March Madness ML

Richard Hardis (077, [rhardis3@gatech.edu](mailto:rhardis3@gatech.edu)),

Arjun Goyal (416, agoyal305@gatech.edu),

Fred Sackfield (772, fsackfield3@gatech.edu),

Anthony Temeliescu (550, anthony.t@gatech.edu)

4/20/2020

# 

# **Abstract**

The NCAA Men’s Basketball Tournament, also known as March Madness, is an annual intercollegiate competition watched by millions across America. Sixty-eight teams are pitted against each other in a ranked, single elimination tournament known for thrilling upsets. We attempted to gain insights on the important predictors of team success in the tournament using regular season team-level statistics for the past five years of tournament results. Six different supervised learning models were tested using cross validation to predict the number of tournament wins by team. The LASSO regression model showed the best performance in win prediction. The model achieved a 52% game prediction accuracy and a mean squared error of 0.85. The LASSO regression model showed that the two most important factors affecting the number of wins are Adjusted Spread and BARTHAG - two widely used statistical measures of regular season performance. We applied our final model to try and predict the outcome of the 2019 March Madness tournament, and it accurately predicted the number of wins for 77% of the teams, including 1 of the 2 correct teams in the championship game.

# **Report**

## **Introduction**

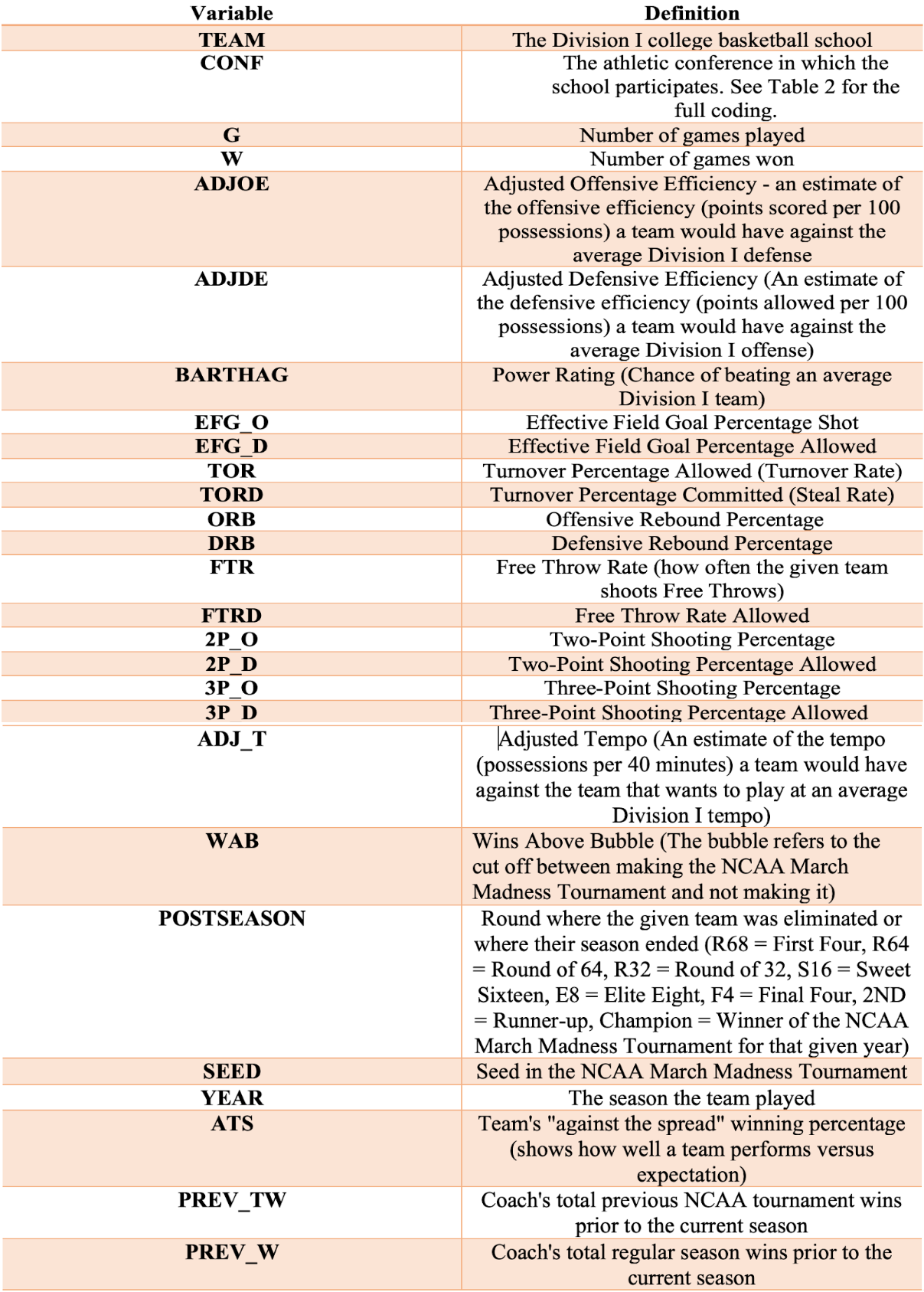
Every year, millions of Americans turn on their televisions and tune into the NCAA men’s basketball tournament, commonly known as March Madness. This tournament pits 68 of the best men’s teams from every conference across Division I against each other in a single elimination style bracket. This tournament is the culmination of a full season’s worth of training and games and the athletic, and financial, stakes could not be higher for every program. Not only does March Madness generate a tremendous media buzz, it also is a topic in which statistical analysis on every level from amateur betters to individuals with Ph.D.’s attempt to predict the outcome of the games. Our group is proposing a novel approach to take on this same task. Using data gathered from various data sources, our project goal is to predict the number of tournament wins for teams participating in the NCAA tournament.

