***Predicting Wins Above the Bubble (WAB) for College Basketball Teams Using the 2019-20 Season***

*Olin Yoder*

**List of Tables** *1 Data Dictionary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .6  
2 Summary Statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .7  
3 Summary Statistics (WAB). . . . . . . . . . . . . . . . . . . . . . . ... . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9  
4 VIF Saturated Model ... . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21*

*5 VIF Naïve Model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25*

*6 VIF Best Subset Model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28*

*7 Model Comparison. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33*

**List of Figures** *1 WAB Distribution. . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9  
2 Adjusted Offensive Efficiency vs WAB . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10  
3 Adjusted Defensive Efficiency vs WAB. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11*

*4 Conference Status vs WAB. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12  
5 ANOVA (WAB by Conference Status) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13*

*6 Adjusted Tempo vs WAB. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14*

*7 Win Percentage vs WAB . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .15*

*8 Win Percentage & Conference Status vs WAB . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17*

*9 Pair Plot. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18*

*10 Saturated Model . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20*

*11 Diagnostic Plots – Saturated Model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . 22*

*12 Potential Outliers – Saturated Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .23*

*13 Naïve Model . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24*

*14 Diagnostic Plots – Naïve Model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .25*

*15 Best Subset Number of Predictors. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27*

*16 Best Subset Model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28*

*17 Diagnostic Plots – Best Subset Model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29*

*18 Distribution of March Madness Champions. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36*

**Title**

Predicting Wins Above the Bubble (WAB) for College Basketball Teams Using the 2019-20 Season

**Abstract**

This analysis explores the relationship between a college basketball team’s performance relative to its peers -- combined with its offensive and defensive statistics -- and how these factors affect its Wins Above the Bubble (WAB). The dataset includes the following variables: Rank, Team, Conference, Games, Wins, Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, BARTHAG, Effective Offensive Field Goal Percentage, Effective Defensive Field Goal Percentage, Turnover Rate, Steal Rate, Offensive Rebound Rate, Defensive Rebound Rate, Free Throw Rate, Free Throw Rate Allowed, Two-Point Shooting Percentage, Two-Point Defensive Percentage, Three-Point Shooting Percentage, Three-Point Defensive Percentage, Adjusted Tempo, and WAB. For detailed definitions of each variable, see **Table 1**. By identifying which variables are strongly correlated with WAB, they can be used to build predictive models. Multiple regression techniques -- including linear, ridge, and LASSO regression -- are applied to assess model performance and predictive capability.

**Introduction**

**The Lost March Madness: A Statistical What-If**

2020 was marked by the onset of COVID-19 and subsequent cancellation of numerous events worldwide. As a basketball fan, one of the most disappointing cancellations was March Madness. a month-long spectacle of 67 games filled with chaos, joy, and heartbreak. Though the tournament was never played, it left behind one of the greatest “what-if” scenarios in college basketball history: *What if the 2020 NCAA Tournament had happened? Which teams would have made the field? Who would have cut down the nets?*

While there is no way to definitively determine the outcome of an event that never happened, especially one with as much unpredictability as March Madness, the 2019-20 season still provided a wealth of data prior to its cancellation.

**Background**

In 2020, 68 NCAA Division 1 (D1) teams were set to qualify for March Madness. Of these teams, 32 teams would receive an automatic bid by winning their conference tournament, while 36 teams would receive an *at-large* bid (the PAC-12 has since dissolved resulting in 31 automatic bids and 37 at-large bids). At-large bids are awarded by a selection committee and based on a team's overall performance up to that point in the season. Naturally, this selection process creates a cutoff, raising the question: *Is the 36th team that makes the tournament truly more deserving than the 37th team that just missed out?* This area of uncertainty is referred to as the "bubble."

In college basketball analytics, there is a statistic called Wins Above the Bubble (WAB) that quantifies how comfortably a team would make the tournament. A team with a high WAB is safely in the tournament, while a team on the bubble has a WAB close to zero -- meaning they have about a 50/50 chance of receiving an at-large bid -- and teams with a WAB below zero are unlikely to receive an at-large bid. For reference, Kansas -- who was ranked No. 1 in numerous rankings at the end of the 2019-20 season -- had a WAB of 10.8, meaning they had 10.8 more wins given their schedule than the average team on the bubble team like Oklahoma St., which had a WAB of 0.0. Conversely, Kennesaw State, a team who finished the year with one win, had a WAB of -21.8.

**Research Question**

Since the tournament was never played, we will never know for certain which teams would have made the field. However, the 2019-20 season still provided most of a season’s worth of data. Using this data, we can attempt to answer the question:

*Using a team’s performance relative to their peers and their offensive/defensive statistics, can a model be developed to predict WAB?*

If successful, such a model could be applied to future seasons to estimate which teams are safely in the tournament, on the bubble, or likely to miss out.

**Materials & Methods**

**Data Sources**

Data was sourced from Kaggle and originally scraped from [barttorvik.com](https://barttorvik.com), a website that compiles and tracks statistics from the college basketball season. In game box scores and statistics are tracked by official scorekeepers and tracking software.

The data set features 353 observations across 22 variables, including Team, which serves as an identification field. The names of the 22 variables as provided are: *RK, TEAM, CONF, G, W, ADJOE, ADJDE, BARTHAG, EFG\_O, EFG\_D, TOR, TORD, ORB, DRB, FTR, FTRD, X2P\_O, X2P\_D, X3P\_O, X3P\_D, ADJ\_T,* and *WAB*. The data set contains no null values.

**Statistical Analysis**

Data was downloaded in the provided.csv format. All analyses are conducted using R (version 4.2.1). The primary objective of the analysis is to identify relationships between the given variables and WAB, specifically to determine how the variables can predict WAB using multiple linear regression. Before training any regression model, each variable is examined individually to understand its relationship with WAB. Next, various multiple regression techniques are then applied, ultimately leading to a final model and the interpretation of said model.

**Model Assumptions**

Unless otherwise noted, all inferences are made using a significance level (α) of 0.05. The variable, *CONF*, is categorical and are dummy encoded during the regression process. Additionally, model assumptions, including the constancy of error variances, normality of the error terms, independence of errors, and linearity of the relationship between the outcome and predictors, are verified before finalizing the estimated fitted regression model.

**Primary Objective Analysis**

Explore each variable and its relationship with WAB. Then, use the variables that best explain WAB to predict and assess each model’s performance.

**Summary of Dataset**

**Table 1**, below, contains the data dictionary.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column** | **Description** | **Data Type** | **Format** | **Example** |
| TEAM | Division 1 College Basketball School | Character | Text | Kansas |
| RK | Team ranking at end of regular season (Barttorvik.com) | Integer | Whole Number | 1 |
| G | Number of games played | Integer | Whole Number | 30 |
| W | Number of games won | Integer | Whole Number | 28 |
| CONF | Athletic conference (e.g., A10, ACC, B12, SEC, etc.) | Category (Character) | Abbreviation | B12 |
| ADJOE | Adjusted Offensive Efficiency (points scored per 100 possessions vs. average D1 defense) | Float | Decimal | 116.1 |
| ADJDE | Adjusted Defensive Efficiency (points allowed per 100 possessions vs. average D1 offense) | Float | Decimal | 87.7 |
| BARTHAG | Power Rating (Chance of beating an average D1 team) | Float | Decimal | 0.9616 |
| EFG\_O | Effective Field Goal Percentage (offense) | Float | Percentage | 53.7 |
| EFG\_D | Effective Field Goal Percentage (defense) | Float | Percentage | 43.7 |
| TOR | Turnover Percentage (offense) | Float | Percentage | 18.7 |
| TORD | Turnover Percentage (defense - steal rate) | Float | Percentage | 18.6 |
| ORB | Offensive Rebound Rate | Float | Percentage | 32.6 |
| DRB | Defensive Rebound Rate (offensive rebounds allowed) | Float | Percentage | 26.4 |
| FTR | Free Throw Rate (offense) | Float | Percentage | 35.8 |
| FTRD | Free Throw Rate Allowed (defense) | Float | Percentage | 23.2 |
| 2P\_O | Two-Point Shooting Percentage (offense) | Float | Percentage | 54.9 |
| 2P\_D | Two-Point Shooting Percentage Allowed (defense) | Float | Percentage | 42.4 |
| 3P\_O | Three-Point Shooting Percentage (offense) | Float | Percentage | 34.1 |
| 3P\_D | Three-Point Shooting Percentage Allowed (defense) | Float | Percentage | 30.5 |
| ADJ\_T | Adjusted Tempo (estimated possessions per 40 minutes) | Float | Decimal | 67.4 |
| WAB | Wins Above Bubble (safety of NCAA Tournament qualification) | Float | Decimal | 10.8 |

**Table 1:** Data Dictionary

**Table 2**, below, contains five figure summary statistics of the dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Min** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max** |
| RK | 1 | 89 | 177 | 177 | 265 | 353 |
| TEAM | (character) | (character) | (character) | (character) | (character) | (character) |
| CONF | (character) | (character) | (character) | (character) | (character) | (character) |
| G | 24.00 | 29.00 | 30.00 | 30.19 | 31.00 | 34.00 |
| W | 1.00 | 13.00 | 16.00 | 16.31 | 20.00 | 31.00 |
| ADJOE | 80.1 | 97.3 | 102.2 | 102.2 | 106.7 | 121.3 |
| ADJDE | 85.6 | 98.0 | 102.0 | 102.2 | 106.4 | 122.7 |
| BARTHAG | 0.0194 | 0.2818 | 0.4804 | 0.4977 | 0.7207 | 0.9616 |
| EFG\_O | 39.30 | 47.60 | 49.60 | 49.57 | 51.50 | 59.70 |
| EFG\_D | 41.20 | 47.60 | 49.50 | 49.61 | 51.50 | 58.40 |
| TOR | 13.60 | 17.80 | 18.80 | 18.92 | 20.30 | 26.60 |
| TORD | 14.0 | 17.4 | 18.7 | 18.9 | 20.3 | 27.8 |
| ORB | 14.20 | 25.20 | 27.90 | 27.89 | 30.60 | 40.10 |
| DRB | 19.00 | 26.00 | 28.00 | 27.97 | 29.90 | 36.90 |
| FTR | 21.70 | 29.10 | 32.60 | 32.64 | 35.80 | 48.10 |
| FTRD | 19.70 | 28.90 | 31.90 | 32.79 | 36.30 | 53.00 |
| X2P\_O | 40.20 | 47.30 | 49.40 | 49.43 | 51.50 | 62.30 |
| X2P\_D | 40.70 | 47.40 | 49.40 | 49.43 | 51.60 | 58.40 |
| X3P\_O | 24.80 | 31.50 | 33.30 | 33.19 | 34.90 | 41.90 |
| X3P\_D | 27.20 | 31.60 | 33.20 | 33.27 | 34.70 | 40.60 |
| ADJ\_T | 59.4 | 66.4 | 68.2 | 68.3 | 69.8 | 77.4 |
| WAB | -25.200 | -12.800 | -7.600 | -7.701 | -3.100 | 10.800 |

**Table 2:** Summary Statistics

**Exploratory Data Analysis & Preprocessing**

**Variable Analysis**

For brevity, not all variables provided are discussed in detail in the variable analysis section (although each can be seen in **Figure 9**).

