# Predicting March Madness

## Using Parametric & Non-Parametric Approaches to Predict Points Scored

Raymond Romaniuk (6047088)

John Lecomte (4999157)

Brock University

MATH 4P82

Due: April 11, 2021

## Table of Contents

## Abstract

With the daily influx of massive amounts of data worldwide having a strong understanding of how to acquire, transform and work with data is imperative to having success in data driven fields like predictive analysis. This project allows us to familiarize ourselves with parametric and non-parametric approaches and utilize these on real world data, which will help us to develop the skills necessary to work with big data in the future.

This project utilizes both parametric and non-parametric regression techniques to create prediction models. Parametric models include least absolute shrinkage and selection operator (LASSO) regression, multiple linear regression, and ridge regression. Cubic B-spline basis functions are used to create our non-parametric regressions, while generalized cross validation and ordinary cross validation are used to select our tuning parameters.

We found that the optimal model to create predictions of March Madness results is dependent on the intended use of the predictions. The optimal model differs for an individual interested in filling out a bracket who only needs which team will win, and an individual who wishes to bet on a game based upon the amount of points each team will score.



### Key words

March Madness, College Basketball, Parametric, Non-Parametric, Multiple Linear Regression, Ridge Regression, LASSO Regression, Points Scored, B-Spline Basis Functions, Cross Validation

## Introduction

The National Basketball Association (NBA) is regarded as one of the four major sports leagues in North America, alongside the National Football League (NFL), National Hockey League (NHL) and Major League Baseball (MLB). Basketball is a team sport played between two teams with five players from each team on the playing surface, known as the court, at any given time. The objective of the game is to shoot the basketball through the opposing teams hoop and accumulate more points than your opponent. Points can be scored in two separate categories, field goals and free throws. A field goal is scored during live play and can be worth either two or three points, called a two or three pointer. The amount of points a field goal is worth is dictated by the position on the court the shooting player shot the ball from. If it was from behind the three-point line (see Figure 1) the shot is worth three points and if it is from inside the three-point line the shot is only worth two points. The second method of scoring is by free throws, these are worth one point each. A free throw occurs when a player commits an infraction, known as a foul, on the opposition and their opponent is awarded a given number of free throws based on the severity of the foul. Free throws take place from the free throw line with no infringement allowed by the opposition. The NBA being a significant focal point in North American sports leads fans to gravitate not just to the NBA, but also to National Collegiate Athletic Association Division I Basketball, NCAA Basketball or College Basketball for short, where the majority of future NBA players come from.

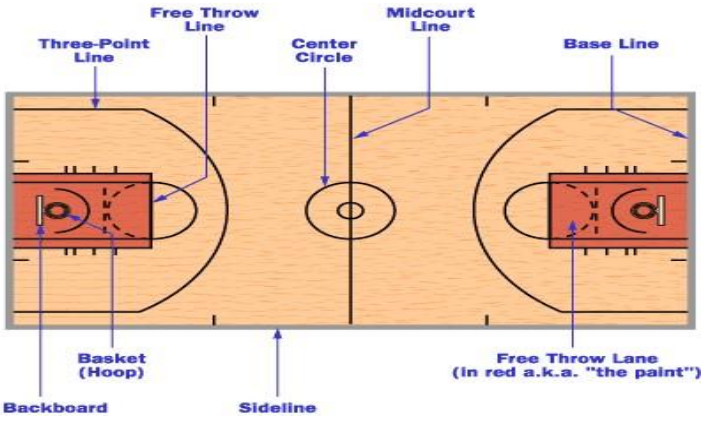


*Figure 1: Diagram of important basketball court locations*

This project aims to explore College Basketball, specifically the NCAA Division I Men's Basketball Tournament, known as March Madness. The March Madness Tournament is the culmination of each College Basketball season and brings 68 of the top teams in the country together to form a bracketed tournament and compete for the distinction of being the best team in Division I College Basketball. Teams gain entry to the tournament through two avenues, automatic bids and at large bids. Of the 68 teams, 32 of them receive automatic bids into the tournament. To receive an automatic bid a team must win their respective conference's playoff tournament. There are 32 conferences, so these 32 teams are the 32 conference champions. The 36 at large bids are comprised of teams who did not win their conference's playoff tournament. These 36 teams are chosen by the Selection Committee who selects the 36 teams they believe are

most deserving of competing in the tournament. Teams are then seeded from 1 to 68 and split into four different regions, East, West, South and Midwest. There are 16 teams allocated to each region and seeded from 1 to 16 within the region. The teams are then bracketed by their seed and the opening round consists of games with the first seed playing the sixteenth seed, second playing fifteenth and so on. Noticeably 16 teams per region does not divide the total 68 teams evenly. An eight-team play-in round, called the First Four, is contested prior to the first round between the four lowest ranked automatic bid teams and the four lowest ranked at large bid teams. The lowest ranked automatic bid teams are usually seeded lower, overall, than the lowest ranked at large bid teams since they are representing weaker conferences. The four automatic bid teams thus play each other for one of two available 16 seed spots, whereas the four at large bid teams play each other for one of two available 11 seed spots. March Madness consists of six rounds and a total of 67 games.


*Figure 2: 2019 March Madness Tournament Bracket*

The data for our project comes from four different sources: ESPN, Fox Sports, the NCAA and the Pomeroy College Basketball Ratings. The data was acquired using the Beautiful Soup package, in Python, to web scrape pertinent information from the four data sources (this was a previous project done outside of school).

After much data cleaning and preparation to align the data in an optimal format for analysis, we have a total of 124 variables. The data gathered corresponds to the 2017-18, 2018-19 and 2020-21 seasons. For our project we will use the 2017-18 and 2018-19 data, for which we have tournament results, as the training data and the 2020-21 data as our testing data.

The dataset that we have includes the following variables: AP (Associated Press) Poll rankings (pre-season and final), win-loss totals (overall, in conference, home, away, neutral and against top 25 opponents), points for/against (overall and in conference), three point shooting totals, field goal shooting totals, free throw shooting totals, strength of schedule (overall and non-conference), along with a variety of analytics from Ken Pomeroy including, offensive and defensive efficiency, adjusted tempo and luck rating. We also have the points scored in the past

March Madness games as our dependent variable. All inputs are available for each team playing in a game.

