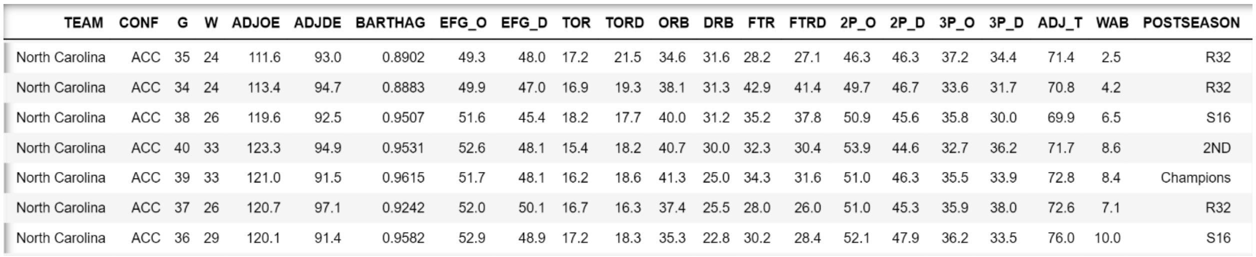
Simeon Kolev
January 15th, 2023

NCAA March Madness Predictive Analysis

In this analysis, Andrew Sundberg's College Basketball Dataset (2014-2019) was used to explore patterns in the stats successfully backing college teams and to predict their expected performance in the NCAA March Madness tournament. This analysis was done in python and incorporates a couple unsupervised and supervised machine learning techniques, including PCA and K-means clustering, and multiclass logistic regression and classification trees. The result of the experiment showed a strong ability to utilize seasonal performance statistics to correctly classify teams into 9 different tournament categories. The models showed ≈ (89% – 92%) predictive accuracy.

An introductory explanatory analysis was showed that two of the dataset's variables, POSTSEASON and SEED, had a significant proportion of N/A inputs. Since these variables are intuitively highly correlated with a march madness performance, I decided encode the N/As in POSTSEASON to represent teams that 'Did Not Qualify" for the tournament, and corresponding those same teams that had N/A in SEED would be assigned a value 18. The dataset was then reordered by ascending team name and year.

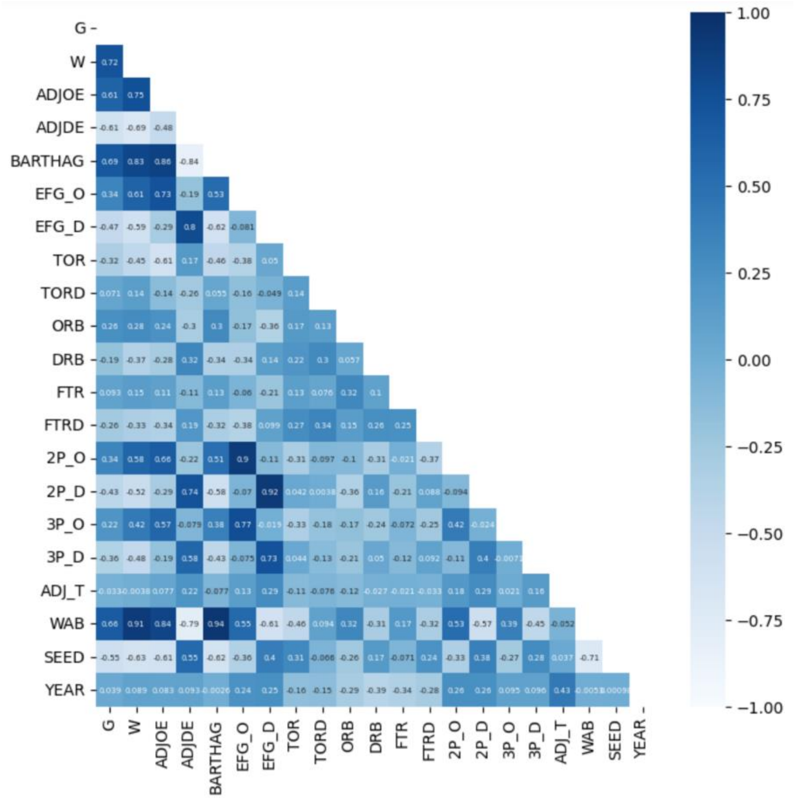
Preview of UNC's stats/data and POSTSEASON performance.