**Preprocessing**

After conducting exploratory statistical analysis, the next step is to use the available variables to predict WAB. However, before building any predictive models, several considerations must be addressed. First, as mentioned earlier, there are no NA or NULL values in the dataset, so no action is needed for missing data.

Generally, stronger offensive statistics (where higher values are better) and stronger defensive statistics (where lower values are better) are expected to be associated with higher WAB values. Additionally, offensive metrics such as Adjusted Offensive Efficiency (ADJOE) and Effective Field Goal Percentage (EFG\_O) are likely to be highly correlated with each other, as are defensive metrics. This could lead to multicollinearity which could pose challenges during modeling and may need to be addressed. In anticipation of this, one option for preprocessing would be to apply dimensionality reduction techniques, such as Principal Component Analysis (PCA) or autoencoders. However, given the nature of the assignment, I elected to retain the original variables and address any multicollinearity during feature selection instead.

For feature engineering, instead of using all 32 distinct conferences individually, teams are grouped conferences into three categories: high major, mid major, and low major. High major teams belong to the Power 6 Conferences in college basketball (ACC, Big Ten, Big 12, Big East, Pac-12, SEC). These conferences typically consist of larger institutions, such as flagship universities (see [CollegeNetWorth.com](https://www.collegenetworth.com/what-are-the-power-6-conferences-in-basketball/) for more information). Mid major teams are classified as the ten highest-rated non-Power 6 conferences by [Rating Percentage Index (RPI) in 2019–20](https://www.teamrankings.com/ncaa-basketball/rpi-ranking/rpi-rating-by-conf?date=2020-03-12), while the remaining conferences are labeled as low major.

Additionally, certain columns -- such as Rank, BARTHAG, Wins, and Games Played -- are excluded from the modeling process. Since the 2019–20 season was canceled partway through, teams played a varying number of games, making raw win totals potentially misleading. A more accurate measure is Win Percentage, calculated as Wins divided by Games Played. Furthermore, Rank and BARTHAG are derived metrics from Barttorvik.com. Because their exact calculation methods are not publicly available, and because the analysis is intended to focus specifically on offensive and defensive team performance, these variables are not included in the modeling process.

**A graph of a graph

Description automatically generatedWAB Distribution**

**Figure 1:** WAB Distribution

**Figure 1** shows the distribution of WAB values. The distribution is approximately normal, with both the mean and median centered around -7.5 (meaning the average team is 7.5 games below the bubble -- or making the tournament). This is expected, as only 64 of the 353 D1 teams qualify for the NCAA tournament (excluding play-in games).

There are no obvious outliers present in the histogram, although a boxplot would be more appropriate for identifying potential outliers. Additionally, summary statistics for WAB are provided in **Table 3** below. The first and third quartiles are roughly equidistant from the median, which supports the observation that the distribution is approximately normal.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Q | Median | Mean | 3rd Q | Max |
| -25.20 | -12.80 | -7.60 | -7.70 | -3.10 | 10.8 |

**Table 3:** Summary Statistics (WAB)

**WAB vs Adjusted Offensive Efficiency (ADJOE)**

A graph with black dots

Description automatically generated

**Figure 2:** ADJOE vs WAB

**Figure 2** shows the relationship between adjusted offensive efficiency and WAB. There is a strong positive linear relationship between adjusted offensive efficiency and WAB, supported by a correlation coefficient of 0.831. This high correlation suggests that adjusted offensive efficiency is likely a strong predictor of WAB. In other words, teams with more efficient offenses are more likely to make the tournament than teams with poor offenses.

**WAB vs Adjusted Defensive Efficiency (ADJDE)**

A graph with black dots

Description automatically generated

**Figure 3:** ADJDE vs WAB

**Figure 3** shows the relationship between adjusted defensive efficiency and WAB. There is a strong negative linear relationship between adjusted defensive efficiency and WAB, supported by a correlation coefficient of 0.807. This high correlation suggests that adjusted defensive efficiency is likely a strong predictor of WAB. In other words, the more points a team gives up on defense, the less likely they are to make the tournament.

**WAB vs Conference Status**

A graph with a row of rectangular objects

Description automatically generated with medium confidence

**Figure 4:** Conference Status vs WAB

**Figure 4** displays the distributions of WAB across the three conference status groups. Each group’s WAB distribution appears approximately normal. As conference prestige increases, WAB generally increases as well. Notably, the lower quartile of WAB for power conferences is higher than the upper quartile for mid major conferences, indicating a clear separation. While mid major conferences have a higher median WAB than low major conferences, their interquartile ranges (IQRs) overlap, suggesting some similarity between those two groups.

Given the differences in WAB distributions across conference statuses, an ANOVA test is appropriate to assess whether the mean WAB differs by group. The null hypothesis (H₀) assumes that the mean WAB is the same across all conference statuses, while the alternative hypothesis (H₁) states that at least one group differs. The ANOVA test returns a p-value of 2e-16, which is far below the 0.05 significance threshold. Thus, the null hypothesis is rejected and there is a statistically significant difference in mean WAB across the different conference statuses.

Although ANOVA confirms that at least one group mean differs, post-hoc t-tests would be helpful to identify which specific pairs of conference statuses differ. Overall, the results suggest that conference status is likely a strong predictor of WAB.

A close-up of numbers

Description automatically generated

**Figure 5:** ANOVA (WAB by Conference Statuses)

**WAB vs Adjusted Tempo (ADJT)**

A graph with black dots

Description automatically generated

**Figure 6:** ADJT vs WAB

**Figure 5** shows the relationship between a team’s adjusted tempo and their WAB. Visually, there appears to be no meaningful relationship between the two variables, which is supported by a low correlation coefficient (r = -0.140). This suggests that teams can be successful in terms of WAB regardless of whether they play at a fast or slow pace. Given the weak and negative correlation, adjusted tempo is likely not a significant predictor of WAB.

**WAB vs Win Percentage**

**A graph with black dots

Description automatically generated**

**Figure 7:** Win Percentage vs WAB

**Figure 6** shows the relationship between win percentage and WAB. There is a strong positive linear relationship between win percentage and WAB, supported by a correlation coefficient of 0.857. This high correlation suggests that win percentage is a strong predictor of WAB.

In other words, the higher percent of games a team wins, the more likely they are to make the tournament.

However, the scatterplot appears to show two roughly parallel trends or clusters. This visual pattern suggests that another variable may be influencing WAB in conjunction with win percentage. Based on the patterns observed in **Figure 4**, it is plausible that these two lines represent different conference statuses, with power conference teams generally forming the higher cluster and mid and low major Conference teams forming the lower one.

If this is the case, an interaction term between win percentage and conference status could be valuable in a regression model to better capture the relationship between these variables and WAB.

**WAB vs Win Percentage & Conference Status**

**A graph with different colored dots

Description automatically generated**

**Figure 8:** Win Percentage & Conference Status vs WAB

**Figure 7** supports the hypothesis that the two distinct lines observed in **Figure 6** are due to differences in conference status. The upper trend line corresponds to teams in power conferences while the lower line represents teams from mid major and low major conferences.

This pattern indicates that two teams with the same win percentage can have significantly different WAB values depending on their conference status. In other words, teams from more prestigious conferences tend to receive more credit -- reflected in higher WAB -- for the same level of on-court success. This finding reinforces the importance of considering conference status as both a main effect and a potential interaction term when modeling WAB.

**A graph of a number of data

Description automatically generated with medium confidencePair Plot**

**Figure 9:** Pair Plot

**Figure 9** is a pair plot, which, in this case, is useful for determining variable correlations with WAB and also correlations with other variables.

As expected, effective field goal percentage on offense is highly correlated (r = 0.693) with adjusted offensive efficiency. Similarly, effective field goal percentage on defense shows a strong correlation (r = 0.823) with adjusted defensive efficiency. These relationships are intuitive, as shooting efficiency is a major component of overall efficiency.

More broadly, the pair plot highlights a trend: offensive metrics tend to be highly correlated with other offensive metrics, and the same holds true for defensive metrics. This clustering of related variables suggests multicollinearity.

For a more larger view of the graph, please download [here](https://github.com/olinyoder2534/CollegeBasketballSimulator/blob/main/PairPlot).

**Modeling**

Several linear regression strategies are employed to model the data, including saturated, naïve, best subset, and regularized approaches using ridge and LASSO regression. For each model, diagnostic plots are examined, and performance is evaluated using accuracy metrics to assess model fit and predictive capability.

**Saturated Model**

A saturated model, or a model using all variables, is used as a baseline model. The R output is as follows:

**A screenshot of a computer

Description automatically generated**

**Figure 10:** Saturated Model

There are several notable issues with the saturated model. First, the majority of the predictors are not statistically significant at the 0.05 level, suggesting that many variables may not meaningfully contribute to the prediction of WAB. While the adjusted R2 value is high (0.9753), indicating that 97.53% of the variance in WAB is explained by the model, this may point to overfitting -- especially given the large number of predictors with insignificant coefficients.

Beyond coefficient estimates and their significance, it is also essential to evaluate other aspects of model performance, such as residual patterns, multicollinearity, and predictive accuracy on unseen data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **GVIF** | **Df** | **GVIF^(1/(2\*Df))** |
| ADJOE | 14.404210 | 1 | 3.795288 |
| ADJDE | 17.863490 | 1 | 4.226522 |
| EFG\_O | 211.833199 | 1 | 14.554491 |
| EFG\_D | 318.443369 | 1 | 17.844982 |
| TOR | 3.941936 | 1 | 1.985431 |
| TORD | 5.216987 | 1 | 2.284072 |
| ORB | 3.822065 | 1 | 1.955010 |
| DRB | 3.432967 | 1 | 1.852827 |
| FTR | 1.544799 | 1 | 1.242900 |
| FTRD | 1.838618 | 1 | 1.355956 |
| X2P\_O | 95.958758 | 1 | 9.795854 |
| X2P\_D | 164.769013 | 1 | 12.836238 |
| X3P\_O | 56.256472 | 1 | 7.500431 |
| X3P\_D | 70.201884 | 1 | 8.378656 |
| ADJ\_T | 1.270744 | 1 | 1.127273 |
| CONF\_STATUS | 4.314812 | 2 | 1.441254 |
| WinPercentage | 6.532866 | 1 | 2.555947 |

**Table 4:** VIF - Saturated Model

The VIF table reveals a high degree of multicollinearity among the predictor variables in the regression model, with most variables having VIF values exceeding 5 and some surpassing 100. This indicates substantial redundancy, meaning that several predictors are highly correlated and provide overlapping information. High multicollinearity complicates the interpretation of regression coefficients because it becomes difficult to isolate the individual effect of each variable on the outcome. As a result, coefficient estimates can become unstable and highly sensitive to small changes in the data, leading to large standard errors and wide confidence intervals. Significance tests for individual predictors may also be misleading, with variables appearing non-significant despite having a true effect due to inflated standard errors. Additionally, the direction and magnitude of coefficients can be counterintuitive or vary drastically depending on which variables are included in the model. While the overall model fit may remain high, the reliability of interpreting individual predictors’ impacts is undermined. Therefore, although multicollinearity does not necessarily reduce predictive power, it significantly hinders the ability to understand and trust the relationships estimated by the model.