The dataset is constructed such that each observation contains data regarding both teams in each game. Table 1 depicts a sample of how the dataset is structured for a single game.

| Team | Wins | Loses | AP Rank | Opponent | Opp Wins | Opp Loses | Opp AP Rank |
|---|---|---|---|---|---|---|---|
| Gonzaga | 26 | 0 | 1 | Baylor | 22 | 2 | 3 |
| Baylor | 22 | 2 | 3 | Gonzaga | 26 | 0 | 1 |

*Table 1: Sample of the dataset structure.*

March Madness consists of 67 games each year, four games in the "First Four" play-in round, 15 games in each of the four regions and three games in the "Final Four". This means we will have 134 observations for each tournament, thus in total, for two tournaments, we have 268 observations in our training dataset.

Our training data, however, is not yet ready for analysis since there are blanks in the AP Poll (pre-season and final) columns, since only the top 25 teams receive a ranking. To combat this, we filled the missing values with a value of 100, to distinguish between the ranked teams and the unranked teams.

Along with the need to fill in the AP Poll columns, our testing data suffers from a COVID-19 related issue. When creating our models with our training data the data is standardized in terms of that same training data. This is no problem since in those two seasons nothing out of the ordinary occurred.

The issue is that this season, due to COVID-19, teams did not play a similar number of games to previous seasons. This is important because due to this lack of games, when the testing data is standardized in terms of the training data, the teams in our testing dataset all perform below average compared to previous seasons. Variables like points for, wins and field goal attempts are examples of what are affected by this lack of games played.

To account for this, we calculated the average number of games played in the previous season (33.14). Having this average, we updated the testing dataset and adjusted each teams' statistics to be in terms of that average number of games played. It is unfortunate that each team in the testing data now has the same number of games played, but since games played is not one of the selected variables, that we find later, this seems like a reasonable solution.

Performing this adjustment allowed us to account for a team like Colgate entering the tournament with only 15 games played and facing Arkansas who had played 28 games. Both teams are very high scoring, but with Colgate only playing half as many games as Arkansas the number of points that Colgate could potentially score would be significantly underestimated.

Ultimately, the goal of the project is to use data from past March Madness tournament's and determine which model will most accurately predict how many points a team will score against a given opponent and eventually determine who is most likely to win each game.

## Method

To assist us in our quest to accurately predict March Madness we will enlist the help of both parametric and non-parametric regression methods to determine whether one of them performs better than the other.

The parametric approaches we will use include least absolute shrinkage and selection (LASSO) regression, multiple linear regression, and ridge regression.

As mentioned in the introduction, our dataset consists of 124 variables. This is far too many for us to explore individually when we reach the non-parametric methods, so we will first use LASSO regression as a variable reduction technique.

LASSO regression is similar to a simple linear regression with the addition of a penalty term that penalizes the model based on the absolute value of the model's coefficients. The LASSO regression model can be written as follows,

$$y_i = \boldsymbol{\beta} \cdot \boldsymbol{x_i} + \lambda \cdot \sum_{j=0}^{v} |\beta_j| + \varepsilon_i$$

where $y_i$ is the $i^{th}$ observation of points scored, $\boldsymbol{\beta}$ is a vector of the estimated coefficients of each independent variable, $\boldsymbol{x_i}$ is a vector of the observed values of the independent variables for the $i^{th}$ observation, $\lambda$ is the chosen constant of our penalty term (if $\lambda = 0$ we are performing a simple linear regression), $\sum_{j=0}^{v}|\beta_j|$ is the sum from 0 to *v*, where *v* is the number of independent variables in our model, of our absolute $\beta$ values and $\varepsilon_i$ is the random error of the $i^{th}$ observation.

In LASSO regression $\beta$ values can be zero. This is dependent on the $\lambda$ in the penalty term and the greater it is the more coefficients will be zero since the model penalizes based on the size of the coefficients.

Our next parametric approach is multiple linear regression. Multiple linear regression is the same as simple linear regression with the lone difference being that we have more than one independent variable. Since we wish to predict points scored, we choose linear regression instead of logistic regression which would be the optimal choice if we were interested in predicting probabilities (e.g. a teams probability of winning the game or the probability a team scores more than 77 points). The multiple linear regression model can be written as follows,

$$y_i = \boldsymbol{\beta} \cdot \boldsymbol{x_i} + \varepsilon_i$$

where $y_i$, $\boldsymbol{\beta}$, $\boldsymbol{x_i}$ and $\varepsilon_i$ are the same variables as the LASSO regression.

We will later use this multiple linear regression approach to create a model with only the variables selected by the LASSO regression.

Our final parametric approach of interest is ridge regression. Ridge regression is very similar to LASSO regression and can be written as follows,

$$y_i = \boldsymbol{\beta} \cdot \boldsymbol{x_i} + \lambda \cdot \sum_{j=0}^{v} \beta_j^2 + \varepsilon_i$$

where the only difference between the two is in the penalty term and all variables are the same as previous models. In the case of ridge regression, we see that the penalty term now contains the sum of the *squared* coefficients rather than the sum of the *absolute* coefficients. This small difference does not allow coefficients to go to zero. Figure 3 illustrates the difference in the two models graphically and we can see that since the absolute value of $\beta$ is not differentiable at all points the LASSO residual sum of squares touches a corner in the figure.

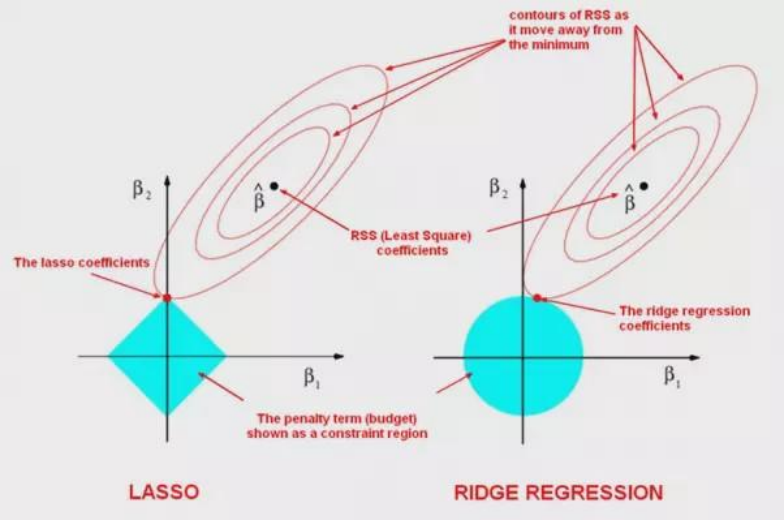


*Figure 3: Depiction of graphical difference between LASSO and ridge regression when estimating coefficients to minimize the penalty.*

Next, we will introduce our non-parametric regression approaches. The non-parametric approaches we will use are cubic B-spline basis functions, with 51 equally spaced knots, and we will select the tuning parameter using generalized cross validation (GCV) and ordinary cross validation (OCV).