## **Data Collection**

The main portion of our data is taken from the Kaggle dataset called College Basketball Dataset by Andrew Sundberg[1]. It contains data on 24 different variables measuring relevant team performance metrics such as adjusted tempo, tournament seed, and defensive rebound percentage. In total, the dataset contains 1757 rows and includes 5 seasons of data from 2015 to 2019. A driving motivation of the project is to work with novel or unconventional data points to explore alternate angles of predicting NCAA tournament performance. To this end, we treated the kaggle dataset as our “base” set, which we then looked to supplement with additional features.

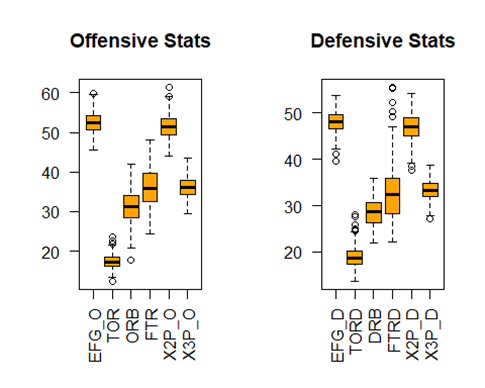
For these additional features, we came up with ideas for quantitative data points that we believed could be useful in predicting post-season performance. One such feature that we added is regular season performance against the spread (ATS). We believe this data point is not widely used in other NCAA tournament modeling efforts, and it provides an interesting perspective on a team’s regular season performance based on market expectations. Another feature that we added was the coach's past performance, both in regular season wins and tournament wins. We believe that a coach’s pedigree and past tournament performance could be a significant factor in tournament performance.

We wrote Python web scraping programs to collect ATS data and coaching data from Covers.com[2] and KenPom.com[3], respectively. These scripts imported the new data into a SQL database, where the data was then cleaned and integrated with our base set from Kaggle. The resulting “full” dataset was then exported from SQL into a csv file, which will be used for the rest of the analysis below.

****

**Table 1: Data Dictionary**

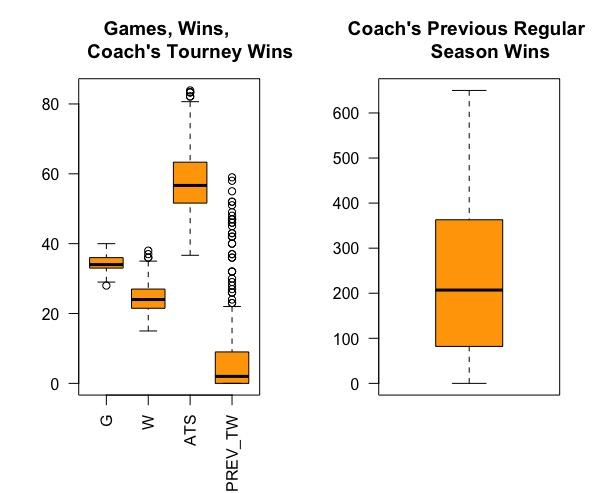
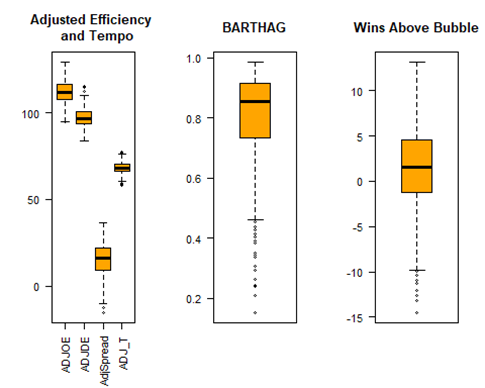
## **Data Exploration**