**A group of graphs with numbers

Description automatically generatedFigure 11:** Diagnostic Plots - Saturated Model

Based on the diagnostic plots, there do not appear to be any major issues affecting the model’s performance. The residuals versus fitted values plot shows residuals randomly distributed around zero, indicating homoscedasticity. The Q-Q plot closely follows the dashed line, suggesting that the residuals are approximately normally distributed. Additionally, the residuals versus leverage plot does not reveal any major influential points. No observations have a Cook’s Distance greater than 0.5, nor do any exceed the threshold expected under the F-distribution given the number of predictors and sample size.

While observation 353 (Chicago State) is a leverage value (df fits greater than 1 and a studentized residual of 3.084), it does not exhibit high influence based on Cook’s Distance or df betas. Furthermore, its residual is not exceptionally large compared to other points (as shown in **Figure 12** below). Therefore, this observation will be retained in the model.

**A graph of a plot

Description automatically generated with medium confidence**

**Figure 12:** Potential Outlier - Saturated Model

**Native Model**

A screenshot of a computer code

Description automatically generatedSaturated models are often overly complex and include variables that are not statistically significant. Removing these non-significant variables typically maintains or improves performance while reducing model complexity. Variable selection is based on insights from exploratory data analysis, with a focus on variables that exhibit strong correlations with WAB but are not highly correlated with each other, thereby avoiding redundancy. For instance, instead of including multiple offensive and defensive metrics, only the variables that best capture overall team offensive and defensive performance -- namely, adjusted efficiency -- are used. Additionally, the interaction term between win percentage and conference status, as identified in **Figure 8**, is included in the model.

**Figure 13:** Naïve Model

The naïve model predicts WAB using adjusted offensive efficiency (ADJOE), adjusted defensive efficiency (ADJDE), win percentage, conference status, and their interactions. Both ADJOE and ADJDE are highly significant predictors: higher offensive efficiency is associated with greater WAB, while higher (i.e., worse) defensive efficiency corresponds to lower WAB. Win percentage also has a strong positive effect on WAB, though this effect varies by conference type. Specifically, mid major and power conference teams see a greater increase in WAB for improved win percentages compared to low major teams. The interaction terms between win percentage and conference status are statistically significant, indicating that the influence of winning percentage differs across conference levels. The main effect for power conference teams is not significant, suggesting no inherent WAB advantage at a win percentage of zero compared to low major teams. However, the positive interaction effect aligns with expectations, as teams -- regardless of conference -- will not qualify for the tournament without winning games. Interestingly, the main effect for mid major teams is negative, indicating that at zero wins, a mid major team is even less likely to make the tournament than a low major team.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Predictor** | **GVIF** | **Df** | **GVIF^1/(2\*Df)** | **Interacts With** | **Other Predictors** |
| ADJOE | 3.327762 | 1 | 1.824215 | -- | ADJDE, WinPercentage, CONF\_STATUS |
| ADJDE | 3.008317 | 1 | 1.734450 | -- | ADJOE, WinPercentage, CONF\_STATUS |
| WinPercentage | 6.566293 | 5 | 1.207069 | CONF\_STATUS | ADJOE, ADJDE |
| CONF\_STATUS | 6.566293 | 5 | 1.207069 | WinPercentage | ADJOE, ADJDE |

**Table 5:** VIF – Naïve Model

Removing highly correlated variables helps reduce multicollinearity, as indicating by no variables having a (GVIF^(1/(2\*Df))) greater than two. Please note, when dealing with categorical variables, it is best to use the GVIF raised to the power of one divided by twice the degrees of freedom (GVIF^(1/(2\*Df))) to properly measure multicollinearity.

A group of graphs with numbers

Description automatically generated with medium confidence

**Figure 14:** Diagnostic Plots – Naïve Model

Once again, the diagnostic plots show no major issues. Although there are a few potential outliers, they have low leverage and do not significantly influence the model’s coefficients like the saturated model.

**Best Subset Regression**

While the naïve model performed well and resolved multicollinearity issues, manually selecting predictors can be time-consuming and prone to human error. In linear regression, several systematic methods for feature selection exist, including best subset regression, ridge regression, and LASSO regression, which are applied in the following analysis.

Using best subset regression to determine the optimal number of predictors, model performance improves considerably up to four predictors based on both adjusted R² and Mallows’ Cp criteria, after which the improvements plateau.

A comparison of a graph

Description automatically generated

**Figure 15:** Best Subset Number of Predictors

According to best subset regression, the optimal four-variable model includes win percentage, adjusted offensive efficiency, adjusted defensive efficiency, and conference status -- closely aligning with the variables selected in the naïve model.

A screenshot of a computer

Description automatically generated  
**Figure 16:** Best Subset Model

All variables are significant at an alpha level of 0.10, with all except the conference status of mid major being significant at the 0.001 level. The coefficients align closely with the trends observed during exploratory data analysis: as teams score more offensively, WAB increases; as defenses allow fewer points, WAB also increases; and as teams win a higher proportion of their games, their WAB rises. Additionally, teams in power conferences tend to have a higher baseline WAB than those in low major conferences, while mid major teams exhibit a slightly higher -- though only marginally significant -- baseline WAB compared to low major teams.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **GVIF** | **Df** | **GVIF^(1/(2×Df))** |
| ADJOE | 3.324352 | 1 | 1.823281 |
| ADJDE | 2.969415 | 1 | 1.723199 |
| WinPercentage | 3.555392 | 1 | 1.885575 |
| CONF\_STATUS | 3.045032 | 2 | 1.320985 |

**Table 6:** VIF – Best Subset Model

Given the similarity to the naïve model, it is unsurprising that there are again no issues with multicollinearity or diagnostic plots, as shown in the table above and figure below.

A group of graphs with numbers

Description automatically generated with medium confidence**Figure 17:** Diagnostic Plots – Best Subset Model

**L1 & L2 Regression**

For both ridge and LASSO regression, three models are fit: Model 1, a first-order additive model including all variables; Model 2, the same as Model 1 but with the interaction term between conference status and win percentage added; and Model 3, only adjusted offensive efficiency, adjusted defensive efficiency, and the interaction between conference status and win percentage as potential variables. Each model is trained using the minimum lambda value determined by 10-fold cross-validation.

**Ridge Regression**

Model 1:

WAB = −14.34811 + 0.28116(ADJOE) − 0.26259(ADJDE) + 0.00440(EFG\_O) − 0.06747(EFG\_D) − 0.08276(TOR) + 0.07563(TORD) + 0.04280(ORB) − 0.01051(DRB) + 0.04944(FTR) − 0.05274(FTRD) + 0.00075(X2P\_O) − 0.05559(X2P\_D) + 0.00199(X3P\_O) − 0.02091(X3P\_D) + 0.02994(ADJ\_T) + 0.33612(CONF\_STATUS: Mid Major) + 2.81331(CONF\_STATUS: Power Conferences) + 14.53620(WinPercentage)

**Model Performance Metrics:**

* **R-squared:** 0.9660044
* **Adjusted R-squared:** 0.9651123
* **Mean Squared Error (MSE):** 1.679534
* **Root Mean Squared Error (RMSE):** 1.295968
* **Optimal Lambda (λ):** 0.578706

Model 2:

WAB = -14.042907397 + 0.262800612(ADJOE) - 0.248219498(ADJDE) + 0.007547853(EFG\_O) - 0.069245023(EFG\_D) - 0.091787294(TOR) + 0.084343590(TORD) + 0.042598257(ORB) - 0.018190422(DRB) + 0.047925619(FTR) - 0.050429185(FTRD) + 0.006317142(X2P\_O) - 0.053659383(X2P\_D) + 0.005507174(X3P\_O) - 0.025950622(X3P\_D) + 0.030778979(ADJ\_T) - 0.424804706(CONF\_STATUS: Mid\_Major) + 1.364335669(CONF\_STATUS: Power\_Conferences) + 13.931894139(WinPercentage) + 1.811113318(CONF\_STATUS: Mid\_Major \* WinPercentage) + 3.311001380(CONF\_STATUS: Power\_Conferences \* WinPercentage)

**Model Performance Metrics:**

* **R-squared:** 0.9673427
* **Adjusted R-squared:** 0.9664858
* **Mean Squared Error (MSE):** 1.645893
* **Root Mean Squared Error (RMSE):** 1.282924
* **Optimal Lambda (λ):** 0.578706

Model 3:

WAB = -15.9917631 + 0.2771062(ADJOE) - 0.2863110(ADJDE) + 15.5018205(WinPercentage) - 0.5338145(CONF\_STATUS: Mid\_Major) + 1.4444827(CONF\_STATUS: Power\_Conferences) + 1.8555683(WinPercentage \* CONF\_STATUS: Mid\_Major) + 3.3798983(WinPercentage \* CONF\_STATUS: Power\_Conferences)

**Model Performance Metrics:**

* **R-squared:** 0.9683307
* **Adjusted R-squared:** 0.9674997
* **Mean Squared Error (MSE):** 1.50639
* **Root Mean Squared Error (RMSE):** 1.227351
* **Optimal Lambda (λ):** 0.578706

**LASSO Regression**

Model 1:

WAB = −18.67999 + 0.42389(ADJOE) − 0.36875(ADJDE) − 0.38367(EFG\_O) + 0.18132(EFG\_D) + 0.26384(TOR) − 0.19477(TORD) − 0.12493(ORB) + 0.08795(DRB) + 0.00612(FTR) − 0.00723(FTRD) − 0.01022(X2P\_O) + 0.01708(X2P\_D) − 0.03275(X3P\_O) + 0.02424(X3P\_D) + 0.04030(ADJ\_T) + 2.16690(CONF\_STATUS: Power Conferences) + 21.97613(WinPercentage)

**Model Performance Metrics:**

* **R-squared:** 0.9764349
* **Adjusted R-squared:** 0.9758166
* **Mean Squared Error (MSE):** 1.210816
* **Root Mean Squared Error (RMSE):** 1.100371
* **Optimal Lambda (λ):** 0.004082656

Model 2:

WAB = -17.065520851 + 0.417035243(ADJOE) - 0.378471866(ADJDE) - 0.374339975(EFG\_O) + 0.188999030(EFG\_D) + 0.246544169(TOR) - 0.200292523(TORD) - 0.122145170(ORB) + 0.091674199(DRB) + 0.008156337(FTR) - 0.005672370(FTRD) - 0.007435317(X2P\_O) + 0.018129422(X2P\_D) - 0.031198754(X3P\_O) + 0.018445979(X3P\_D) + 0.039424464(ADJ\_T) - 1.153099285(CONF\_STATUS: Mid\_Major) + 0.576406914(CONF\_STATUS: Power\_Conferences) + 20.757144208(WinPercentage) + 2.165754413(CONF\_STATUS: Mid\_Major\*WinPercentage) + 2.806063283(CONF\_STATUS: Power\_Conferences\*WinPercentage)