B-spline basis functions can be described as a group of piecewise polynomial functions over a specified range and allow the model to be more flexible and better follow trends in the data. Figure 4 is an example of what the B-spline basis functions look like when there are six equally spaced knots. Evidently four of these six knots are interior knots and since the order of our basis functions are M = 4, we can determine the number of basis functions as follows,

$$\#\ of\ basis\ functions = (total\ knots - 1) + (M - 1)$$

$$= (6 - 1) + (4 - 1) = 5 + 3 = 8$$

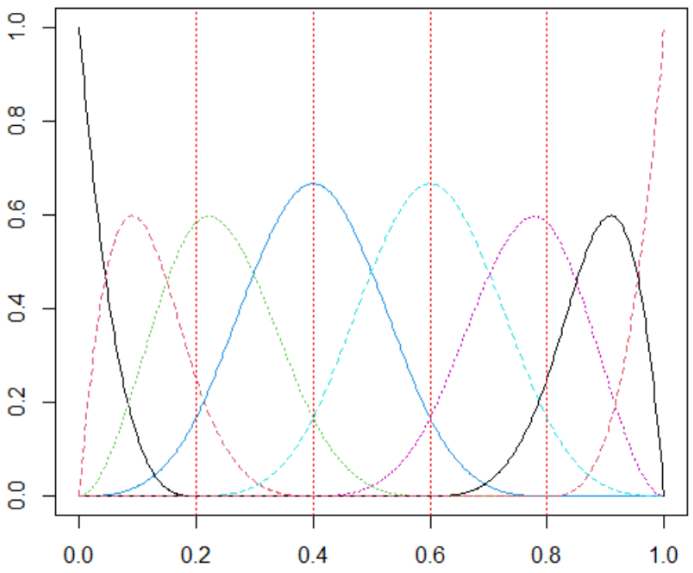


*Figure 4: Example of cubic B-spline basis functions with six equally spaced knots.*

The non-parametric regression model can be written as follows,

$$y_i = f(x_i) + \varepsilon_i = \sum_{j=1}^{J} c_j \cdot \varphi_j(x_i) + \varepsilon_i$$

where $c_j$ can be estimated by minimizing the least squares equation

$$\sum_{i=1}^{n} (y_i - \sum_{j=1}^{J} c_j \cdot \varphi_j(x_i))^2$$

and $\varphi_j(x_i)$ are our known basis functions of the $i^{th}$ observation.

A potential issue that we must account for is our model overfitting the training data and being too "wiggly". To account for this, we will use GCV and OCV to find the optimal tuning parameter and improve our predictions.

Using the OCV (ordinary cross validation) method we can test a specified number of λ values (our tuning parameter) and solve for the minimum of the following function,

$$OCV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \sum_{j=1}^{J} \widehat{\beta}_j \cdot B_j(x_i)}{1 - L_{ii}} \right)^2$$

where $B_j(x_i)$ is a matrix of our cubic B-spline basis functions, $\widehat{\beta}_j$ are our estimated coefficients and $L_{ii}$ are the diagonal values of our $L$ matrix, $L = B(B^T B + \lambda \mathrm{V})^{-1} B^T Y$.

Similarly, for the GCV (generalized cross validation) method the optimal λ value can be found by minimizing,

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \sum_{j=1}^{J}\widehat{\beta}_j \cdot B_j(x_i)}{1 - \frac{v}{n}}\right)^2$$

where the only difference from OCV is that we substitute $\frac{v}{n}$ for $L_{ii}$ where *v* is the trace of *L*, so we are substituting the diagonal of the *L* matrix for the average of the diagonal of the *L* matrix.

# Results

### Parametric Approaches

First, and most importantly, we can use R to fit our LASSO regression model to reduce the number of variables we will use in our subsequent models. Doing so, we can obtain our optimal model and its coefficients are displayed in Table 2.

| Variable | Coefficient | Variable | Coefficient |
|---|---|---|---|
| Points For | 2.58 | Opponent AP Rank | 0.05 |
| Opponent Adjusted Defensive Efficiency Rank | 1.42 | Opponent Home Loses | 0.02 |
| Opponent Non-Conference Strength of Schedule | 1.02 | Opponent Top 25 Wins | -0.07 |
| Adjusted Efficiency Margin | 0.81 | Opponent Adjusted Offensive Efficiency Rank | -0.09 |
| Three Point Attempts | 0.64 | Opponent Luck Rank | -0.36 |
| Non-Conference Strength of Schedule | 0.45 | Average Opponents Defensive Efficiency Rank | -0.52 |
| AP Rank | 0.39 | Adjusted Tempo Rank | -0.88 |
| Average Opponents Offensive Efficiency Rank | 0.27 | Opponent Adjusted Tempo Rank | -1.18 |
| Opponent Average Opponents Offensive Efficiency Rank | 0.24 | Adjusted Offensive Efficiency Rank | -1.41 |
| Top 25 Loses | 0.23 | Opponent Neutral Site Wins | -1.75 |

*Table 2: List of the variables selected by our optimal LASSO regression model.*

The optimal λ value we found for the LASSO model is approximately 0.5129. A higher λ value would provide us with a model with fewer coefficients, whereas a lower λ would result in more coefficients in our model.

Using this LASSO model to predict this year's tournament we predicted 43 (64.18%) of the 67 games correctly. We also found that the mean squared error of our predicted points scored was approximately 15772.9494. This seems like a large number, but we will have to see how our other models perform to know for sure.