**Figure 1:** Boxplots for Offensive and Defensive Percentage Metrics

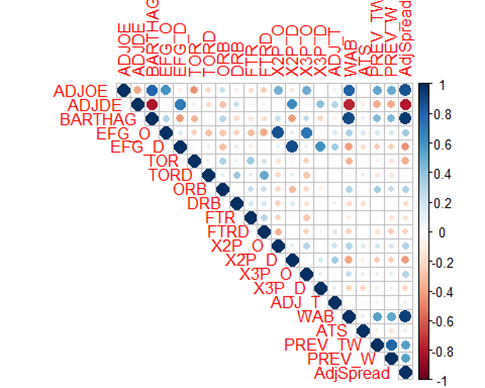
The offensive percentage statistics (Fig. 1) were analyzed separately from the defensive percentage statistics to view outliers and trends. We see one team that shot at a higher field goal percentage (EFG\_O) of 60% compared to other teams. There are a few teams that had a higher turnover rate (TOR) closer to 25% compared to themedian. One team had a significantly lower offensive rebound percentage (ORB), under 20%, compared to the median near 30%. The median two-point shooting percentage (X2P\_0) is near 50% while two teams shot closer to 60%. We see that on defense, there were a few teams that caused more turnovers (TORD) compared to the median. While the median free throw rate allowed (FTRD) was near 30%, 6 teams allowed much higher rates near 50-55%. Two teams were good defenders against the two-point shooting of their opponents (X2P\_D), only allowing under 40% compared to the median closer to 50%.

To qualify for the tournament, most teams have a high adjusted offensive efficiency (ADJOE) that is higher than their adjusted defensive efficiency (ADJDE), indicating they should be able to score on division I teams effectively while preventing their opponents from scoring on them. This is evidenced by the plot for the AdjSpread, the difference between a team’s ADJOE and ADJDE, where teams with a negative spread had a worse defensive efficiency than their offense.The boxplot for BARTHAG indicates that the median chance of defeating the average division I opponent is roughly 90%. However, there are many outliers that are below 50% given that the tournament is open to 64 teams, many of which are from small basketball programs and less-competitive conferences.



**Figure 2:** Boxplots for Quantitative Metrics

Most teams that qualify for the tournament play and win roughly the same amount of games during the season. Some teams achieve a higher ATS score than the median since the tournament frequently sees lower-seeded teams advance against the odds. We see that the median number a coach’s previous tournament wins (PREV\_TW) is near 0, however there are a few teams who possess coaches with 20-60 games of winning experience in the tournament. A coach’s previous regular season wins sits at a median of 200 with no outliers indicated.



**Figure 3: Correlation Matrix of Quantitative Features**

From the correlation plot of the quantitative variables (Fig. 3), we see that a team’s adjusted offensive efficiency (ADJOE) has a positive correlation with some of the team’s offensive percentage stats such as EFG\_O, X2P\_O, X3P\_O. There is a strong positive correlation with the coach’s previous tournament and regular season wins, as well as the BARTHAG rating. Teams with a high ADJOE have a strongly correlated WAB as they are more likely to make the tournament. Expectedly, the adjusted defensive efficiency (ADJDE) is negatively correlated with these explanatory variables given that a lower ADJDE is desired. EFG\_O is strongly positively correlated with the team’s two-point and three-point shooting percentage (X2P\_O and X3P\_O), and conversely the team’s EFG\_D is strongly positively correlated with the team’s ability to defend the two-point and three-point shooting of their opponent (X2P\_D, X3P\_D). Many of the explanatory variables are only weakly correlated with each other. We do see a moderate positive correlation between the WAB and the coach’s previous tournament and regular season wins, perhaps indicating the impact of a coach’s previous experience on a team’s performance.

## **Clustering**

In the spirit of better understanding which team qualities correspond to higher chances of success, we can use clustering methods to group teams together based on similar metrics. We use the most common clustering technique, which is K-Means clustering. K-Means clustering starts with a supplied parameter k to specify the number of clusters we want to group the data into. A small set of data points are chosen randomly to serve as the “initial” cluster centers. The k-means algorithm then iterates through the dataset, calculates the Euclidean distance between each point and the k cluster centers, assigns the new point to the closest cluster, and then recalculates the cluster centers after each data point is added.

One important consideration with clustering is to avoid the “curse of dimensionality”. In other words, we should be selective in what features we use for clustering because distance-based algorithms suffer when the number of features increases. So we have a few options:

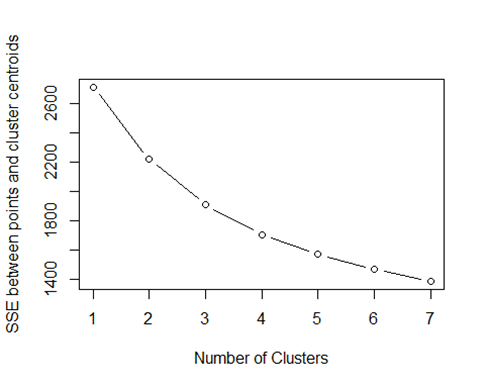
a) use domain knowledge to pick out relevant and non-redundant features; this allows for better interpretability because original features are preserved

b) correlation-based selection, i.e. eliminate variables from pairs of correlated ones; this is more systematic, and still preserves features

c) do PCA first before clustering; this is more efficient for the task at hand, but takes away from interpretability

