensive clusters. We would repeat this process for the defensive cluster regression and the 3-part-position regression. Our final strength rating would be the sum value of all three of these regressions. This process was repeated to calculate a strength rating for each team to appear in each round of the NCAA tournament. Below are the equations that should provide clarity of our process:

$$RoundOf32\ Rating = OffClusterRoundof32Strength + DefClusterRoundof32Strength + 3PartPosRoundof32Strength$$

$$Sweet16Rating = OffClusterSweet16Strength + DefClusterRSweet16Strength + 3PartPosSweet16Strength$$

$$Elite8Rating = OffClusterElite8Strength + DefClusterRElite8Strength + 3PartPosElite8Strength$$

$$Final4Rating = OffClusterFinal4Strength + DefClusterFinal4trength + 3PartPosFinal4Strength$$

$$Final\ Rating = OffClusterFinalStrength + DefClusterFinalStrength + 3PartPosFinalStrength$$

$$ChampionRating = OffClusterChampionStrength + DefClusterChampionStrength + 3PartPosChampionStrength$$

We used these ratings to create a bracket prediction for each NCAA tournament from 2010 to 2019. The winner of each round was determined by their strength rating to appear in the next round. Our predicted 2014 bracket is shown below in figure **6**, and all of our yearly bracket predictions can be found at

https://docs.google.com/spreadsheets/d/1UhEyaJzIro4vSKHtLECGTLRMmOfbJDlb4t4M2ZXulww/edit#gid=841879642

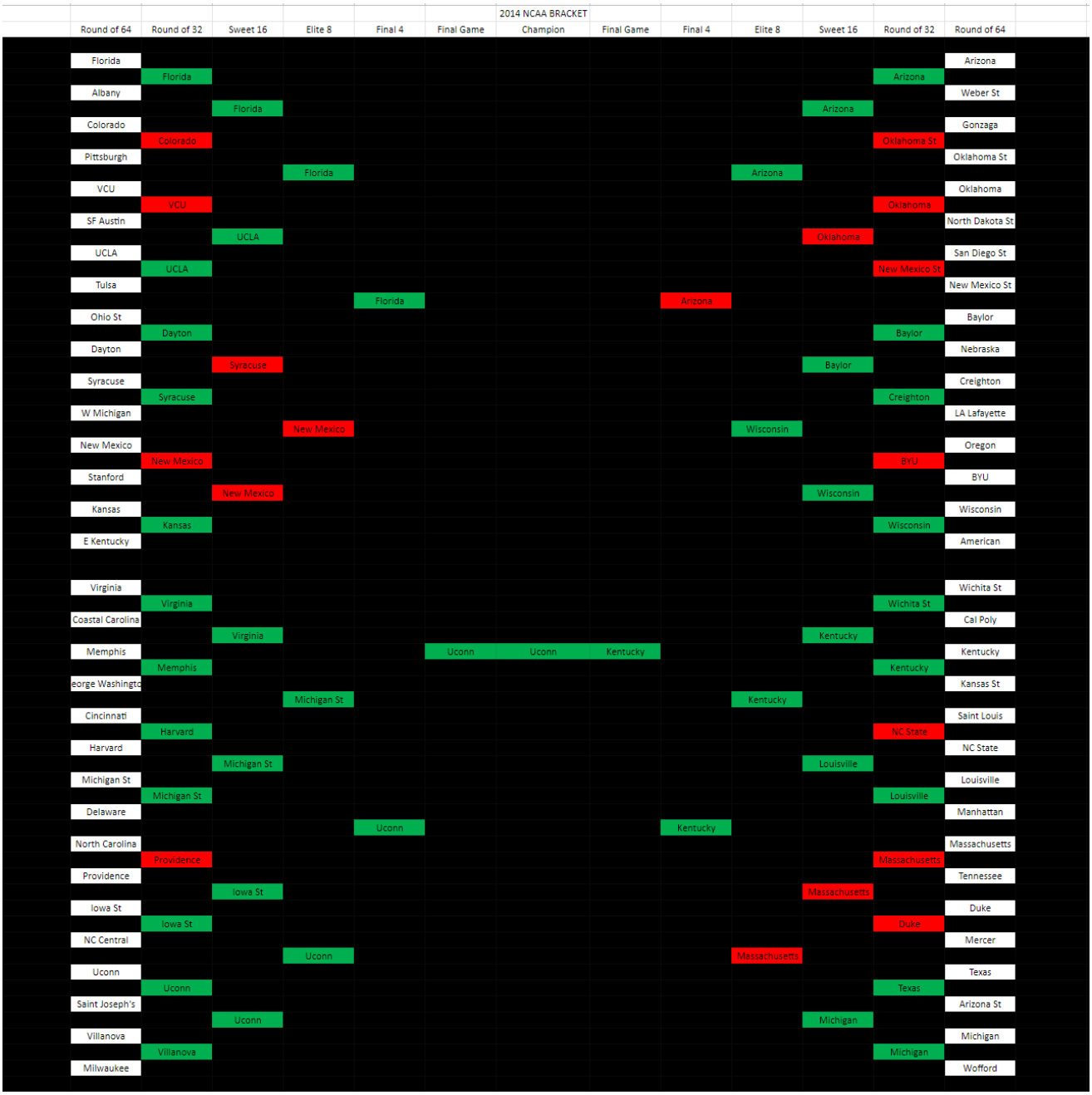

Figure 6: 2014 March Madness Prediction

We also show our round by round success rates in the image below (figure **7**) which should show how successful our model is which in turn validates our methods to create an ideal roster for March Madness.