Worth noting is that 12 of the 20 variables selected by the LASSO regression are advanced statistics indicating that we may need to rely on metrics that can't be easily seen while watching the game. This may contribute to the reason that it is so difficult to predict each game.

Next, we will use R to fit a multiple linear regression model for the 20 variables selected in our LASSO regression. Table 3 shows the three maximal and minimal coefficients estimated by this linear model.

| Variable | Coefficient | Variable | Coefficient |
|---|---|---|---|
| Points For | 5.42 | Opponent Neutral Site Wins | -3.15 |
| Three Point Attempts | 2.01 | Opponent Adjusted Tempo Rank | -2.49 |
| Average Opponents Offensive Efficiency Rank | 1.79 | Average Opponents Defensive Efficiency Rank | -1.72 |
| Opponent Non-Conference Strength of Schedule | 1.31 | Adjusted Offensive Efficiency Rank | -1.30 |
| Top 25 Loses | 1.15 | Opponent Adjusted Offensive Efficiency Rank | -1.25 |

*Table 3: Maximal and minimal coefficients of our multiple linear regression model.*

This model correctly predicts 43 (64.18%) of the games and has a mean squared error of 19544.4437. Although this linear model less accurately predicts the true number of points scored, it is still able to predict the same number of games correctly as the LASSO model.

Half of the 20 variables in this model are statistically significant at the 5% level and the model has an adjusted R-squared value of 0.4295. This adjusted R-squared value indicates that our model does not do a great job fitting to the data, but since it is not close to zero, we can have some optimism that does do a satisfactory job and should not be discarded.

We can also observe that six of the ten maximal and minimal coefficients are advanced statistics, which notably is the same proportion of advanced statistics as our model is ingesting (12 of 20 variables are advanced statistics).

Our final parametric approach is ridge regression. As mentioned earlier ridge regression is quite similar to LASSO regression with the main difference being the squared value in its penalty term. Fitting a ridge regression model to our 20 variables of interest, we obtain the maximal and minimal coefficients in Table 4.

| Variable | Coefficient | Variable | Coefficient |
|---|---|---|---|
| Points For | 2.76 | Opponent Neutral Site Wins | -2.29 |
| Three Point Attempts | 2.31 | Opponent Adjusted Temp Rank | -1.87 |
| Adjusted Efficiency Margin | 1.36 | Adjusted Offensive Efficiency Rank | -1.32 |
| Average Opponents Offensive Efficiency | 1.35 | Average Opponents Defensive Efficiency Rank | -1.26 |
| Opponent Non-Conference Strength of Schedule | 1.30 | Adjusted Tempo Rank | -1.19 |

*Table 4: Maximal and minimal coefficients of our optimal ridge regression model.*

The optimal $\lambda$ value for our ridge regression was approximately 1.2589. This time we see an increase in advanced statistics in our group of maximal and minimal coefficients, now accounting for 70% of them.

This optimal ridge regression model was the least accurate of our three parametric models, only predicting 40 (59.7%) of the games correctly. It's mean squared error on the other hand was approximately 17383.9814, which is less than that of our linear regression, but higher than our LASSO regression.

Interestingly, although this ridge regression model was less accurate, according to correct predictions, it performed better than the linear regression when predicting the number of points teams will score.

**Non-Parametric Approaches**

Now we will explore the non-parametric side of things. As mentioned in the Method section we will be using cubic B-spline basis functions to help create our non-parametric regression models. To ensure that we don't create a model that overfits our training data we will use generalized cross validation and ordinary cross validation to select a tuning parameter that assures our model is not too "wiggly".

To reduce our analysis to only the most influential variables we will use the three maximal and three minimal coefficients found in our LASSO regression. These variables include: Points For, Opponent Adjusted Defensive Efficiency Rank, Opponent Non-Conference Strength of Schedule, Opponent Adjusted Tempo Rank, Adjusted Offensive Efficiency Rank and Opponent Neutral Site Wins.

These six variables contain both advanced and non-advanced statistics, along with several variables pertaining to a team's opponent. Intuitively we may expect these variables for a team's opponent may not be the best way to predict the amount of points a team will score, but we will need to confirm this hypothesis.

First, we will explore the four variables related to a team's opponent. These variables all seem to be relevant to predicting a March Madness game. Opponent Adjusted Defensive Efficiency Rank captures how strong the opponent is defensively, Opponent Non-Conference Strength of Schedule captures the difficulty level of the teams the opponent played that were not in their conference (better teams will usually test themselves against top teams), Opponent Adjusted Tempo Rank captures the speed at which the opponent plays (a slower team leads to fewer possessions and fewer points scored) and Opponent Neutral Site Wins captures the number of wins a team has on a neutral court, since March Madness is played at a neutral site this may be important.

Creating our non-parametric regression model for Opponent Adjusted Defensive Efficiency Rank using GCV and OCV we find the estimated functions and 95% confidence bands in Figures 5 (GCV) and 6 (OCV).

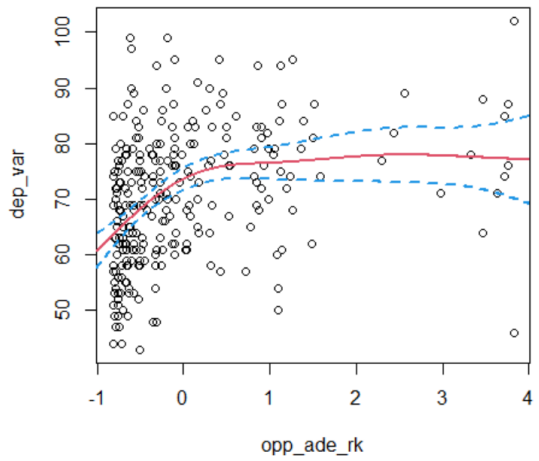


*Figure 5: Estimated function to predict points scored plotted with a 95% confidence band for Opponent Adjusted Defensive Efficiency Rank, using generalized cross validation to select the tuning parameter.*

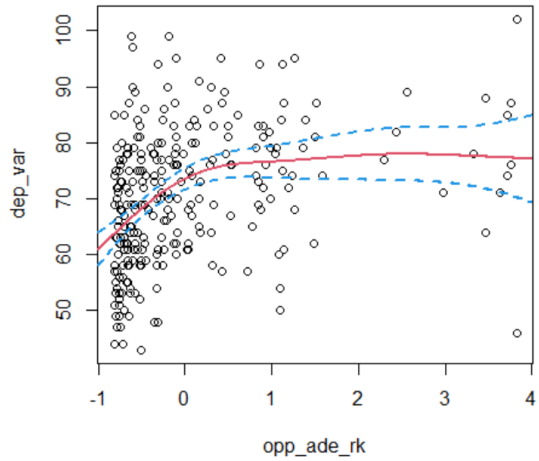


*Figure 6: Estimated function to predict points scored plotted with a 95% confidence band for Opponent Adjusted Defensive Efficiency Rank, using ordinary cross validation to select the tuning parameter.*

The variances of these two estimated functions of points scored are 119.793 for GCV and 119.8367 for OCV.