**Model Performance Metrics:**

* **R-squared:** 0.9775159
* **Adjusted R-squared:** 0.9769259
* **Mean Squared Error (MSE):** 1.17061
* **Root Mean Squared Error (RMSE):** 1.081948
* **Optimal Lambda (λ):** 0.003389493

Model 3:

WAB = -16.5278340 + 0.2812066(ADJOE) - 0.2915845(ADJDE) + 16.8228314(WinPercentage) - 0.1748484(CONF\_STATUS: Mid\_Major) + 1.3592401(CONF\_STATUS: Power\_Conferences) + 1.0439572(WinPercentage \* CONF\_STATUS: Mid\_Major) + 3.2964891(WinPercentage \* CONF\_STATUS: Power\_Conferences)

**Model Performance Metrics:**

* **R-squared:** 0.9693218
* **Adjusted R-squared:** 0.9685168
* **Mean Squared Error (MSE):** 1.494235
* **Root Mean Squared Error (RMSE):** 1.222389
* **Optimal Lambda (λ):** 0.01985233

**Model Comparison**

The saturated model is used as a benchmark (despite its issues with multicollinearity), alongside the naïve model, the best subset model, and the ridge and LASSO models with the highest adjusted R². However, when comparing models, it is important to recognize that R² is not always a reliable metric due to differences in the number of predictor variables across models. As a result, each model is evaluated using 10-fold cross-validation, with the primary focus placed on the Root Mean Squared Error (RMSE) as the key performance metric.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R²** | **Adjusted R²** | **R² (10-fold CV)** | **RMSE (10-fold CV)** |
| Saturated Model | 0.9765 | 0.9753 | 0.9742 | 1.089592 |
| Naive Model 1 | 0.9697 | 0.9691 | 0.9698 | 1.206808 |
| Best Subset | 0.9681 | 0.9676 | 0.9684 | 1.228271 |
| Best LASSO (Model 2) | 0.9775 | 0.9769 | 0.9755 | 1.066139 |
| Best Ridge (Model 3) | 0.9683 | 0.9675 | 0.9668 | 1.417429 |

**Table 7:** Model Comparison

All models perform similarly in terms of R² and adjusted R², ranging between 0.96 and 0.98. However, in terms of RMSE, the ridge regression model performs the worst, with an RMSE of 1.42 -- an unexpected outcome given ridge regression’s typical strength in handling multicollinearity. The naïve and best subset models fall in the middle, with RMSE values of 1.20 and 1.23, respectively. The LASSO and saturated models yield the lowest RMSEs at 1.06 and 1.08. Despite the strong performance of the saturated model, its complexity and high degree of multicollinearity limit its practical usefulness. As such, the LASSO model emerges as the best-performing model based on RMSE. However, a closer inspection of the LASSO model’s coefficients reveals some counterintuitive results. For instance, TOR (offensive turnover rate) has a positive coefficient of 0.247, implying that more turnovers lead to higher WAB -- an illogical conclusion. Similarly, EFG\_O (effective field goal percentage on offense) has a negative coefficient of -0.374, suggesting worse shooting is associated with better outcomes. These inconsistencies suggest that multicollinearity may still be influencing the model. Therefore, a simpler and more interpretable model -- such as the naïve model -- may offer more reliable insights.

**Conclusions**

Through both the graphs and the linear regression models, WAB is effectively explained by a team’s offensive and defensive statistics, along with their conference status and win percentage. Although many offensive and defensive statistics told similar stories, using just the primary offensive and defensive metrics (adjusted efficiencies) combined with conference status and win percentage explains approximately 97% of the variation in WAB.

For prediction purposes, the LASSO regression model offers the best balance between fit and generalization, making it the most suitable model for predicting WAB on new data. However, due to some counterintuitive coefficients present in the LASSO model, the naïve model is recommended for practical use. The naïve model can be interpreted as follows:

WAB = -15.85081 + 0.28483(ADJOE) - 0.29804(ADJDE) + 16.12351(WinPercentage) - 1.16389(CONF\_STATUS (Mid Major)) + 0.97667(CONF\_STATUS (Power Conferences)) + 2.73451((WinPercentage \* CONF\_STATUS: Mid Major)) + 3.93473((WinPercentage \* CONF\_STATUS: Power Conferences))

* **Intercept (-15.85081)**
  + The baseline WAB when all predictors are zero and conference status is low major.
* **ADJOE (0.28483)**
  + For each additional point scored per 100 possessions, on average, WAB increases by 0.28483, holding other variables constant.
* **ADJDE (-0.29804)**
  + For each additional point given up per 100 possessions, on average, WAB decreases by 0.29804, holding other variables constant.
* **WinPercentage (16.12351)**
  + For each one-unit (100%) increase in win percentage, WAB increases by 16.12351 for Low Major Conferences.
* **CONF\_STATUS: Mid Major (-1.16389)**
  + Being in a mid major conference decreases WAB by 1.16389 compared to the being in a low major conference -- assuming win percentage is zero.
* **CONF\_STATUS: Power Conferences (0.97667)**:
  + Being in a power conference increases WAB by 0.97667 compared to the being in a low major conference -- assuming win percentage is zero.
* **WinPercentage \* Mid Major (2.73451)**
  + The effect of win percentage on WAB is 2.73451 wins stronger for mid major teams compared to low major conference teams.
* **WinPercentage \* Power Conferences (3.93473)**
  + The effect of win percentage on WAB is 3.93473 wins stronger for power conference teams compared to low major conference teams.

**Sources of Error & Next Steps**

All analysis and models are trained using data from the 2019-20 season. To make the model more robust across different years, additional data from other seasons could be incorporated. This would help adjust for variations in playing style and other temporal effects. Not only would expanding the dataset across years add value, but including more granular data from the same year could also improve predictions. For example, a team’s strength of schedule (a metric which reflects the quality of their opponents) could be a strong predictor of WAB. Although strength of schedule may be correlated with conference status since higher-status conferences generally face tougher competition, including it explicitly could refine the model.

While scaling predictors is not always necessary in linear regression, it may enhance model performance. Separately, given the observed multicollinearity issues and the success of LASSO regression, exploring Elastic Net regression -- which combines the strengths of ridge and LASSO -- could further improve results. Additionally, although linear regression was a natural choice due to the apparent linear relationships in the data, alternative machine learning methods such as random forests or boosting algorithms may offer improved predictive accuracy. Further exploration of interaction terms could also be valuable; for example, investigating whether conference status itself predicts WAB directly, or if it primarily acts through influencing team statistics.

While this modeling effort successfully predicts WAB, it naturally leads to the question: who would have won the 2019-20 NCAA tournament had it taken place? To explore this, I created a tournament simulation and ran it 1,000 times. The results are available in the cbbSimulation zip file on [this GitHub repository](https://github.com/olinyoder2534/CollegeBasketballSimulator/tree/main).

Across the 1,000 simulations, one seeds won approximately 45% of the time (in the same simulation for 2024-25, one seeds won 55% of the time), which is slightly lower than the historical average of 63% since the tournament expanded to 64 teams in 1985. Among the one seeds, Kansas -- the overall number one seed -- won most frequently, with a 13.7% share of tournament victories. Conversely, two seeds won slightly more often in the simulation than historically observed, at 26% in the simulation compared to 13% historically.

As a college basketball fan, it was surprising to learn how frequently one seeds have won relative to two seeds historically. While the one and two seed winning percentages in the simulation differ somewhat from historical trends, it is worth noting that only 41 tournaments have been played since the expansion to 64 teams. It is entirely plausible that over the next several decades, the distribution of wins by seed could shift to more closely resemble the simulation outcomes.

Below are the distributions of winning teams and their seeds for the simulated 2019-20 tournament.

**Figure 18:** Distribution of March Madness Champions

Below is an example of a *single* simulation of the 2019-20 tournament in which, in a rare occurrence, a seven seed is the champion. For each round of the simulation, the matchups are presented alongside the projected win probabilities for each team, the game outcomes, and the teams advancing to the next round.

----------------------------------------------------------

Simulating Round 1...

Matchups for Round 1:

KANSAS (Seed: 1.0, Tournament Rank: 1.0) vs NORFOLK ST. (Seed: 16.0, Tournament Rank: 64.0)

GONZAGA (Seed: 1.0, Tournament Rank: 2.0) vs PRAIRIE VIEW A&M (Seed: 16.0, Tournament Rank: 63.0)

BAYLOR (Seed: 1.0, Tournament Rank: 3.0) vs ST. FRANCIS PA (Seed: 16.0, Tournament Rank: 62.0)

DAYTON (Seed: 1.0, Tournament Rank: 4.0) vs SIENA (Seed: 16.0, Tournament Rank: 61.0)

DUKE (Seed: 2.0, Tournament Rank: 5.0) vs WINTHROP (Seed: 15.0, Tournament Rank: 60.0)

SAN DIEGO ST. (Seed: 2.0, Tournament Rank: 6.0) vs WRIGHT ST. (Seed: 15.0, Tournament Rank: 59.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0) vs COLGATE (Seed: 15.0, Tournament Rank: 58.0)

OHIO ST. (Seed: 2.0, Tournament Rank: 8.0) vs NORTH DAKOTA ST. (Seed: 15.0, Tournament Rank: 57.0)

LOUISVILLE (Seed: 3.0, Tournament Rank: 9.0) vs BELMONT (Seed: 14.0, Tournament Rank: 56.0)

WEST VIRGINIA (Seed: 3.0, Tournament Rank: 10.0) vs HOFSTRA (Seed: 14.0, Tournament Rank: 55.0)

MARYLAND (Seed: 3.0, Tournament Rank: 11.0) vs UC IRVINE (Seed: 14.0, Tournament Rank: 54.0)

CREIGHTON (Seed: 3.0, Tournament Rank: 12.0) vs STEPHEN F. AUSTIN (Seed: 14.0, Tournament Rank: 53.0)

BYU (Seed: 4.0, Tournament Rank: 13.0) vs TEXAS ST. (Seed: 13.0, Tournament Rank: 52.0)

HOUSTON (Seed: 4.0, Tournament Rank: 14.0) vs NEW MEXICO ST. (Seed: 13.0, Tournament Rank: 51.0)

FLORIDA ST. (Seed: 4.0, Tournament Rank: 15.0) vs AKRON (Seed: 13.0, Tournament Rank: 50.0)

MICHIGAN (Seed: 4.0, Tournament Rank: 16.0) vs LIBERTY (Seed: 13.0, Tournament Rank: 49.0)

OREGON (Seed: 5.0, Tournament Rank: 17.0) vs VERMONT (Seed: 12.0, Tournament Rank: 48.0)

VILLANOVA (Seed: 5.0, Tournament Rank: 18.0) vs NORTHERN COLORADO (Seed: 12.0, Tournament Rank: 47.0)