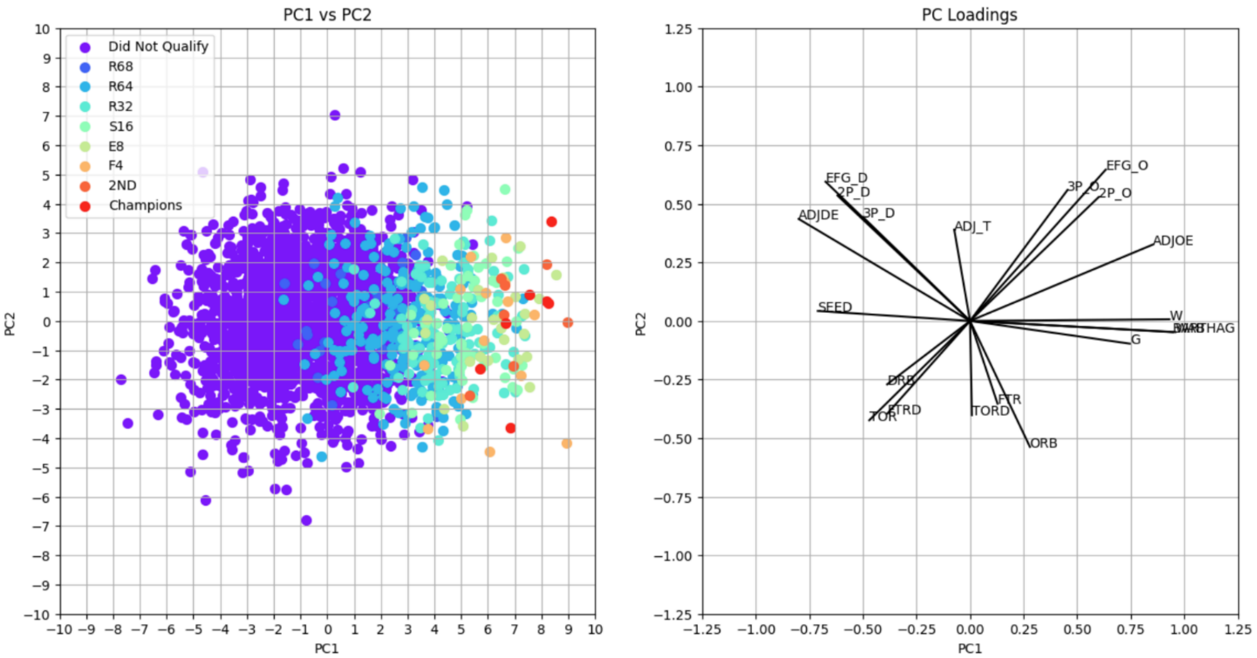
TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	TORD	ORB	DRB	FTR	FTRD	2P_O	2P_D	3P_O	3P_D	ADJ_T	WAB	POSTSEASON
North Carolina	ACC	35	24	111.6	93.0	0.8902	49.3	48.0	17.2	21.5	34.6	31.6	28.2	27.1	46.3	46.3	37.2	34.4	71.4	2.5	R32
North Carolina	ACC	34	24	113.4	94.7	0.8883	49.9	47.0	16.9	19.3	38.1	31.3	42.9	41.4	49.7	46.7	33.6	31.7	70.8	4.2	R32
North Carolina	ACC	38	26	119.6	92.5	0.9507	51.6	45.4	18.2	17.7	40.0	31.2	35.2	37.8	50.9	45.6	35.8	30.0	69.9	6.5	S16
North Carolina	ACC	40	33	123.3	94.9	0.9531	52.6	48.1	15.4	18.2	40.7	30.0	32.3	30.4	53.9	44.6	32.7	36.2	71.7	8.6	2ND
North Carolina	ACC	39	33	121.0	91.5	0.9615	51.7	48.1	16.2	18.6	41.3	25.0	34.3	31.6	51.0	46.3	35.5	33.9	72.8	8.4	Champions
North Carolina	ACC	37	26	120.7	97.1	0.9242	52.0	50.1	16.7	16.3	37.4	25.5	28.0	26.0	51.0	45.3	35.9	38.0	72.6	7.1	R32
North Carolina	ACC	36	29	120.1	91.4	0.9582	52.9	48.9	17.2	18.3	35.3	22.8	30.2	28.4	52.1	47.9	36.2	33.5	76.0	10.0	S16