The optimal tuning parameters found with the GCV and OCV methods are 1.8045 (ln(1.8045) = 0.59) and 2.0967 (ln(2.0967) = 0.74) respectively. Figures 7 and 8 show the minimal natural logarithm of the tuning parameter for the GCV and OCV cases.

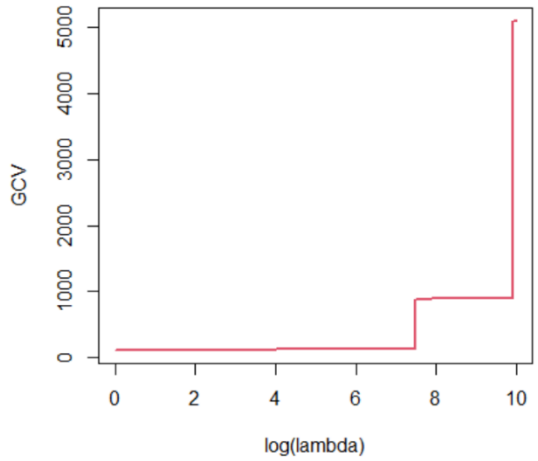


*Figure 7: Plot of the natural logarithm of the tuning parameter and its minimized value for generalized cross validation.*

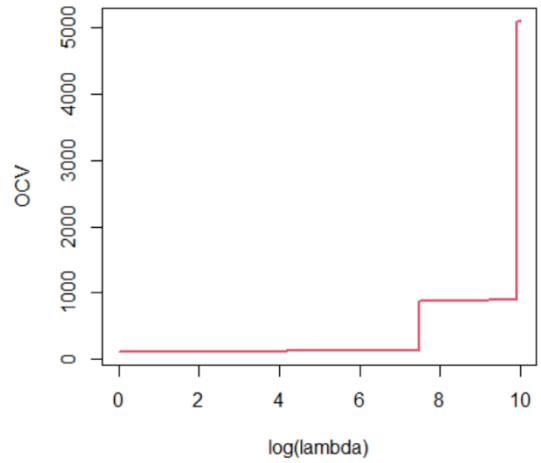


*Figure 8: Plot of the natural logarithm of the tuning parameter and its minimized value for ordinary cross validation.*

Using our first two non-parametric regression models to make predictions, we find that these models only predict 37 (55.22%) and 38 (56.72%) games correctly, our worst result thus far. These models also report the highest mean squared errors of 22165.3489 and 21983.3087, respectively.

It is evident that these two models using Opponent Adjusted Defensive Efficiency Rank do not outperform our parametric approaches, so we will move on to the next variable interest.

Next, we will perform the same non-parametric approach for the Opponent Non-Conference Strength of Schedule variable. Figures 9 and 10 show the estimated functions to predict points scored and 95% confidence bands for GCV and OCV.

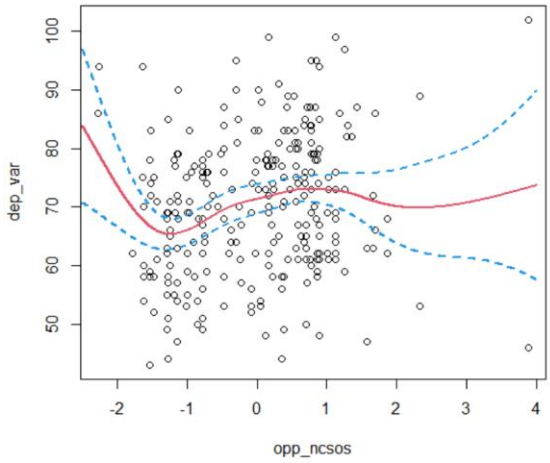


*Figure 9: Estimated function to predict points scored plotted with a 95% confidence band for Opponent Non-Conference Strength of Schedule, using generalized cross validation to select the tuning parameter.*

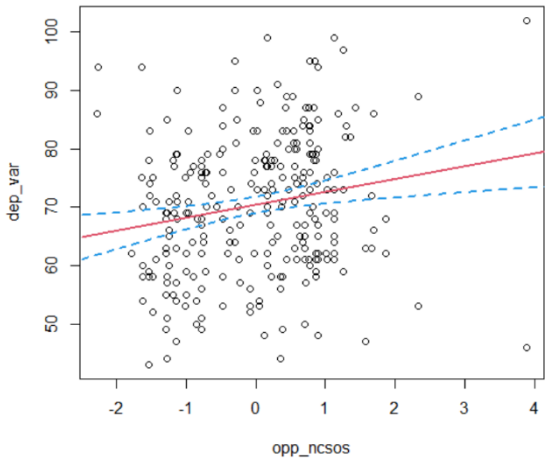


*Figure 10: Estimated function to predict points scored plotted with a 95% confidence band for Opponent Non-Conference Strength of Schedule, using ordinary cross validation to select the tuning parameter.*

The variances of these two estimated functions of points scored are 133.8866 for GCV and 137.7207 for OCV.

The optimal tuning parameters found with the GCV and OCV methods are 0.5002 (ln(0.5002) = -0.6927) and 4226.6905 (ln(4226.6905) = 8.3492) respectively. Notably, the OCV method provides quite a large tuning parameter value compared to GCV and therefore creates a much smoother function in Figure 10 than the more "wiggly" function in Figure 9. Figures 11 and 12 show the minimal natural logarithm of the tuning parameter for the GCV and OCV cases.