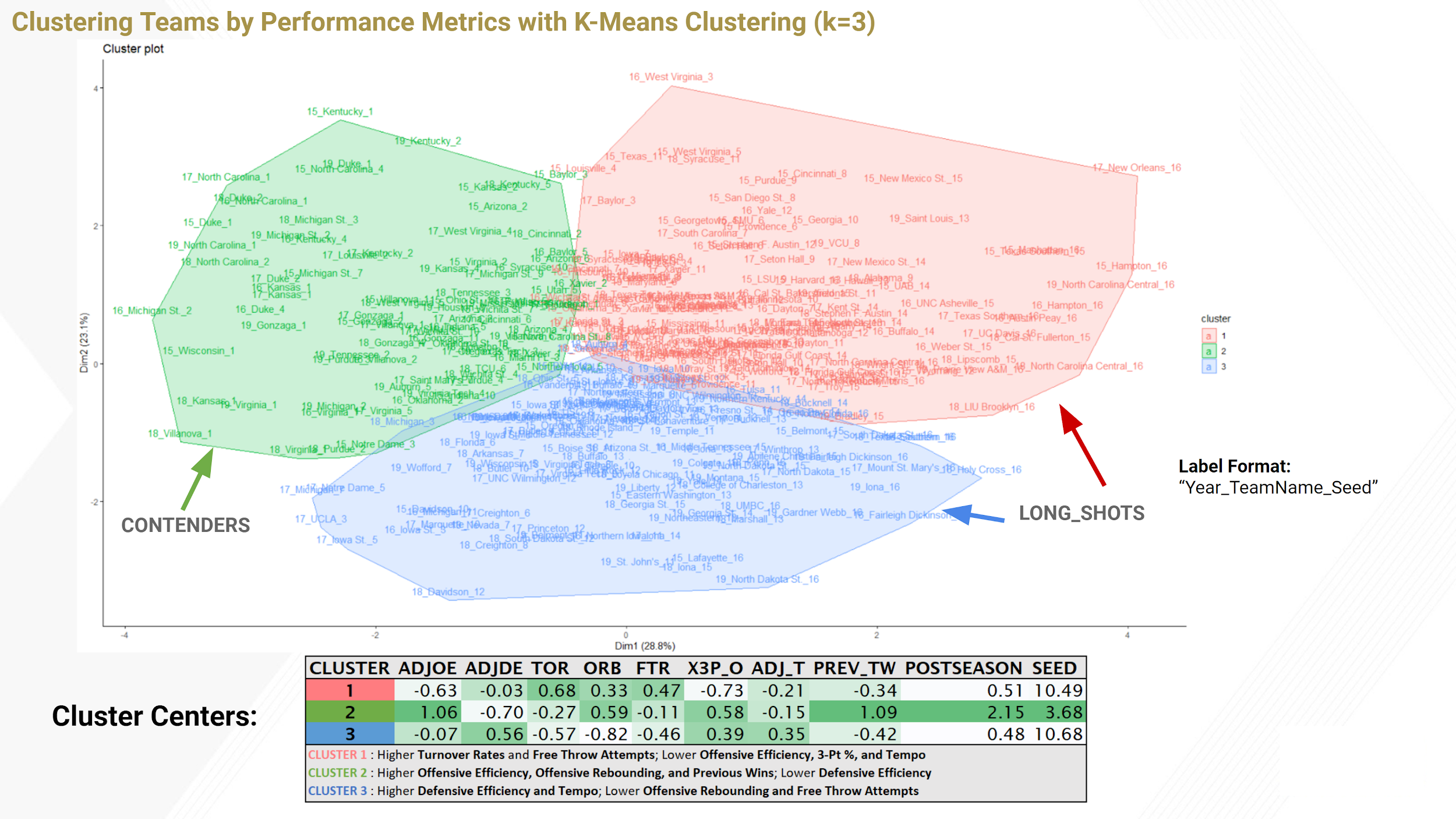
Since we prefer interpretability of our model, we’ll choose a combination of options a and b. First we’ll use domain knowledge to pick out features we think are relevant and not redundant. The hand-picked features are a reduced subset of the gameplay metrics, including off/def efficiency, offensive rebounds, 3pt, turnovers, free throws, and experience. We can also look at the reduced set of variables on the correlation matrix above to see if any others need to be removed.

As we can see from the correlation matrix, none of the remaining variables have a high correlation, so we can proceed with clustering using the selected subset of features. As mentioned above, the k-means algorithm needs a parameter k to specify the number of clusters. To determine the optimal number of clusters to group our data into, we can plot the within-cluster sum of squared errors as a function of k. The optimal k value should be at the “elbow point”: where the marginal decrease of within-SSE levels off. However, based on the plot in Figure 4 below, there isn’t a clear elbow point, so we’ll try clustering with both k=3 and k=4.