A correlation matrix was also used to explore some noticeable trends in the data before performing the unsupervised ML techniques. The correlation heatmap showed some striking high correlations between dependent predictors or predictors that had common derivations, such as G (# games), W (# wins), and WAB (# wins above bubble), as well as 2P_O (2-point shooting %) and EFG_O (FG shooting %), as well as 2P_D (2-point shooting % allowed) and EFG_D (FG shooting % allowed).

However, relevant relationships exist between the 2P_O, 2P_D, ADJOE (offensive efficiency), ADJDE (defensive efficiency), BARTHAG, and SEED predictors.



A more detailed analysis of the correlation matrix will show that a relevant proportion of variables shared above 50% correlation with a select group of intuitively important variables such as ADJOE, ADJDE, etc. This gave reason to further explore some relationships in the data before making classifications. An eigenvector decomposition of the covariance matrix was performed to significantly reduce the dimensionality to a 3D space within R20 using the first 3 PCs. Prior to this analysis the variables were rescaled, and TEAM, CONF, YEAR, and POSTSEAON were removed for simplicity. The python sklearn library was used to perform PCA and results showed that a combined ≈ 65% of the total variance was explained using just 3 PC's, with ≈ 40% PVE from PC1, ≈ 17% PVE from PC2 and ≈ 8% PVE from PC3. The POSTSEAON performances were then merged and plotted with the PC scores of the observations and the loadings of the first two PC's. The PC plots showed an observable correlation between successful tournament performances, PC1, and the most significant predictors composing PC1.

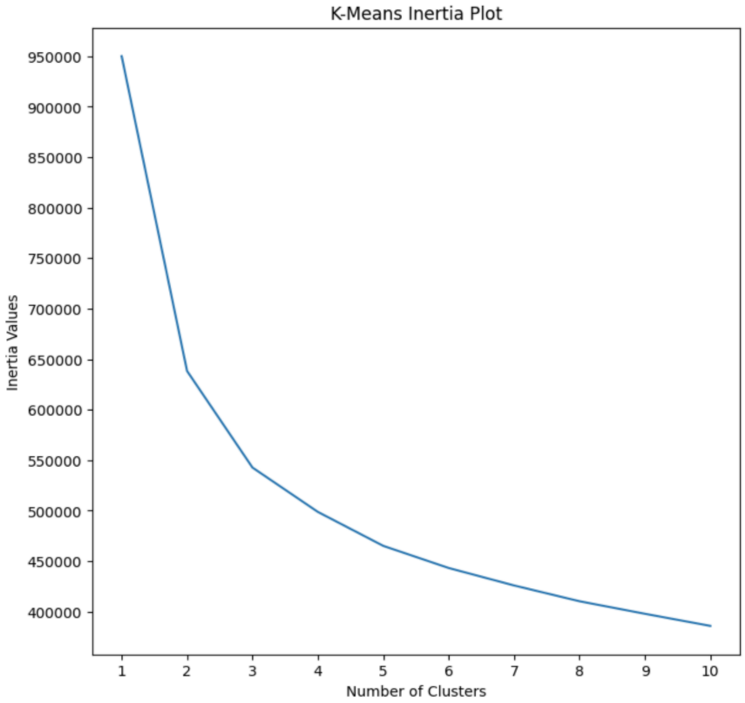


Three performances-based groupings can be visibly made into DNQ, (R68 - R32), and (S16 – Champions).