ARIZONA (Seed: 5.0, Tournament Rank: 19.0) vs LOUISIANA TECH (Seed: 12.0, Tournament Rank: 46.0)

SETON HALL (Seed: 5.0, Tournament Rank: 20.0) vs RHODE ISLAND (Seed: 12.0, Tournament Rank: 45.0)

WISCONSIN (Seed: 6.0, Tournament Rank: 21.0) vs ARIZONA ST. (Seed: 11.0, Tournament Rank: 44.0)

IOWA (Seed: 6.0, Tournament Rank: 22.0) vs TEXAS (Seed: 11.0, Tournament Rank: 43.0)

BUTLER (Seed: 6.0, Tournament Rank: 23.0) vs YALE (Seed: 11.0, Tournament Rank: 42.0)

PENN ST. (Seed: 6.0, Tournament Rank: 24.0) vs EAST TENNESSEE ST. (Seed: 11.0, Tournament Rank: 41.0)

RUTGERS (Seed: 7.0, Tournament Rank: 25.0) vs USC (Seed: 10.0, Tournament Rank: 40.0)

KENTUCKY (Seed: 7.0, Tournament Rank: 26.0) vs NORTHERN IOWA (Seed: 10.0, Tournament Rank: 39.0)

ILLINOIS (Seed: 7.0, Tournament Rank: 27.0) vs RICHMOND (Seed: 10.0, Tournament Rank: 38.0)

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0) vs VIRGINIA (Seed: 10.0, Tournament Rank: 37.0)

AUBURN (Seed: 8.0, Tournament Rank: 29.0) vs UTAH ST. (Seed: 9.0, Tournament Rank: 36.0)

INDIANA (Seed: 8.0, Tournament Rank: 30.0) vs WICHITA ST. (Seed: 9.0, Tournament Rank: 35.0)

COLORADO (Seed: 8.0, Tournament Rank: 31.0) vs SAINT MARY'S (Seed: 9.0, Tournament Rank: 34.0)

OKLAHOMA (Seed: 8.0, Tournament Rank: 32.0) vs LSU (Seed: 9.0, Tournament Rank: 33.0)

Projected win probabilities for KANSAS vs NORFOLK ST.:

KANSAS: 0.99

NORFOLK ST.: 0.01

Projected win probabilities for GONZAGA vs PRAIRIE VIEW A&M:

GONZAGA: 0.98

PRAIRIE VIEW A&M: 0.02

Projected win probabilities for BAYLOR vs ST. FRANCIS PA:

BAYLOR: 0.98

ST. FRANCIS PA: 0.02

Projected win probabilities for DAYTON vs SIENA:

DAYTON: 0.98

SIENA: 0.02

Projected win probabilities for DUKE vs WINTHROP:

DUKE: 0.94

WINTHROP: 0.06

Projected win probabilities for SAN DIEGO ST. vs WRIGHT ST.:

SAN DIEGO ST.: 0.94

WRIGHT ST.: 0.06

Projected win probabilities for MICHIGAN ST. vs COLGATE:

MICHIGAN ST.: 0.94

COLGATE: 0.06

Projected win probabilities for OHIO ST. vs NORTH DAKOTA ST.:

OHIO ST.: 0.93

NORTH DAKOTA ST.: 0.07

Projected win probabilities for LOUISVILLE vs BELMONT:

LOUISVILLE: 0.89

BELMONT: 0.11

Projected win probabilities for WEST VIRGINIA vs HOFSTRA:

WEST VIRGINIA: 0.88

HOFSTRA: 0.12

Projected win probabilities for MARYLAND vs UC IRVINE:

MARYLAND: 0.89

UC IRVINE: 0.11

Projected win probabilities for CREIGHTON vs STEPHEN F. AUSTIN:

CREIGHTON: 0.88

STEPHEN F. AUSTIN: 0.12

Projected win probabilities for BYU vs TEXAS ST.:

BYU: 0.83

TEXAS ST.: 0.17

Projected win probabilities for HOUSTON vs NEW MEXICO ST.:

HOUSTON: 0.84

NEW MEXICO ST.: 0.16

Projected win probabilities for FLORIDA ST. vs AKRON:

FLORIDA ST.: 0.83

AKRON: 0.17

Projected win probabilities for MICHIGAN vs LIBERTY:

MICHIGAN: 0.83

LIBERTY: 0.17

Projected win probabilities for OREGON vs VERMONT:

OREGON: 0.79

VERMONT: 0.21

Projected win probabilities for VILLANOVA vs NORTHERN COLORADO:

VILLANOVA: 0.78

NORTHERN COLORADO: 0.22

Projected win probabilities for ARIZONA vs LOUISIANA TECH:

ARIZONA: 0.76

LOUISIANA TECH: 0.24

Projected win probabilities for SETON HALL vs RHODE ISLAND:

SETON HALL: 0.76

RHODE ISLAND: 0.24

Projected win probabilities for WISCONSIN vs ARIZONA ST.:

WISCONSIN: 0.72

ARIZONA ST.: 0.28

Projected win probabilities for IOWA vs TEXAS:

IOWA: 0.70

TEXAS: 0.30

Projected win probabilities for BUTLER vs YALE:

BUTLER: 0.70

YALE: 0.30

Projected win probabilities for PENN ST. vs EAST TENNESSEE ST.:

PENN ST.: 0.71

EAST TENNESSEE ST.: 0.29

Projected win probabilities for RUTGERS vs USC:

RUTGERS: 0.63

USC: 0.37

Projected win probabilities for KENTUCKY vs NORTHERN IOWA:

KENTUCKY: 0.66

NORTHERN IOWA: 0.34

Projected win probabilities for ILLINOIS vs RICHMOND:

ILLINOIS: 0.63

RICHMOND: 0.37

Projected win probabilities for MARQUETTE vs VIRGINIA:

MARQUETTE: 0.62

VIRGINIA: 0.38

Projected win probabilities for AUBURN vs UTAH ST.:

AUBURN: 0.55

UTAH ST.: 0.45

Projected win probabilities for INDIANA vs WICHITA ST.:

INDIANA: 0.54

WICHITA ST.: 0.46

Projected win probabilities for COLORADO vs SAINT MARY'S:

COLORADO: 0.54

SAINT MARY'S: 0.46

Projected win probabilities for OKLAHOMA vs LSU:

OKLAHOMA: 0.52

LSU: 0.48

Results for Round 1:

KANSAS 83 - NORFOLK ST. 53 | Winner: KANSAS

GONZAGA 90 - PRAIRIE VIEW A&M 55 | Winner: GONZAGA

BAYLOR 84 - ST. FRANCIS PA 62 | Winner: BAYLOR

DAYTON 87 - SIENA 48 | Winner: DAYTON

DUKE 81 - WINTHROP 93 | Winner: WINTHROP

SAN DIEGO ST. 83 - WRIGHT ST. 59 | Winner: SAN DIEGO ST.

MICHIGAN ST. 84 - COLGATE 78 | Winner: MICHIGAN ST.

OHIO ST. 73 - NORTH DAKOTA ST. 63 | Winner: OHIO ST.

LOUISVILLE 67 - BELMONT 66 | Winner: LOUISVILLE

WEST VIRGINIA 90 - HOFSTRA 51 | Winner: WEST VIRGINIA

MARYLAND 79 - UC IRVINE 52 | Winner: MARYLAND

CREIGHTON 84 - STEPHEN F. AUSTIN 62 | Winner: CREIGHTON

BYU 75 - TEXAS ST. 87 | Winner: TEXAS ST.

HOUSTON 66 - NEW MEXICO ST. 55 | Winner: HOUSTON

FLORIDA ST. 83 - AKRON 57 | Winner: FLORIDA ST.

MICHIGAN 88 - LIBERTY 56 | Winner: MICHIGAN

OREGON 72 - VERMONT 82 | Winner: VERMONT

VILLANOVA 68 - NORTHERN COLORADO 59 | Winner: VILLANOVA

ARIZONA 71 - LOUISIANA TECH 50 | Winner: ARIZONA

SETON HALL 73 - RHODE ISLAND 55 | Winner: SETON HALL

WISCONSIN 72 - ARIZONA ST. 80 | Winner: ARIZONA ST.

IOWA 66 - TEXAS 68 | Winner: TEXAS

BUTLER 69 - YALE 77 | Winner: YALE

PENN ST. 68 - EAST TENNESSEE ST. 65 | Winner: PENN ST.

RUTGERS 66 - USC 64 | Winner: RUTGERS

KENTUCKY 70 - NORTHERN IOWA 83 | Winner: NORTHERN IOWA

ILLINOIS 67 - RICHMOND 63 | Winner: ILLINOIS

MARQUETTE 67 - VIRGINIA 62 | Winner: MARQUETTE

AUBURN 74 - UTAH ST. 78 | Winner: UTAH ST.

INDIANA 60 - WICHITA ST. 79 | Winner: WICHITA ST.

COLORADO 70 - SAINT MARY'S 78 | Winner: SAINT MARY'S

OKLAHOMA 79 - LSU 74 | Winner: OKLAHOMA

Teams advancing to Round 2:

KANSAS (Seed: 1.0, Tournament Rank: 1.0)

GONZAGA (Seed: 1.0, Tournament Rank: 2.0)

BAYLOR (Seed: 1.0, Tournament Rank: 3.0)

DAYTON (Seed: 1.0, Tournament Rank: 4.0)

WINTHROP (Seed: 15.0, Tournament Rank: 60.0)

SAN DIEGO ST. (Seed: 2.0, Tournament Rank: 6.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0)

OHIO ST. (Seed: 2.0, Tournament Rank: 8.0)

LOUISVILLE (Seed: 3.0, Tournament Rank: 9.0)

WEST VIRGINIA (Seed: 3.0, Tournament Rank: 10.0)

MARYLAND (Seed: 3.0, Tournament Rank: 11.0)

CREIGHTON (Seed: 3.0, Tournament Rank: 12.0)

TEXAS ST. (Seed: 13.0, Tournament Rank: 52.0)

HOUSTON (Seed: 4.0, Tournament Rank: 14.0)

FLORIDA ST. (Seed: 4.0, Tournament Rank: 15.0)

MICHIGAN (Seed: 4.0, Tournament Rank: 16.0)

VERMONT (Seed: 12.0, Tournament Rank: 48.0)

VILLANOVA (Seed: 5.0, Tournament Rank: 18.0)

ARIZONA (Seed: 5.0, Tournament Rank: 19.0)

SETON HALL (Seed: 5.0, Tournament Rank: 20.0)

ARIZONA ST. (Seed: 11.0, Tournament Rank: 44.0)

TEXAS (Seed: 11.0, Tournament Rank: 43.0)

YALE (Seed: 11.0, Tournament Rank: 42.0)

PENN ST. (Seed: 6.0, Tournament Rank: 24.0)

RUTGERS (Seed: 7.0, Tournament Rank: 25.0)

NORTHERN IOWA (Seed: 10.0, Tournament Rank: 39.0)

ILLINOIS (Seed: 7.0, Tournament Rank: 27.0)

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

UTAH ST. (Seed: 9.0, Tournament Rank: 36.0)

WICHITA ST. (Seed: 9.0, Tournament Rank: 35.0)

SAINT MARY'S (Seed: 9.0, Tournament Rank: 34.0)

OKLAHOMA (Seed: 8.0, Tournament Rank: 32.0)

----------------------------------------------------------

Simulating Round 2...