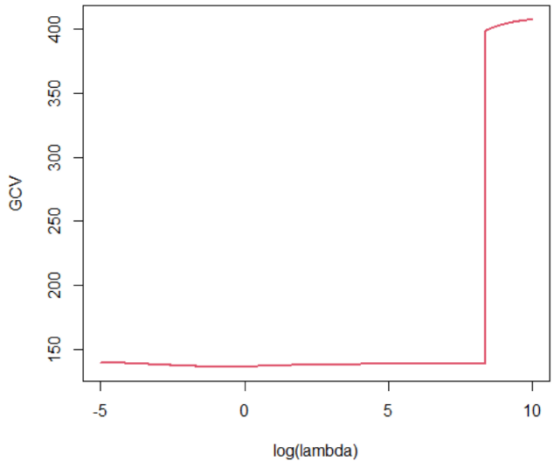


*Figure 11:Plot of the natural logarithm of the tuning parameter and its minimized value for generalized cross validation.*

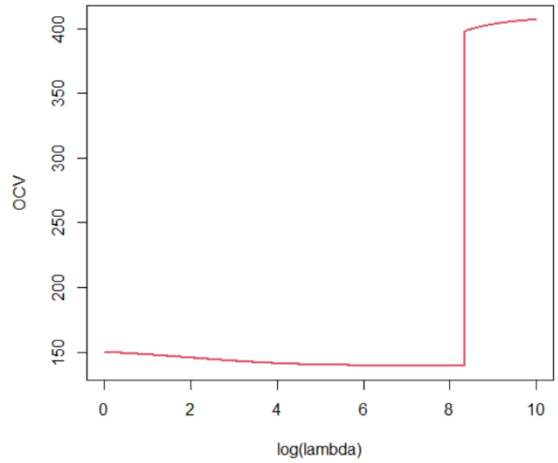


*Figure 12:Plot of the natural logarithm of the tuning parameter and its minimized value for ordinary cross validation.*

These two non-parametric regression models only predicted 34 (50.75%) of the games correctly, performing even worse than our previous non-parametric models. These two models for Opponent Non-Conference Strength of Schedule did, however, have lower mean squared errors than the Opponent Adjusted Defensive Efficiency Rank models, obtaining MSE's of approximately 20212.0280 and 21171.4300 for GCV and OCV.

Two thirds of the way through our non-parametric models and we have yet to find a model that outperforms the parametric models. Luckily, we are saving, what we believe to be, the most promising variables for last, so we will continue with another variable pertaining to a team's opponent.

Our third set of non-parametric models will explore how well the Opponent Adjusted Tempo Rank predicts how many points a team will score. Figures 13 and 14 show the estimated functions to predict points scored and 95% confidence bands for GCV and OCV.

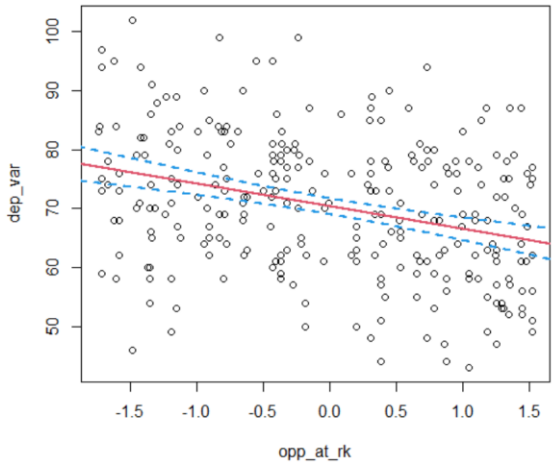


*Figure 13: Estimated function to predict points scored plotted with a 95% confidence band for Opponent Adjusted Tempo Rank, using generalized cross validation to select the tuning parameter.*

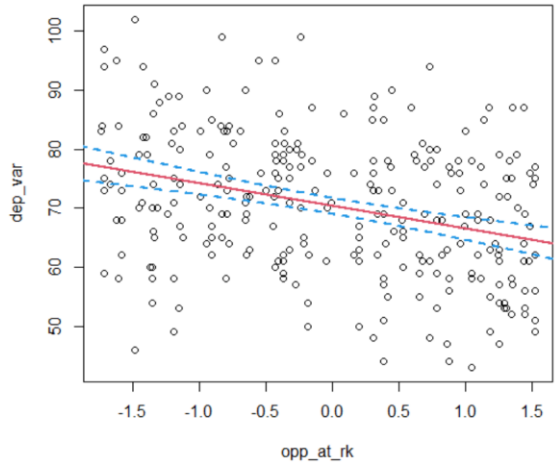


*Figure 14: Estimated function to predict points scored plotted with a 95% confidence band for Opponent Adjusted Tempo Rank, using ordinary cross validation to select the tuning parameter.*

The variances of these two estimated functions of points scored are both 127.8422 as the same tuning parameter was selected for both.

The optimal tuning parameters found with the GCV and OCV methods are both 3115.1171 (ln(3115.1171) = 8.0440). This tuning parameter is quite high, and we see that reflected in Figures 13 and 14 where the line seems to be linear. Figure 15 shows the minimal natural logarithm of the tuning parameter for the GCV case since it is identical to the OCV case.

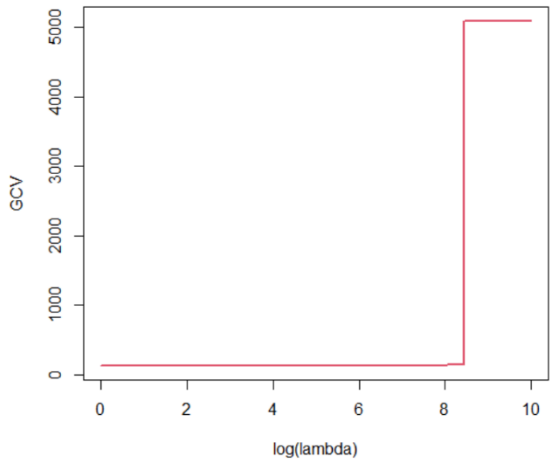


*Figure 15: Plot of the natural logarithm of the tuning parameter and its minimized value for generalized cross validation.*

Unfortunately, these two models are even less accurate than our previous models, only predicting 32 (47.76%) of the games correctly. Strangely these models returned the lowest mean squared error for a non-parametric model of approximately 19805.3004, which rivals that of our multiple linear regression.