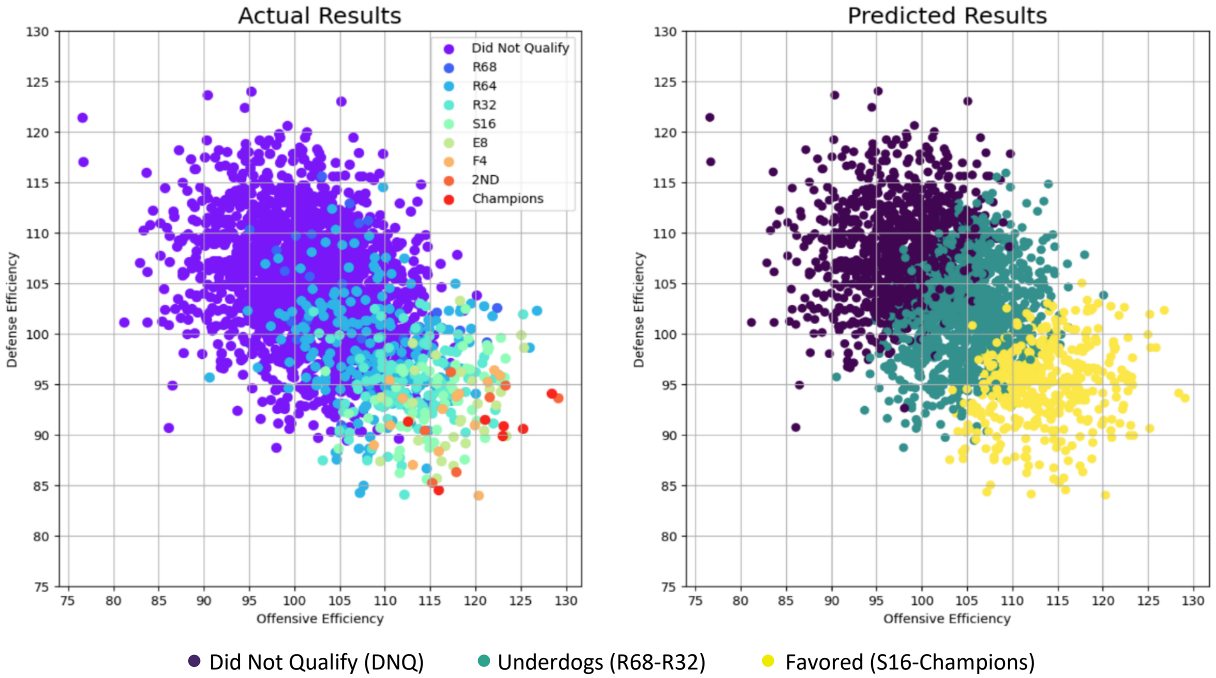
The G, W, WAB, ADJOE, ADJDE, and SEED predictors had strong PC1 loadings, with the ADJOE and ADJDE showing a relatively equal loading in PC1 and PC2. Surprisingly, the color encoded observations are distributed in a way to show very strong progressive correlation in the direction of PC1 and a very subtle interactive correlation between PC1*PC2 but only for the upper rightmost observations (reds). This correlation aligns well with the PC loadings of ADJOE and inversely with ADJDE (this is only the case desirable ADJDE scores are low, and desirable ADJOE scores are high). These two variables inspired further analysis and attempts to observe the relative success of unsupervised classification methods using K-Means clustering techniques, plotted on a ADJOE x ADJDE scale.

Note: The BARTHAG variable was ignored in this analysis. Although it showed strong values in the exploratory analysis, PCA, and Random Forest, it is a "power rating" assigned to each team based on that team's ability to beat an "average D1" team. Since this is not very explicitly defined, I assumed it to be heavily dependent on other predictors, so its relevance was disregarded.