Matchups for Round 2:

KANSAS (Seed: 1.0, Tournament Rank: 1.0) vs OKLAHOMA (Seed: 8.0, Tournament Rank: 32.0)

GONZAGA (Seed: 1.0, Tournament Rank: 2.0) vs SAINT MARY'S (Seed: 9.0, Tournament Rank: 34.0)

BAYLOR (Seed: 1.0, Tournament Rank: 3.0) vs WICHITA ST. (Seed: 9.0, Tournament Rank: 35.0)

DAYTON (Seed: 1.0, Tournament Rank: 4.0) vs UTAH ST. (Seed: 9.0, Tournament Rank: 36.0)

WINTHROP (Seed: 15.0, Tournament Rank: 60.0) vs MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

SAN DIEGO ST. (Seed: 2.0, Tournament Rank: 6.0) vs ILLINOIS (Seed: 7.0, Tournament Rank: 27.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0) vs NORTHERN IOWA (Seed: 10.0, Tournament Rank: 39.0)

OHIO ST. (Seed: 2.0, Tournament Rank: 8.0) vs RUTGERS (Seed: 7.0, Tournament Rank: 25.0)

LOUISVILLE (Seed: 3.0, Tournament Rank: 9.0) vs PENN ST. (Seed: 6.0, Tournament Rank: 24.0)

WEST VIRGINIA (Seed: 3.0, Tournament Rank: 10.0) vs YALE (Seed: 11.0, Tournament Rank: 42.0)

MARYLAND (Seed: 3.0, Tournament Rank: 11.0) vs TEXAS (Seed: 11.0, Tournament Rank: 43.0)

CREIGHTON (Seed: 3.0, Tournament Rank: 12.0) vs ARIZONA ST. (Seed: 11.0, Tournament Rank: 44.0)

TEXAS ST. (Seed: 13.0, Tournament Rank: 52.0) vs SETON HALL (Seed: 5.0, Tournament Rank: 20.0)

HOUSTON (Seed: 4.0, Tournament Rank: 14.0) vs ARIZONA (Seed: 5.0, Tournament Rank: 19.0)

FLORIDA ST. (Seed: 4.0, Tournament Rank: 15.0) vs VILLANOVA (Seed: 5.0, Tournament Rank: 18.0)

MICHIGAN (Seed: 4.0, Tournament Rank: 16.0) vs VERMONT (Seed: 12.0, Tournament Rank: 48.0)

Projected win probabilities for KANSAS vs OKLAHOMA:

KANSAS: 0.78

OKLAHOMA: 0.22

Projected win probabilities for GONZAGA vs SAINT MARY'S:

GONZAGA: 0.80

SAINT MARY'S: 0.20

Projected win probabilities for BAYLOR vs WICHITA ST.:

BAYLOR: 0.80

WICHITA ST.: 0.20

Projected win probabilities for DAYTON vs UTAH ST.:

DAYTON: 0.80

UTAH ST.: 0.20

Projected win probabilities for WINTHROP vs MARQUETTE:

WINTHROP: 0.25

MARQUETTE: 0.75

Projected win probabilities for SAN DIEGO ST. vs ILLINOIS:

SAN DIEGO ST.: 0.67

ILLINOIS: 0.33

Projected win probabilities for MICHIGAN ST. vs NORTHERN IOWA:

MICHIGAN ST.: 0.80

NORTHERN IOWA: 0.20

Projected win probabilities for OHIO ST. vs RUTGERS:

OHIO ST.: 0.70

RUTGERS: 0.30

Projected win probabilities for LOUISVILLE vs PENN ST.:

LOUISVILLE: 0.60

PENN ST.: 0.40

Projected win probabilities for WEST VIRGINIA vs YALE:

WEST VIRGINIA: 0.80

YALE: 0.20

Projected win probabilities for MARYLAND vs TEXAS:

MARYLAND: 0.80

TEXAS: 0.20

Projected win probabilities for CREIGHTON vs ARIZONA ST.:

CREIGHTON: 0.80

ARIZONA ST.: 0.20

Projected win probabilities for TEXAS ST. vs SETON HALL:

TEXAS ST.: 0.25

SETON HALL: 0.75

Projected win probabilities for HOUSTON vs ARIZONA:

HOUSTON: 0.55

ARIZONA: 0.45

Projected win probabilities for FLORIDA ST. vs VILLANOVA:

FLORIDA ST.: 0.53

VILLANOVA: 0.47

Projected win probabilities for MICHIGAN vs VERMONT:

MICHIGAN: 0.81

VERMONT: 0.19

Results for Round 2:

KANSAS 81 - OKLAHOMA 87 | Winner: OKLAHOMA

GONZAGA 86 - SAINT MARY'S 71 | Winner: GONZAGA

BAYLOR 73 - WICHITA ST. 77 | Winner: WICHITA ST.

DAYTON 85 - UTAH ST. 52 | Winner: DAYTON

WINTHROP 61 - MARQUETTE 82 | Winner: MARQUETTE

SAN DIEGO ST. 77 - ILLINOIS 62 | Winner: SAN DIEGO ST.

MICHIGAN ST. 79 - NORTHERN IOWA 59 | Winner: MICHIGAN ST.

OHIO ST. 66 - RUTGERS 51 | Winner: OHIO ST.

LOUISVILLE 87 - PENN ST. 77 | Winner: LOUISVILLE

WEST VIRGINIA 76 - YALE 61 | Winner: WEST VIRGINIA

MARYLAND 76 - TEXAS 52 | Winner: MARYLAND

CREIGHTON 80 - ARIZONA ST. 81 | Winner: ARIZONA ST.

TEXAS ST. 101 - SETON HALL 95 | Winner: TEXAS ST.

HOUSTON 88 - ARIZONA 73 | Winner: HOUSTON

FLORIDA ST. 84 - VILLANOVA 77 | Winner: FLORIDA ST.

MICHIGAN 71 - VERMONT 73 | Winner: VERMONT

Teams advancing to Round 3:

OKLAHOMA (Seed: 8.0, Tournament Rank: 32.0)

GONZAGA (Seed: 1.0, Tournament Rank: 2.0)

WICHITA ST. (Seed: 9.0, Tournament Rank: 35.0)

DAYTON (Seed: 1.0, Tournament Rank: 4.0)

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

SAN DIEGO ST. (Seed: 2.0, Tournament Rank: 6.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0)

OHIO ST. (Seed: 2.0, Tournament Rank: 8.0)

LOUISVILLE (Seed: 3.0, Tournament Rank: 9.0)

WEST VIRGINIA (Seed: 3.0, Tournament Rank: 10.0)

MARYLAND (Seed: 3.0, Tournament Rank: 11.0)

ARIZONA ST. (Seed: 11.0, Tournament Rank: 44.0)

TEXAS ST. (Seed: 13.0, Tournament Rank: 52.0)

HOUSTON (Seed: 4.0, Tournament Rank: 14.0)

FLORIDA ST. (Seed: 4.0, Tournament Rank: 15.0)

VERMONT (Seed: 12.0, Tournament Rank: 48.0)

----------------------------------------------------------

Simulating Round 3...

Matchups for Round 3:

OKLAHOMA (Seed: 8.0, Tournament Rank: 32.0) vs VERMONT (Seed: 12.0, Tournament Rank: 48.0)

GONZAGA (Seed: 1.0, Tournament Rank: 2.0) vs FLORIDA ST. (Seed: 4.0, Tournament Rank: 15.0)

WICHITA ST. (Seed: 9.0, Tournament Rank: 35.0) vs HOUSTON (Seed: 4.0, Tournament Rank: 14.0)

DAYTON (Seed: 1.0, Tournament Rank: 4.0) vs TEXAS ST. (Seed: 13.0, Tournament Rank: 52.0)

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0) vs ARIZONA ST. (Seed: 11.0, Tournament Rank: 44.0)

SAN DIEGO ST. (Seed: 2.0, Tournament Rank: 6.0) vs MARYLAND (Seed: 3.0, Tournament Rank: 11.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0) vs WEST VIRGINIA (Seed: 3.0, Tournament Rank: 10.0)

OHIO ST. (Seed: 2.0, Tournament Rank: 8.0) vs LOUISVILLE (Seed: 3.0, Tournament Rank: 9.0)

Projected win probabilities for OKLAHOMA vs VERMONT:

OKLAHOMA: 0.67

VERMONT: 0.33

Projected win probabilities for GONZAGA vs FLORIDA ST.:

GONZAGA: 0.64

FLORIDA ST.: 0.36

Projected win probabilities for WICHITA ST. vs HOUSTON:

WICHITA ST.: 0.35

HOUSTON: 0.65

Projected win probabilities for DAYTON vs TEXAS ST.:

DAYTON: 0.91

TEXAS ST.: 0.09

Projected win probabilities for MARQUETTE vs ARIZONA ST.:

MARQUETTE: 0.66

ARIZONA ST.: 0.34

Projected win probabilities for SAN DIEGO ST. vs MARYLAND:

SAN DIEGO ST.: 0.58

MARYLAND: 0.42

Projected win probabilities for MICHIGAN ST. vs WEST VIRGINIA:

MICHIGAN ST.: 0.56

WEST VIRGINIA: 0.44

Projected win probabilities for OHIO ST. vs LOUISVILLE:

OHIO ST.: 0.54

LOUISVILLE: 0.46

Results for Round 3:

OKLAHOMA 74 - VERMONT 88 | Winner: VERMONT

GONZAGA 86 - FLORIDA ST. 78 | Winner: GONZAGA

WICHITA ST. 73 - HOUSTON 66 | Winner: WICHITA ST.

DAYTON 80 - TEXAS ST. 68 | Winner: DAYTON

MARQUETTE 75 - ARIZONA ST. 73 | Winner: MARQUETTE

SAN DIEGO ST. 78 - MARYLAND 88 | Winner: MARYLAND

MICHIGAN ST. 77 - WEST VIRGINIA 69 | Winner: MICHIGAN ST.

OHIO ST. 69 - LOUISVILLE 81 | Winner: LOUISVILLE

Teams advancing to Round 4:

VERMONT (Seed: 12.0, Tournament Rank: 48.0)

GONZAGA (Seed: 1.0, Tournament Rank: 2.0)

WICHITA ST. (Seed: 9.0, Tournament Rank: 35.0)

DAYTON (Seed: 1.0, Tournament Rank: 4.0)

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

MARYLAND (Seed: 3.0, Tournament Rank: 11.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0)

LOUISVILLE (Seed: 3.0, Tournament Rank: 9.0)

----------------------------------------------------------

Simulating Round 4...