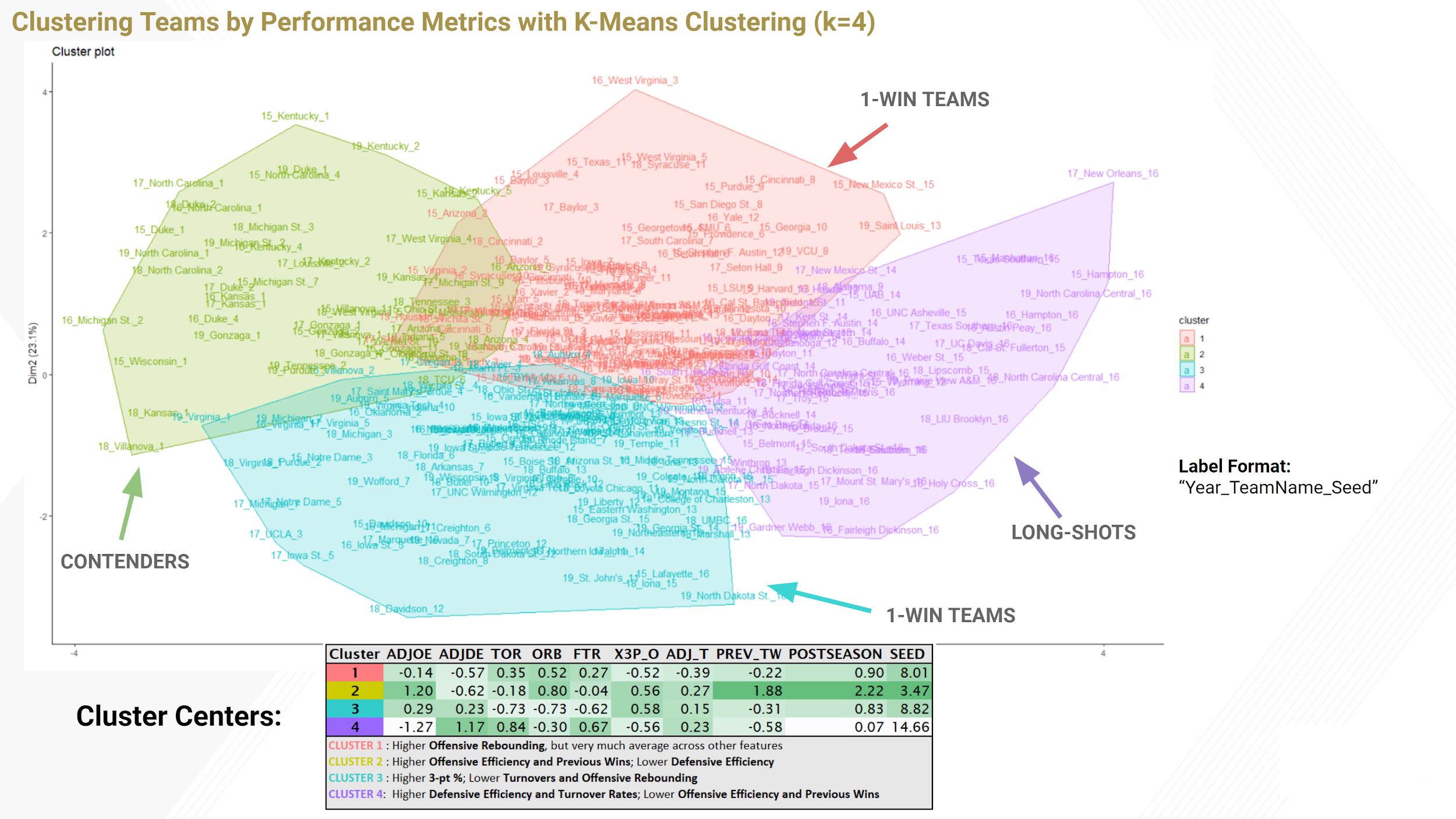


**Figure 4:** “Elbow Plot” (within SSE vs. number of clusters)

Below in Figures 5 and 6 are the PCA-reduced cluster visualizations generated from the k means algorithm with k = 3 and k = 4, respectively. If we have a priori or historical knowledge about different basketball teams over the past five years, we can visually inspect the plots to either confirm or question previously held beliefs about similar-style basketball teams. For example, we can see that the 2016-2019 Virginia basketball teams are all located closely together in the same cluster. This confirms our previously held belief that coach Tony Bennett instills a distinct and consistent style of play that is different from most other basketball teams.