Prior to implementing the K-Means clustering techniques, the data was reverted to its original scaling to ensure the distribution of observations kept its original shape and preserve the Euclidean distance between the observations to assist the clustering algorithm. An initial loop was run to observe the clustering performance at different cluster counts. A K-Means Inertia Plot was made to display the trend between the summed distance between the data points and their assigned cluster centroid as the cluster count increased. The plot was not very helpful in determining the optimal number of clusters used as there were no obvious kinks in the inertia line.



Since slight kinks can be observed at the cluster levels 2 and 3, I decided to test out the visible performance of K-Means clustering using 3 clusters since as mentioned in the analysis of the PCA scatterplot, there appears to be a potential clustering of teams into Did Not Qualify (DNQ), Underdogs (R68-R32), and Favored (S16-Champions). The scatterplot did not provide convincing results to further pursue clustering classification, but with general ability, it reaffirmed the trends observed in PCA.

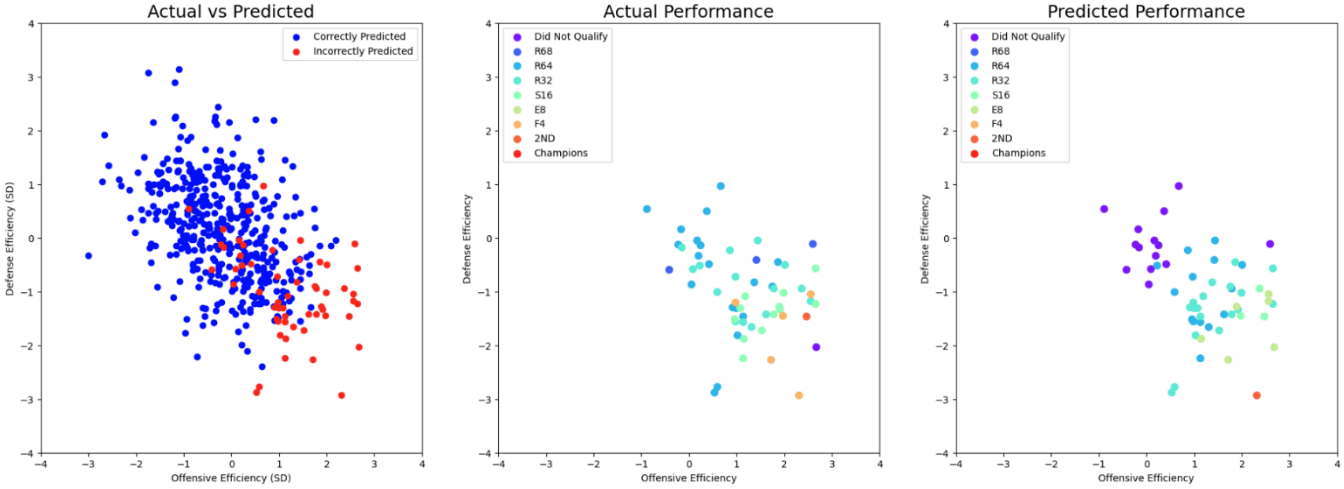


Since K-Means initially randomly assigns each observation to 1 of 3 classes, to be used as a classification technique, it relies heavily on the true distributions to contain an equal number of observations in each class. Since the natural structure of this tournament is that there are many more teams that DNQ than Qualify, it is difficult to implement standard K-Means packages. I also could not figure out how to change the algorithm and manually set the max of observations it assigns to R68-Champions.

A few transformations were made in preparation for the supervised learning portion of the analysis. First, the character encoded variables CONF, and YEAR were transformed to indicator variables, creating many new variables, one for each conference, and one for each year. Then the remaining non indicator numerical variables were standardized using similar methods in PCA. Finally, TEAM variable was then completely removed for this remaining analysis. Since no likely logical reason exists for a team's name or reputation associated with its name should have significant impact on tournament performance.

The data was then split 5 ways, using 20% as the test data and then performing K-Fold cross validation using 5 folds on the training data (remaining 80%). The multiclass logistic regression was fit to the cross validated training data and then used to make classifications predictions on the test data.

Multiclass Logistic Regression showed Prediction Accuracy of ≈ 89% with Standard Deviation of ≈ 0.6%.