Matchups for Round 4:

VERMONT (Seed: 12.0, Tournament Rank: 48.0) vs LOUISVILLE (Seed: 3.0, Tournament Rank: 9.0)

GONZAGA (Seed: 1.0, Tournament Rank: 2.0) vs MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0)

WICHITA ST. (Seed: 9.0, Tournament Rank: 35.0) vs MARYLAND (Seed: 3.0, Tournament Rank: 11.0)

DAYTON (Seed: 1.0, Tournament Rank: 4.0) vs MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

Projected win probabilities for VERMONT vs LOUISVILLE:

VERMONT: 0.22

LOUISVILLE: 0.78

Projected win probabilities for GONZAGA vs MICHIGAN ST.:

GONZAGA: 0.56

MICHIGAN ST.: 0.44

Projected win probabilities for WICHITA ST. vs MARYLAND:

WICHITA ST.: 0.31

MARYLAND: 0.69

Projected win probabilities for DAYTON vs MARQUETTE:

DAYTON: 0.73

MARQUETTE: 0.27

Results for Round 4:

VERMONT 78 - LOUISVILLE 69 | Winner: VERMONT

GONZAGA 76 - MICHIGAN ST. 88 | Winner: MICHIGAN ST.

WICHITA ST. 60 - MARYLAND 65 | Winner: MARYLAND

DAYTON 72 - MARQUETTE 75 | Winner: MARQUETTE

Teams advancing to Round 5:

VERMONT (Seed: 12.0, Tournament Rank: 48.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0)

MARYLAND (Seed: 3.0, Tournament Rank: 11.0)

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

----------------------------------------------------------

Simulating Round 5...

Matchups for Round 5:

VERMONT (Seed: 12.0, Tournament Rank: 48.0) vs MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0) vs MARYLAND (Seed: 3.0, Tournament Rank: 11.0)

Projected win probabilities for VERMONT vs MARQUETTE:

VERMONT: 0.34

MARQUETTE: 0.66

Projected win probabilities for MICHIGAN ST. vs MARYLAND:

MICHIGAN ST.: 0.56

MARYLAND: 0.44

Results for Round 5:

VERMONT 68 - MARQUETTE 73 | Winner: MARQUETTE

MICHIGAN ST. 83 - MARYLAND 69 | Winner: MICHIGAN ST.

Teams advancing to Round 6:

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0)

----------------------------------------------------------

Simulating Round 6...

Matchups for Round 6:

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0) vs MICHIGAN ST. (Seed: 2.0, Tournament Rank: 7.0)

Projected win probabilities for MARQUETTE vs MICHIGAN ST.:

MARQUETTE: 0.33

MICHIGAN ST.: 0.67

Results for Round 6:

MARQUETTE 85 - MICHIGAN ST. 77 | Winner: MARQUETTE

Teams advancing to Round 7:

MARQUETTE (Seed: 7.0, Tournament Rank: 28.0)

The winner of the tournament is: MARQUETTE

**References**

Bobbitt, Zach. 2019. “How to Calculate Variance Inflation Factor (VIF) in R.” Statology. May 9, 2019. https://www.statology.org/variance-inflation-factor-r/.

Dutta, Sreejata. 2019. Review of Predicting Electrical Power Output in a Combined Cycle Power Plant.

Huang, Yibi. n.d. “STAT 224 Lecture 18 Ridge and Lasso Regressions.” https://www.stat.uchicago.edu/~yibi/teaching/stat224/L18.pdf.

“NCAA College Basketball RPI Rankings & Ratings 2025.” NCAA College Basketball RPI Rankings (updated today). Accessed May 1, 2025. https://www.teamrankings.com/ncaa-basketball/rpi-ranking/rpi-rating-by-conf?date=2020-03-12.

“NCAA Tournament Championships by Seed,” n.d. https://www.printyourbrackets.com/ncaa-tournament-wins-by-seed.html.

“T-Rank  - Customizable College Basketball Tempo Free Stats - T-Rank,” n.d. https://barttorvik.com/trank.php?year=2020&sort=&conlimit=#.

“Understanding Lasso and Ridge Regression | R-Bloggers.” 2020. R Bloggers. June 16, 2020. https://www.r-bloggers.com/2020/06/understanding-lasso-and-ridge-regression/.

**Appendix (R code)**

library(tidyverse)

library(MASS)

library(GGally)

library(caret)

library(glmnet)

library(car)

#library(keras)

df <- read.csv(“cbb20.csv")

#kaggle link

#https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset?resource=download&select=cbb20.csv

head(df)

#SUMMARY

head(df, 25)

tail(df)

summary(df)

dim(df)

length(unique(df$CONF)) #32 conference champs -- 32 at large bids (not doing first four)

unique(df$CONF)

table(df$CONF)

sum(is.na(df)) #no NAs

head(df[order(df$WAB, decreasing = FALSE), ], 25)

subset(df, WAB == 0)

#PREPROCESSING

#creating CONF\_STATUS (high, mid, low major)

#High major = power 6, mid major = 10 highest rated conferences according to link below, all else = low major

#https://www.teamrankings.com/ncaa-basketball/rpi-ranking/rpi-rating-by-conf?date=2020-03-12

df$CONF\_STATUS <- ifelse(df$CONF %in% c("B12", "B10", "P12", "SEC", "ACC", "BE"), "Power Conferences",

ifelse(df$CONF %in% c("AMER", "A10", "WCC", "MWC", "MVC", "MAC", "SC", "SB", "Ivy", "CAA"), "Mid Major",

"Low Major"))

#write.csv(df, file = "cbb1.csv", row.names = FALSE)

#creating a win% column (for the simulation, I do this separately in Python)

df$WinPercentage = df$W/df$G

df[order(-df$WinPercentage), ]

#PLOTS

ggplot(df, aes(x = WAB)) +

geom\_histogram(binwidth = 5, fill = "white", color = "black") +

labs(

title = "Histogram of WAB",

x = "Wins Above the Bubble (WAB)",

y = "Count"

) +

scale\_x\_continuous(breaks = seq(0, max(df$WAB), by = 20)) +

theme(

plot.title = element\_text(hjust = 0.5, size = 16, face = "bold"),

axis.title.x = element\_text(size = 12),

axis.title.y = element\_text(size = 12),

axis.text = element\_text(size = 10),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank()

)

#ADJOE

ggplot(df, aes(x = ADJOE, y = WAB)) +

geom\_point() +

labs(

title = "WAB vs ADJOE",

x = "Adjusted Offensive Efficiency (ADJOE)",

y = "WAB"

)

cor(df$ADJOE, df$WAB)

#ADJDE

ggplot(df, aes(x = ADJDE, y = WAB)) +

geom\_point() +

labs(

title = "WAB vs ADJDE",

x = "Adjusted Defensive Efficiency (ADJDE)",

y = "WAB"

)

cor(df$ADJDE, df$WAB)

ggplot(df, aes(x = CONF\_STATUS, y = WAB))+

geom\_boxplot()+ # would assume that conference status is a strong predictor of WAB from the graph, or is it just that teams have better statistics in Power Conferences?

labs(

title = "WAB vs Conference Status",

x = "Conference Status",

y = "WAB"

)

anova\_score <- aov(WAB ~ CONF\_STATUS, data = df)

summary(anova\_score)

#ADJT

ggplot(df, aes(x = ADJ\_T, y = WAB)) +

geom\_point() +

labs(

title = "WAB vs ADJT",

x = "Adjusted Tempo (ADJT)",

y = "WAB"

)

cor(df$ADJ\_T, df$WAB)

#Win %

ggplot(df, aes(x = WinPercentage, y = WAB)) +

geom\_point() +

labs(

title = "WAB vs Win Percentage",

x = "Win Percentage",

y = "WAB"

)

cor(df$WinPercentage, df$WAB)

#by conf\_status

ggplot(df, aes(x = WinPercentage, y = WAB, color = CONF\_STATUS)) +

geom\_point() +

labs(

title = "WAB vs Win Percentage by Conf Status",

x = "Win Percentage",

y = "WAB",

color = "Conference Status"

) +

theme\_minimal() +

theme(

plot.title = element\_text(hjust = 0.5, size = 16, face = "bold"),

axis.title.x = element\_text(size = 12),

axis.title.y = element\_text(size = 12),

legend.title = element\_text(size = 12),

legend.text = element\_text(size = 10)

)

ggpairs(df[, !(names(df) %in% c("TEAM", "CONF", "RK", "G", "W", "BARTHAG"))])

#MODELING

#excluding TEAM (same as ID) and all non-offensive/defensive statistics (except Win Percentage as well) as calculated values from Barttorvik (RK and BARTHAG)

##########################

#Model 1 (saturated model)

lm1 <- lm(WAB ~ . - TEAM - RK - W - G - CONF - BARTHAG, df)

summary(lm1)

vif(lm1)

#diagnostic plots

#par(mfrow = c(2, 2))

plot(lm1)

df[c(353, 311, 134),] #points of interest from the plots

#Cook's Distance

which(abs(cooks.distance(lm1)) > .5) #no points with a Cook's Distance > .5

cd\_lm1 <- cooks.distance(lm1)

cd\_lm1[c(353, 311, 134)]

k <- length(coefficients(lm1)) - 1

n <- nrow(model.frame(lm1))

numerator\_df <- k

denominator\_df <- n - k

q <- pf(cooks.distance(lm1),numerator\_df,denominator\_df)

which(q>.1)

which(q>.2)

ols\_plot\_cooksd\_bar(lm1)

#hat values

which(hatvalues(lm1) > .5)

#dffits

which(dffits(lm1) > 1)

dffits\_lm1 <- dffits(lm1)

dffits\_lm1[353]

#dfbetas

which(dfbetas(lm1) > 1)

n <- nrow(df)

thresh <- 2/sqrt(n)

which(dfbetas(lm1) > thresh)

#outliers

which(abs(rstudent(lm1)) > 3) #point 353 (Chicago St. worst team statistically) has a abs studentized resid > 3, going to remove it

studentized\_res\_lm1 <- rstudent(lm1)

studentized\_res\_lm1[c(353)]

plot(resid(lm1),

main = "Residual Plot for Saturated Model",

xlab = "Observation Index",

ylab = "Residuals",

ylim = c(-3.5, 3.5))

points(c(353), resid(lm1)[c(353)], col = "red", pch = 19)

text(c(353), resid(lm1)[c(353)], labels = c(353), pos = 3, col = "red")

##########################

#Model 2

lm2 <- lm(WAB ~ ADJOE + ADJDE + (WinPercentage \* CONF\_STATUS), df)

summary(lm2)

vif(lm2, type=c("predictor"))

#diagnostic plots

plot(lm2) #a few potential outliers, but nothing too egregious

boxplot(resid(lm2))

#Cook's Distance

which(abs(cooks.distance(lm2)) > .5) #no points with a Cook's Distance > .5

cd\_lm1 <- cooks.distance(lm2)

cd\_lm1[c(353, 311, 134)]

k <- length(coefficients(lm2)) - 1

n <- nrow(model.frame(lm2))

numerator\_df <- k

denominator\_df <- n - k

q <- pf(cooks.distance(lm2),numerator\_df,denominator\_df)

which(q>.1)

which(q>.2)

ols\_plot\_cooksd\_bar(lm2)