Next, we will create the final set of models for our variables related to opponent characteristics. Figures 16 and 17 show the estimated functions to predict points scored and 95% confidence bands for GCV and OCV, with respect to the Opponent Neutral Site Wins variable.

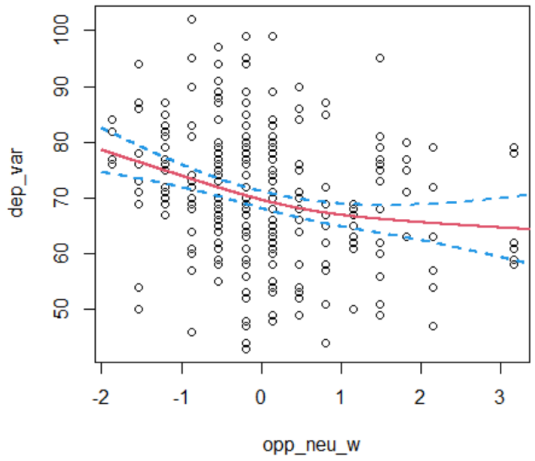


*Figure 16: Estimated function to predict points scored plotted with a 95% confidence band for Opponent Neutral Site Wins, using generalized cross validation to select the tuning parameter.*

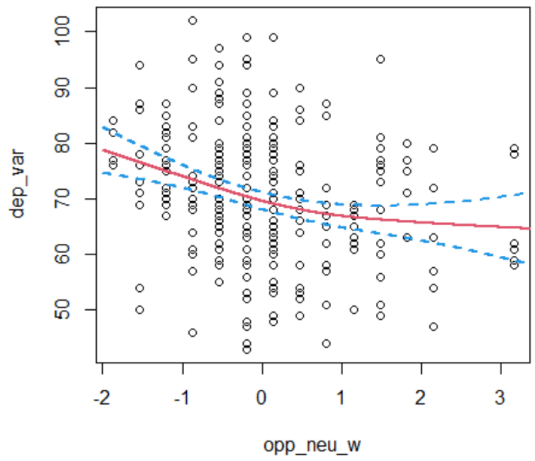


*Figure 17: Estimated function to predict points scored plotted with a 95% confidence band for Opponent Neutral Site Wins, using ordinary cross validation to select the tuning parameter.*

The variances of these two estimated functions of points scored are 132.835 for GCV and 132.8075 for OCV.

The optimal tuning parameters found with the GCV and OCV methods are 28.4081 (ln(28.4081) = 3.3467) and 23.6079 (ln(23.6079) = 3.1616) respectively. Figures 18 and 19 show the minimal natural logarithm of the tuning parameter for the GCV and OCV cases.

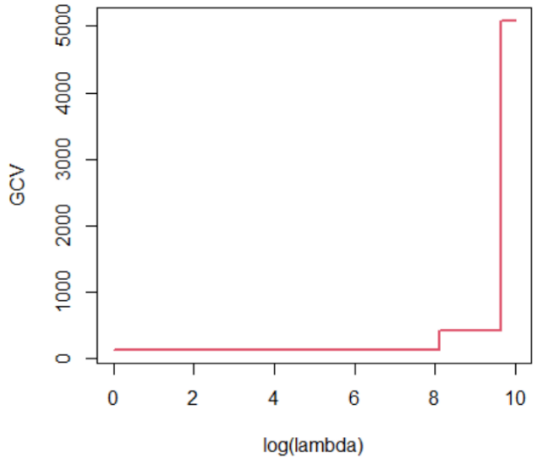


*Figure 18: Plot of the natural logarithm of the tuning parameter and its minimized value for generalized cross validation.*

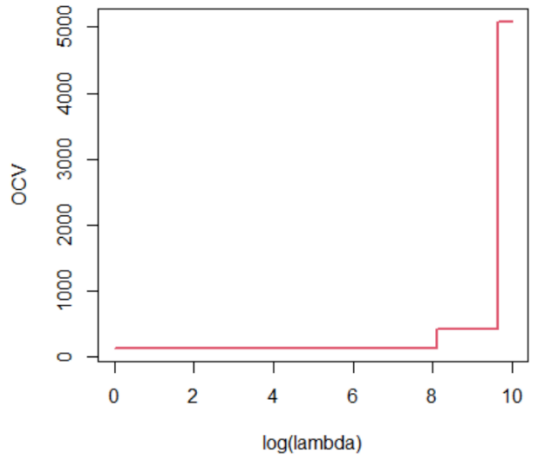


*Figure 19: Plot of the natural logarithm of the tuning parameter and its minimized value for ordinary cross validation.*

These two non-parametric models predicted 39 (58.21%) games correctly, the most of any of our non-parametric models. However, this increase in accuracy did not translate to more accurately predicting the number of points scored as we found the mean squared errors to be approximately 21750.3951 and 21805.2785.

With the models that were expected to not be the most accurate out of the way we can now explore the non-parametric models that, intuitively, seemed to be the most promising. The first set of non-parametric regressions we will test are centered around the Points For variable, which is the sum of the points the scored in each game throughout the season.

Figures 20 and 21 show the estimated functions to predict points scored and 95% confidence bands for GCV and OCV.

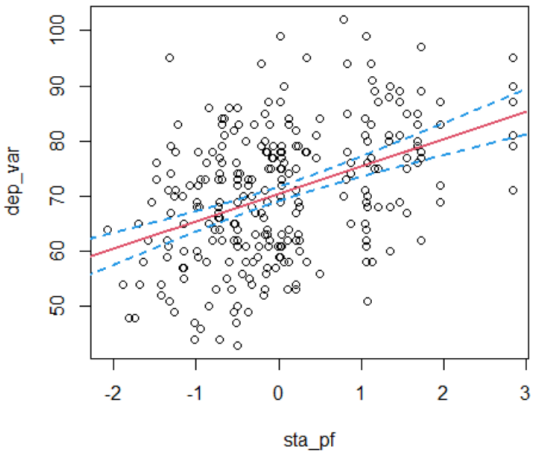


*Figure 20: Estimated function to predict points scored plotted with a 95% confidence band for Points For, using generalized cross validation to select the tuning parameter.*