**Figure 5:** K-Means Cluster Visualization (k = 3)



**Figure 6:** K-Means Cluster Visualization (k = 4)

## 

## **Methods**

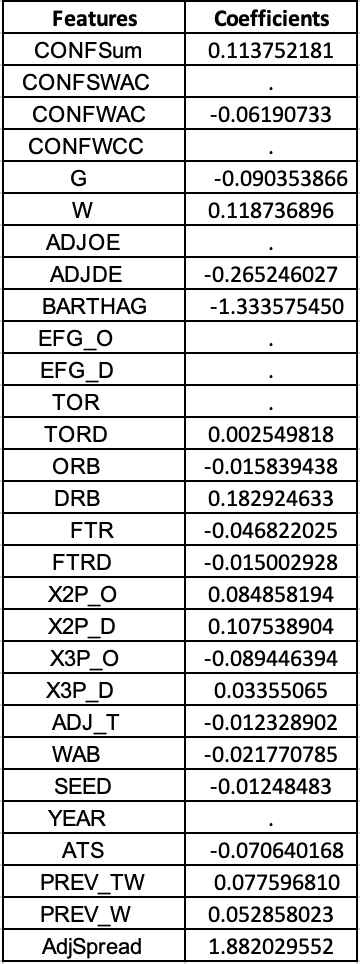
In order to predict the number of tournament games won by a given team, several models were fit and assessed. The models were built in R using libraries such as glmnet and xgboost. This prediction is a regression-type problem as the target variable is continuous over the range [0,6]. The six model types chosen for this task were lasso regression, ridge regression, multilinear regression, Multivariate Adaptive Regression Splines (MARS), boosting, and random forest. For models with varying parameter values, such as lambda in lasso regression, an initial cross validation was run with built in cross validation functionality to find the best parameter values. Once the best parameter models were gathered, all of the models were compared against each other using ten fold K-folds cross validation. Two separate metrics were used to score the models. The first is the mean squared error (MSE) of the model’s prediction against the actual number of NCAA tournament games won. The second is a custom metric called “accuracy” that measures the percentage of games correctly predicted within plus or minus a half game of the true value. This metric was developed because it may have more applicability to game winner prediction if this model is used to fill out a March Madness Bracket than a pure MSE score. Each model was scored and the average score for the ten cross validation folds is displayed in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | LASSO | Ridge | MLR | MARS | Boosting | Random Forest |
| MSE | 0.850 | 0.862 | 0.869 | 0.873 | 0.872 | 0.885 |
| +/- 0.5 Game Prediction Accuracy | 0.524 | 0.478 | 0.487 | 0.577 | 0.469 | 0.493 |

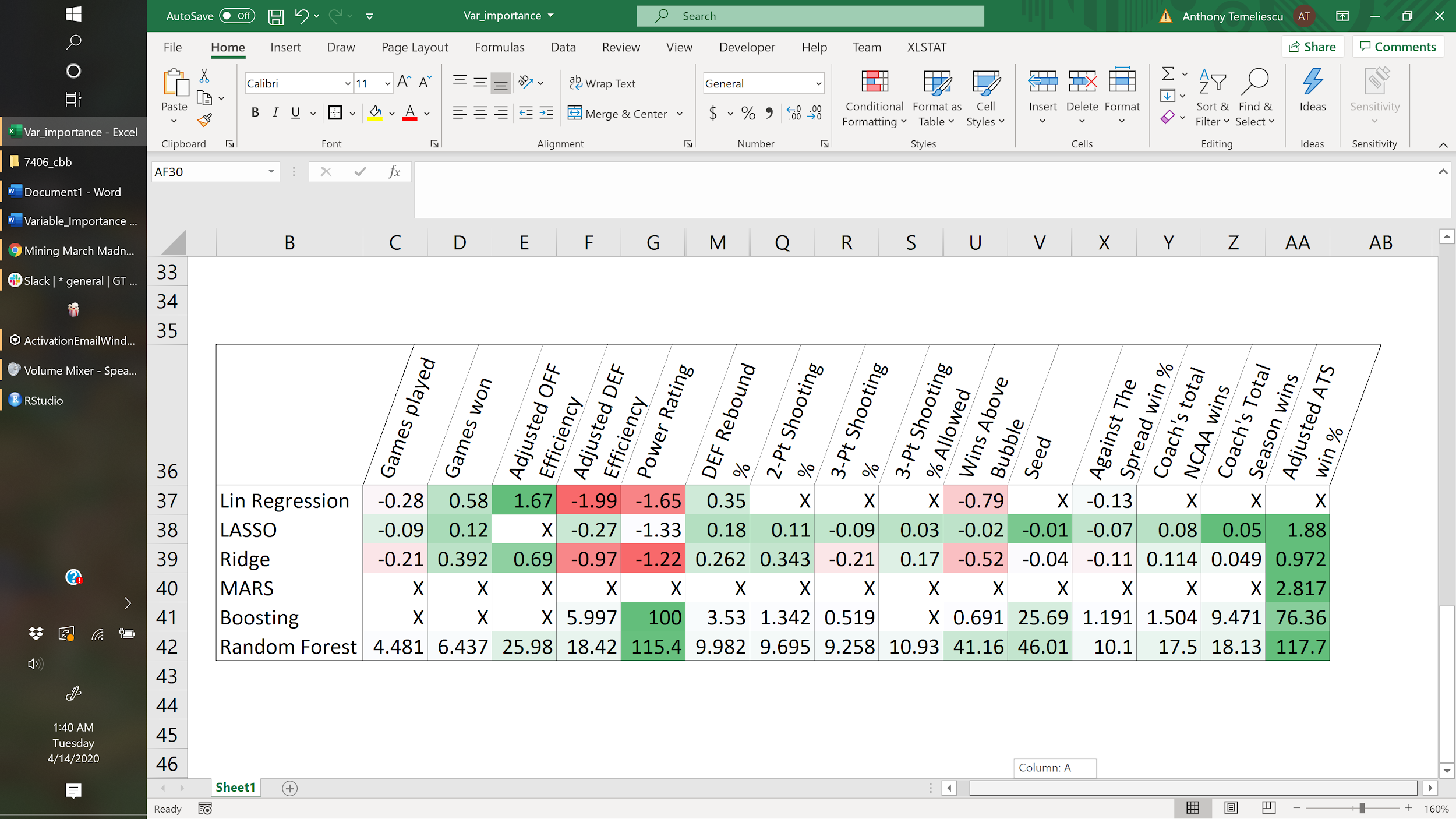
**Table 2: MSE and Accuracy Results of Models Tested**