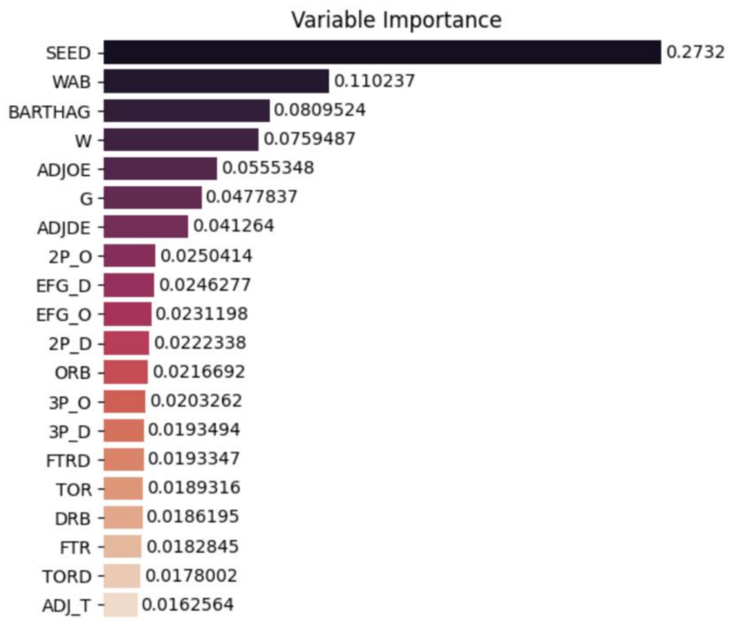


From the Actual vs Predicted scatterplot, as worrying misclassification rate increase can be observed. Most of the misclassifications occurred at the higher ranged of the rescaled ADJOE and lower ranges of the rescaled ADJDE, indicating that misclassification of tournament performance is more likely to be observed in generally "better" teams. However, a closer look regarding the actual degree of misclassification shows that these misclassified observations were 1. Outliers or 2. Barely Misclassified. Outliers such as teams that experienced tournament upsets resulting in early eliminations or underdogs winning big games they have a statistically disadvantage in are not a worry to the utility of this analysis as irreducible error is inevitable. The remaining misclassification, however, show that 11% error may not be as Boolean as it may seem. A closer comparison of each observations Actual Performance vs Predicted Performance shows that teams were typically misclassified by an average of ≈ ± 1.5 classes.

Ex. Predicted R32 vs Actual S16, or Predicted F4 vs Actual 2ND

The final portion of this analysis included a random forest using classification trees was constructed using the same training and test data split used in the preparation for the multiclass logistic regression. Random Forests yielded slightly better and seemingly more consistent results than the multiclass logistic regression, with a Test Prediction Accuracy of ≈ 92%.

A graph of the relative predictor imporance showed that SEED was overwhelmingly the more imporant stat when projecting tournament preformance, and as suspected ADJOE and ADJDE showed the second most importance after the obvious G, W, and WAB since they will naturally be correlated with tournament performance as #G and #W increases as a team advances in the tournament.



Possible improvements for a second run through the data would be to further investigate the BARTHAG and WAB predictors as it is porbable they artificially improved the predictive ability of the model and can likely be removed with confidence.

Other improvments could include disccounting the G and W stats that added due to a teams progression in the tournament. However, since not all teams across all conferences have the same number of scheduled season games, it might be very tedious to manually add/alter the data.

The Classification Tree diagram was included as it was extremely difficult to read, extremely large, and non-aesthetic.

In conclusion, the various machine learning algorithms used showed a relatively strong predictive power in terms of projecting a team's march madness performance using those years prior seasonal stats/data. There can be various improvements made to these models as the packages I used have many more features than I am familiar with or comfortable tweaking. There are also many other techniques that could have been used such as K-Nearest Neighbors, Random Forests with bagging or boosting, or SVM, or implement other regression-based techniques such as Lasso/Ridge Regression, Natural/Smoothing Splines, or Linear/Polynomial Regressions when exploring other variables such as BARTHAG.