#hat values

which(hatvalues(lm2) > .5)

#dffits

which(dffits(lm2) > 1)

dffits\_lm1 <- dffits(lm2)

dffits\_lm1[353]

#dfbetas

which(dfbetas(lm2) > 1)

n <- nrow(df)

thresh <- 2/sqrt(n)

which(dfbetas(lm2) > thresh, arr.ind = TRUE)

#outliers

which(abs(rstudent(lm2)) > 3) #point 353 (Chicago St. worst team statistically) has a abs studentized resid > 3, going to remove it

studentized\_res\_lm1 <- rstudent(lm2)

studentized\_res\_lm1[c(311, 352, 353)]

plot(resid(lm2),

main = "Residual Plot for Saturated Model",

xlab = "Observation Index",

ylab = "Residuals",

ylim = c(-5.5, 5.5))

points(c(311, 352, 353), resid(lm2)[c(311, 352, 353)], col = "red", pch = 19)

text(c(311, 352, 353), resid(lm2)[c(311, 352, 353)], labels = c(311, 352, 353), pos = 3, col = "red")

##########################

#Model 2.1 (naive model, w/o outliers)

df.1 <- df[-c(353),]

lm3 <- lm(WAB ~ ADJOE + ADJDE + WinPercentage \* CONF\_STATUS, df)

summary(lm3)

vif(lm3, type=c("predictor"))

#diagnostic plots

plot(lm3) #a few potential outliers, but nothing too egregious

boxplot(resid(lm3))

#Cook's Distance

which(abs(cooks.distance(lm3)) > .5) #no points with a Cook's Distance > .5

cd\_lm1 <- cooks.distance(lm3)

k <- length(coefficients(lm3)) - 1

n <- nrow(model.frame(lm3))

numerator\_df <- k

denominator\_df <- n - k

q <- pf(cooks.distance(lm3),numerator\_df,denominator\_df)

which(q>.1)

which(q>.2)

ols\_plot\_cooksd\_bar(lm3)

#hat values

which(hatvalues(lm3) > .5)

#dffits

which(dffits(lm3) > 1)

dffits\_lm1 <- dffits(lm3)

dffits\_lm1[352]

#dfbetas

which(dfbetas(lm3) > 1)

n <- nrow(df)

thresh <- 2/sqrt(n)

which(dfbetas(lm3) > thresh)

#outliers

which(abs(rstudent(lm3)) > 3) #point 353 (Chicago St. worst team statistically) has a abs studentized resid > 3, going to remove it

studentized\_res\_lm1 <- rstudent(lm3)

studentized\_res\_lm1[c(121, 311, 352)]

plot(resid(lm3),

main = "Residual Plot for Saturated Model",

xlab = "Observation Index",

ylab = "Residuals",

ylim = c(-5.5, 5.5))

points(c(121, 311, 352), resid(lm2)[c(121, 311, 352)], col = "red", pch = 19)

text(c(121, 311, 352), resid(lm2)[c(121, 311, 352)], labels = c(121, 311, 352), pos = 3, col = "red")

##########################

#Model 3 (Using best subset)

#Only adjusted offensive/defense, tempo, conference, & win % (Strength of Schedule would be a good variable in this model, too)

lm4 <- regsubsets(WAB ~. - TEAM - RK - W - G - CONF - BARTHAG, df, nvmax = NULL)

slm4 <-summary(lm4)

slm4$adjr2

slm4$which

adjr2\_vals <- slm4$adjr2

cp\_vals <- slm4$cp

# Create the plot

plot(adjr2\_vals, type = "b", pch = 19, col = "black",

xlab = "Number of Predictors", ylab = "Adjusted R¬≤",

main = "Adjusted R¬≤ vs Number of Predictors")

plot(cp\_vals, type = "b", pch = 19, col = "black",

xlab = "Number of Predictors", ylab = "Mallows' Cp",

main = "Mallows' Cp vs Number of Predictors")

lm4 <- lm(WAB ~ ADJOE + ADJDE + WinPercentage + CONF\_STATUS, df)

summary(lm4)

vif(lm4)

#diagnostic plots

plot(lm4) #a few potential outliers, but nothing too egregious

boxplot(resid(lm4))

#Cook's Distance

which(abs(cooks.distance(lm4)) > .5) #no points with a Cook's Distance > .5

cd\_lm1 <- cooks.distance(lm4)

k <- length(coefficients(lm4)) - 1

n <- nrow(model.frame(lm4))

numerator\_df <- k

denominator\_df <- n - k

q <- pf(cooks.distance(lm4),numerator\_df,denominator\_df)

which(q>.1)

which(q>.2)

ols\_plot\_cooksd\_bar(lm4)

#hat values

which(hatvalues(lm4) > .5)

#dffits

which(dffits(lm4) > 1)

dffits\_lm1 <- dffits(lm4)

dffits\_lm1[352]

#dfbetas

which(dfbetas(lm4) > 1)

n <- nrow(df)

thresh <- 2/sqrt(n)

which(dfbetas(lm4) > thresh)

#outliers

which(abs(rstudent(lm4)) > 3) #point 353 (Chicago St. worst team statistically) has a abs studentized resid > 3, going to remove it

studentized\_res\_lm1 <- rstudent(lm4)

studentized\_res\_lm1[c(121, 311, 352)]

plot(resid(lm4),

main = "Residual Plot for Saturated Model",

xlab = "Observation Index",

ylab = "Residuals",

ylim = c(-5.5, 5.5))

points(c(121, 311, 352), resid(lm2)[c(121, 311, 352)], col = "red", pch = 19)

text(c(121, 311, 352), resid(lm2)[c(121, 311, 352)], labels = c(121, 311, 352), pos = 3, col = "red")

##########################

#TRAINING ALL MODELS ON DATA WITHOUT POINT 353, SINCE THEY ARE SO FAR OFF THE BUBBLE, I AM COMFORTABLE REMOVING THEM, WE ARE REALLY INTERESTING ON APPLYING THIS TO TEAMS ON THE BUBBLE

## COMPARE ALL MODELS USING RMSE FROM 10 FOLD CV: SATURATED, NAIVE (MULTIPLE JUST TO SHOW ANOVA), STEPWISE, LASSO

train\_control <- trainControl(method = "cv", number = 10)

#lm1

lm1\_cv <- train(WAB ~ . - TEAM - RK - W - G - CONF - BARTHAG, df, method = "lm", trControl = train\_control)

print(lm1\_cv)

kansas\_row <- df[1,]

predict(lm1, newdata = kansas\_row)

#lm2

lm2\_cv <- train(WAB ~ ADJOE + ADJDE + WinPercentage \* CONF\_STATUS, df, method = "lm", trControl = train\_control)

print(lm2\_cv)

predict(lm2, newdata = kansas\_row)

#lm3

lm3\_cv <- train(WAB ~ ADJOE + ADJDE + WinPercentage + CONF\_STATUS, df, method = "lm", trControl = train\_control)

print(lm3\_cv)

predict(lm3, newdata = kansas\_row)

#lasso

y <- df$WAB

#x <- model.matrix(WAB ~ . - TEAM - RK - W - G - CONF - BARTHAG, data = df)[, -1]

x <- model.matrix(WAB ~ . - TEAM - RK - W - G - CONF - BARTHAG - CONF\_STATUS - WinPercentage + (CONF\_STATUS \* WinPercentage), data = df)[, -1]

#x <- model.matrix(WAB ~ ADJOE + ADJDE + (WinPercentage \* CONF\_STATUS), data = df)[, -1]

set.seed(123)

cv\_model\_lasso <- cv.glmnet(x, y,

alpha = 1,

nfolds = 10,

standardize = TRUE)

#plot(cv\_model\_lasso)

lambda\_min\_lasso <- cv\_model\_lasso$lambda.min

#lambda\_min\_lasso

lambda\_1se\_lasso <- cv\_model\_lasso$lambda.1se

#lambda\_1se\_lasso

best\_model\_lasso <- glmnet(x, y,

alpha = 1,

lambda = lambda\_min\_lasso,

standardize = TRUE)

#coef(best\_model\_lasso)

#rmse

mse\_lasso <- mean(cv\_model\_lasso$cvm[cv\_model\_lasso$lambda == lambda\_min\_lasso])

rmse\_lasso <- sqrt(mse\_lasso)

rmse\_lasso

#calc r^2

y\_pred\_lasso <- predict(best\_model\_lasso, newx = x)

r\_squared\_lasso <- 1 - (sum((y - y\_pred\_lasso)^2) / sum((y - mean(y))^2))

#adj r^2

adjusted\_r\_squared\_lasso <- 1 - (1 - r\_squared\_lasso) \* ((n - 1) / (n - p - 1))

summary\_info\_lasso <- list(

r\_squared = r\_squared\_lasso,

adjusted\_r\_squared = adjusted\_r\_squared\_lasso,

mse = mse\_lasso,

rmse = rmse\_lasso,

lambda = lambda\_min\_lasso,

coefficients = coef(best\_model\_lasso)

)

print(summary\_info\_lasso)

#10-fold cv

cv\_lasso <- train(x = x, y = y,

method = "glmnet",

trControl = train\_control,

tuneGrid = expand.grid(alpha = 1, lambda = lambda\_min\_lasso))

print(cv\_lasso)

#ridge

cv\_model\_ridge <- cv.glmnet(x, y,

alpha = 0,

nfolds = 10,

standardize = TRUE)

#plot(cv\_model\_ridge)

lambda\_min\_ridge <- cv\_model\_ridge$lambda.min

#lambda\_min\_ridge

lambda\_1se\_ridge <- cv\_model\_ridge$lambda.1se

#lambda\_1se\_ridge

best\_model\_ridge <- glmnet(x, y,

alpha = 0,

lambda = lambda\_min\_ridge,

standardize = TRUE)

#coef(best\_model\_ridge)

#rmse

mse\_ridge <- mean(cv\_model\_ridge$cvm[cv\_model\_ridge$lambda == lambda\_min\_ridge])

rmse\_ridge <- sqrt(mse\_ridge)

rmse\_ridge

#calc r^2

y\_pred\_ridge <- predict(best\_model\_ridge, newx = x)

r\_squared\_ridge <- 1 - (sum((y - y\_pred\_ridge)^2) / sum((y - mean(y))^2))

#adj r^2

adjusted\_r\_squared\_ridge <- 1 - (1 - r\_squared\_ridge) \* ((n - 1) / (n - p - 1))

summary\_info\_ridge <- list(

r\_squared = r\_squared\_ridge,

adjusted\_r\_squared = adjusted\_r\_squared\_ridge,

mse = mse\_ridge,

rmse = rmse\_ridge,

lambda = lambda\_min\_ridge,

coefficients = coef(best\_model\_ridge)

)

print(summary\_info\_ridge)

#10-fold cv

cv\_ridge <- train(x = x, y = y,

method = "glmnet",

trControl = train\_control,

tuneGrid = expand.grid(alpha = 1, lambda = lambda\_min\_ridge))

print(cv\_ridge)