## **Results**

It can be seen from the above table that the best model by MSE score was Lasso Regression. Lasso also has the second highest accuracy after MARS and is still above 50% accurate as measured by the custom metric. Due to its high sores in both measurement categories, LASSO Regression was chosen as the model type for the final predictions. It should be noted that the accuracy of this model for predicting any given team’s tournament wins is barely above 50%, only slightly above random guessing. In fact, when our model was used to generate a predicted winners bracket for the 2019 tournament and compared against the actual results it could be seen that most of the games that our model incorrectly predicted were cases where the two teams were extremely evenly matched measured by predicted wins. Either team could have won the game that day and that is what makes March Madness simultaneously so entertaining and so difficult to predict. Further insights can be derived from the final LASSO model trained on the full dataset to inform game predictions and coaching changes. The coefficients for the model are found in the output below.



**Table 3:** coefficients from cross validation of LASSO model



**Figure 7:** Heatmap of scaled significant explanatory variables of each model

The two most important factors were Adjusted Spread and BARTHAG. Adjusted spread, the difference between a team’s offensive and defensive efficiency, scores teams that are excellent on offense and defense more than teams that only excel in one of the two areas. BARTHAG, or the probability of a team beating a statistically “average” NCAA Division I team in a single game, has a strong negative effect on predicted wins. This is surprising as one would expect a team with a better probability of winning against an average team to perform better in the tournament. This, however, is not the case in the LASSO model fitted to the data. It seems that the artificial BARTHAG measurement does not correspond with winning in the presence of the other included factors. Other important factors include Adjusted Defensive Efficiency, Defensive Rebounding Rate, wins, and conference.

## **Conclusions**

Even with advanced data mining techniques, predicting March Madness team wins remains difficult to do with good accuracy. Despite this, these techniques are still valuable in two applications. The first is for team coaches. The March Madness performance ridge regression model coefficients can be used to determine which factors are useful in improving team tournament wins. All other factors held constant, several factors provide outside improvements or hindrances on win numbers. Teams from stronger conferences such as the ACC or Big 12 have higher win predictions. Conversely, conferences such as Big Sky or CAA are negatively impacted by their conference. This is likely due to top conferences putting teams through a more difficult regular season schedule, having better talent overall, and having good coaches. For a coach of a team in a less competitive conference, this signifies that the team may get some benefit from scheduling early-season out of conference games against opponents from strong conferences to help diminish the effect of playing conference games against weaker opponents.

In the future, these modeling techniques can be expanded upon and improved. More granular data such as player level statistics and composition of team positions could be added. More data beyond 5 years of tournaments would also help reduce the variance of this model. There are two different ways this model could be extended. The first direction uses the same target variable: number of tournament wins. This choice of target variable reflects on the whole how good a given team is in a year and how well that performance is expected to translate into tournament play. This is a more general measure of success. The other, more specific method, is to predict a tournament’s outcome game by game. This would require more granular data and relies on predicting outcomes of specific head-to-head matchups between two different teams. The seeding and order of games played has a much greater effect on this method’s predictions than using the tournament wins prediction method. Both are useful. A coach may be more interested in how generally good his or her team is heading into a tournament or when recruiting specific types of players to improve important metrics such as defensive rebounding or offensive efficiency. Someone who is betting on a March Madness bracket may be more interested in the game-by-game winner prediction modeling technique as it forecasts a single specific bracket outcome. Overall, the pursuit of better NCAA Tournament win predictions is important to fans and coaching staff alike. Even though the crystal ball of perfect win prediction remains elusive, better data and analytical techniques are bringing that closer to reality.