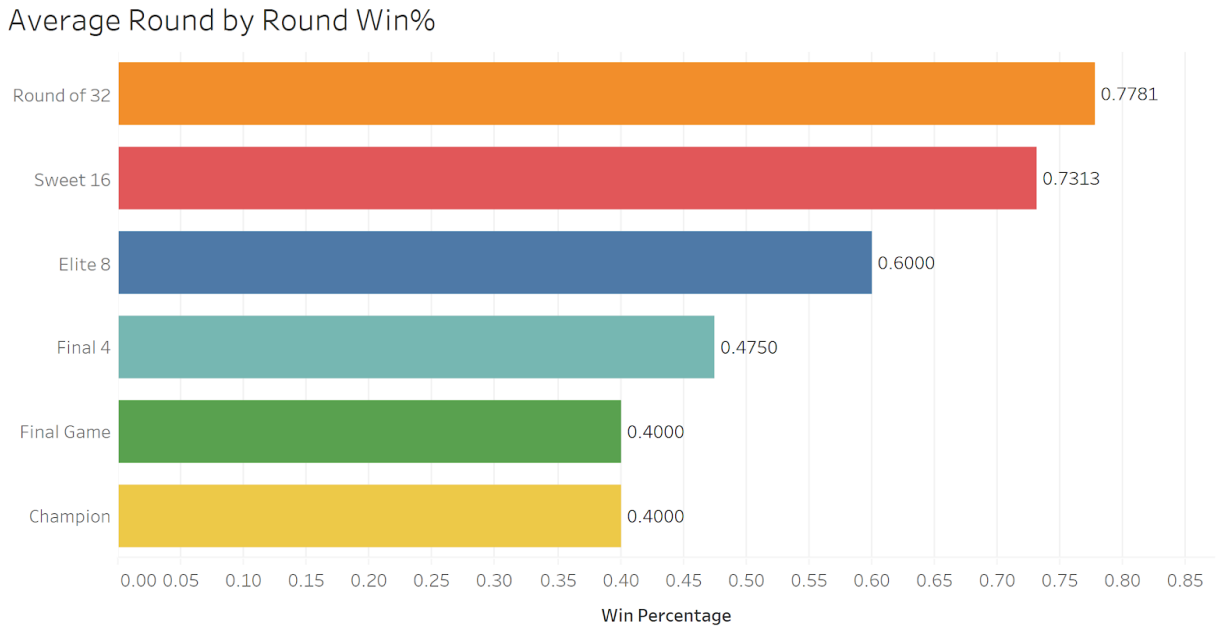

Figure 7: Average Round by Round Success Rates

Based on the percentages of correct picks above, our average bracket would include:

- 24.9/32 Round of 32
- 11.7/16 Sweet 16
- 4.8/8 Elite 8
- 1.9/4 Final 4
- 0.8/2 Final Game
- 0.4/1 Champions
- 44.5/63 Correct Picks

The maximum number of teams we predicted correctly to appear in each round is as follows:

- Round of 32: 29 (2013, 2018)
- Sweet 16: 13 (2013, 2019)
- Elite 8: 6 (2014, 2015)
- Final 4: 3 (2012, 2014, 2015)
- Final Game: 2 (2014)
- Champion: 1 (2010, 2011, 2012, 2014)
- Correct Picks: 50/63 (2013)

This model either correctly picked or nearly correctly picked a large amount of noteworthy upsets and Cinderella runs over the years, some of the most impressive including:

- (16) UMBC defeating (1) Virginia (2018)
- (15) Florida Gulf Coast defeating (2) Georgetown and (7) San Diego St (2013)
- (7) UConn defeating (8) Kentucky in the Championship Game (2014)

4. Discussion and Conclusion

The implications of this project could result in millions of dollars being paid to a conference for a team making a large run throughout the tournament. If a team bases their recruiting on finding players who would conform to the clusters our model predicts as significant and positive towards winning, they could largely increase their probability to go on a tournament run. Each tournament win for a team results in the addition of one "unit" to be given to their conference. For an example of how much a tournament run can be worth, in 2011, VCU's Final 4 run from the play-in game earned the Colonial Athletic Association $9.12 million (USA Today).

While the model's predictions of later rounds is not poor, it can be clearly seen that the model has worse results in the later rounds as compared to the earlier rounds. We believe this would be due to the smaller sample size of teams who have appeared in the later rounds, for example, we have observed eight-times more teams in the Round of 32 as compared to the Final 4. Considering the data has this discrepancy, we would expect the model to be better suited for predicting early round success. If we used more past data, it would be expected that the model's win percentage would improve, as a larger sample size would increase the strength of the model.

Based on our results, we believe that if teams were to decide their high school and transfer recruiting targets based on their conformity to our highly successful clusters, they would drastically increase their chances to go on a deep run in the NCAA Tournament.

Bibliography

https://www.usatoday.com/story/sports/ncaab/2019/03/26/whats-an-ncaa-tournament-unit-worth-millions-heres-how/39256149/

https://barttorvik.com/

https://www.sports-reference.com/cbb/

https://cran.r-project.org/web/packages/mclust/mclust.pdf

http://www.sloansportsconference.com/wp-content/uploads/2020/02/Kalman_NBA_Line_up_Analysis.pdf

https://github.com/lwinter819/Capstone/blob/master/Code/Data_Collection/Bart%20Torvik%20Data.ipynb