## **Lessons Learned**

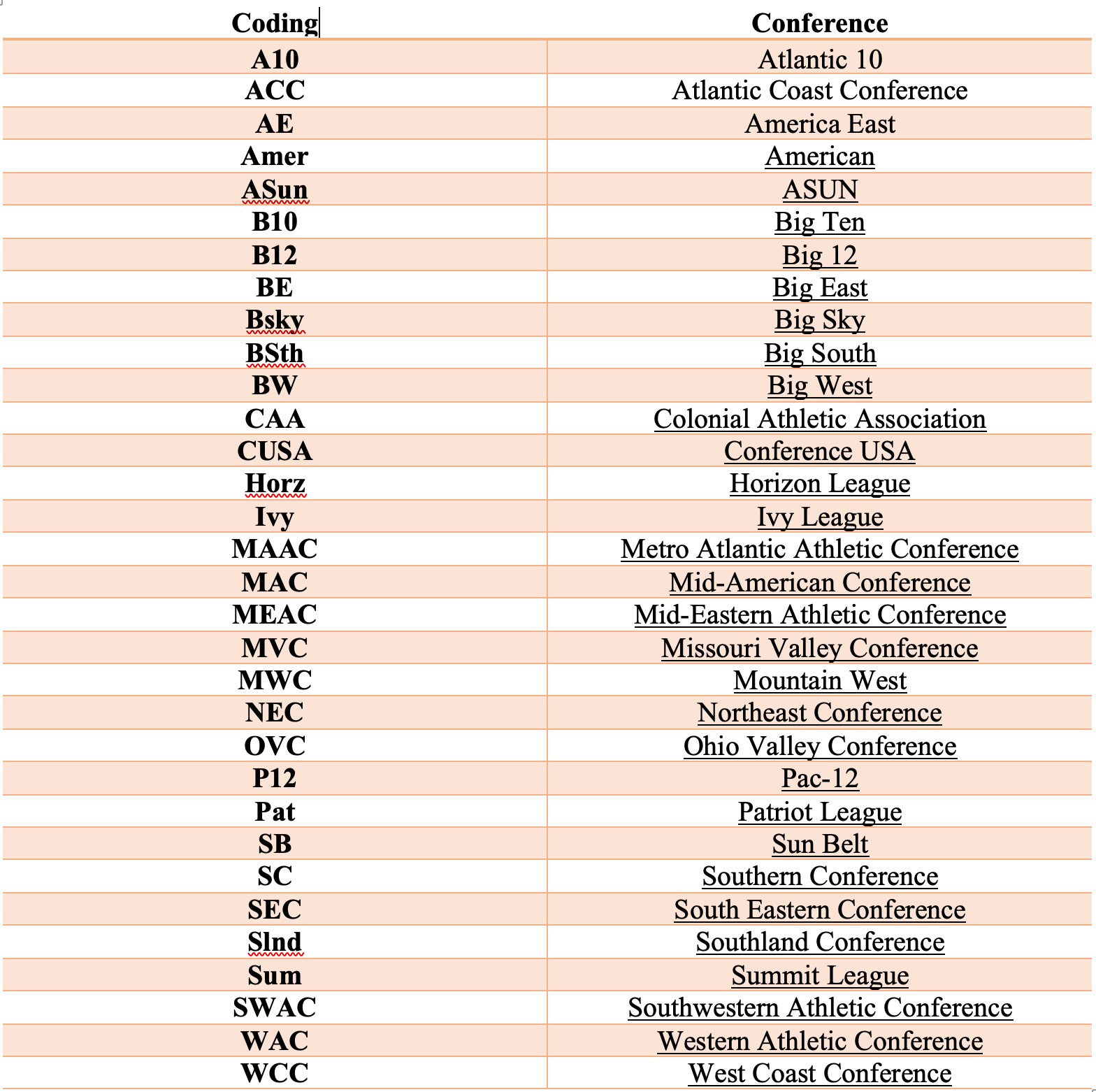
An important lesson that we came across was understanding the meaning and implications of the data we are working with. This came up when we had already developed the models on our joined dataset, and were interpreting the predictions created by our models. The models reported accuracies in the 80% range. But we realized that we included a “Games Played” metric that was a count of all games played by a team throughout a season, which included the March Madness tournament. Given that this metric includes the outcome of the tournaments (since a team will play more games in a season as they advance through the tournament), we realized this was not valid for a model who is attempting to predict future tournament results. We changed that column to be the number of conference tournament games played before March Madness, which ultimately led to a less accurate- but valid- model.

Additionally, we learned how to develop efficient cross validation methods that make use of parallel computing to allow the validation of models across a much finer set of parameter values. The benefits to using cross validation for objective model selection were also heavily reinforced.

Furthermore, we evaluated our models based on two metrics: MSE and +/- 0.5 Prediction Accuracy. Although the LASSO model performed the best in regards to MSE, it did not perform as well as the MARS model when considering prediction accuracy. We resolved this discrepancy in performance by considering a main objective of our analysis: providing coaches with interpretable insight into the factors that lead to more tournament wins. Therefore, we decided to select the LASSO model because we could obtain clearly identifiable feature importance. We learned that model selection requires an examination of all considerations of the analysis.

## 

## **Appendix**

****

**Table 4: Coding for CONF Variable**

## 

## 

## **Bibliography**

[1] Sundberg, A. (2020, March 31). College Basketball Dataset. Retrieved from <https://www.kaggle.com/andrewsundberg/college-basketball-dataset#cbb.csv>

[2] NCAAB - 2019-2020 Betting Statistics: Covers. (n.d.). Retrieved from <https://www.covers.com/sport/basketball/ncaab/statistics>

[3] 2020 Pomeroy College Basketball Ratings. (n.d.). Retrieved from https://kenpom.com/