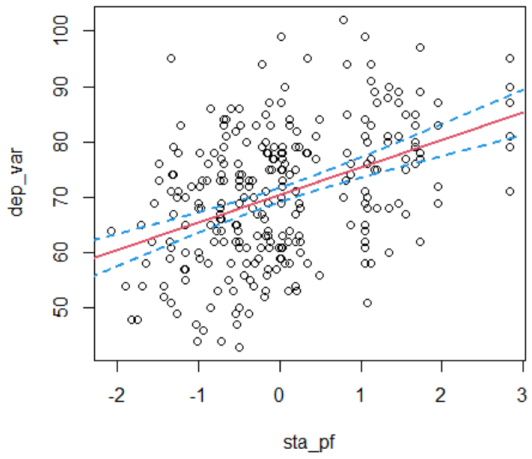


*Figure 21: Estimated function to predict points scored plotted with a 95% confidence band for Points For, using ordinary cross validation to select the tuning parameter.*

The variances of these two estimated functions of points scored are 117.8161 for both the GCV and OCV models, since the same tuning parameter was selected for both.

The optimal tuning parameters found with the GCV and OCV methods are both 4142.9550 (ln(4142.9550) = 8.3292). This tuning parameter is quite high, and we see that reflected in Figures 20 and 21 where the line seems to be linear. Figure 22 shows the minimal natural logarithm of the tuning parameter for the GCV case since it is identical to the OCV case.

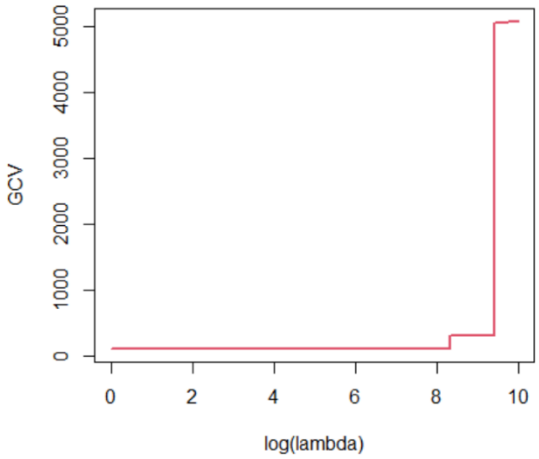


*Figure 22: Plot of the natural logarithm of the tuning parameter and its minimized value for generalized cross validation.*

These non-parametric regression models successfully predicted 44 (65.67%) of the games correctly. This model more accurately predicted winning teams than any of our other parametric and non-parametric models. With this increased accuracy it also reported the lowest mean squared error of all the non-parametric regression models at approximately 18331.4679.

Our final set of non-parametric regression models will use the Adjusted Offensive Efficiency Rank variable. Figures 23 and 24 show the estimated functions to predict points scored and 95% confidence bands for GCV and OCV.

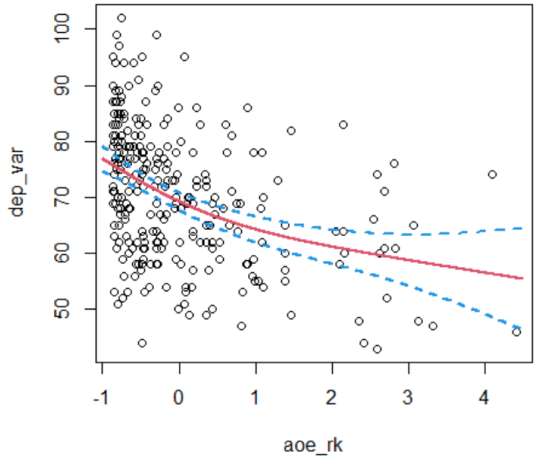


*Figure 23: Estimated function to predict points scored plotted with a 95% confidence band for Adjusted Offensive Efficiency Rank, using generalized cross validation to select the tuning parameter.*

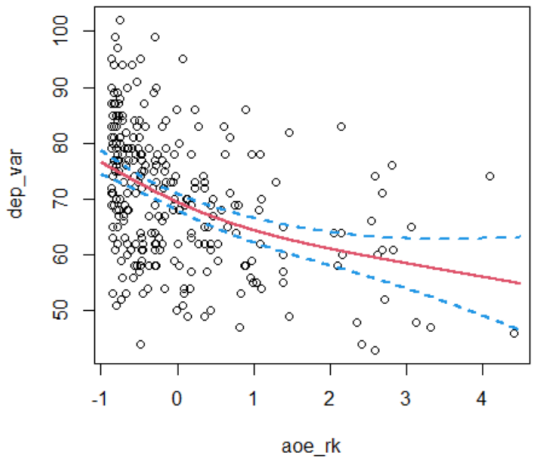


*Figure 24: Estimated function to predict points scored plotted with a 95% confidence band for Adjusted Offensive Efficiency Rank, using ordinary cross validation to select the tuning parameter.*

The variances of these two estimated functions of points scored are 118.8281 for GCV and 118.9358 for OCV.

The optimal tuning parameters found with the GCV and OCV methods are 13.4140 (ln(13.4140) = 2.5963) and 20.4199 (ln(20.4199) = 3.0165) respectively. Figures 25 and 26 show the minimal natural logarithm of the tuning parameter for the GCV and OCV cases.

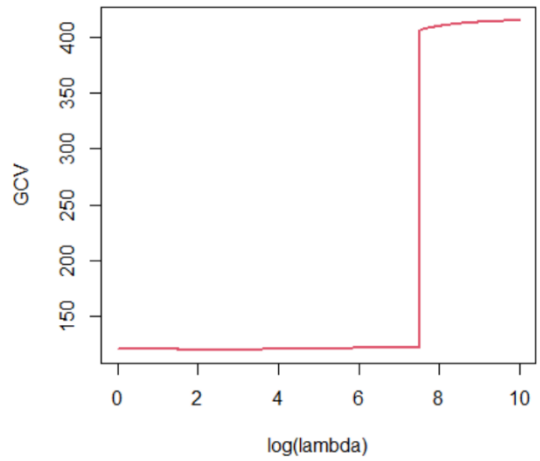


*Figure 25: Plot of the natural logarithm of the tuning parameter and its minimized value for generalized cross validation.*

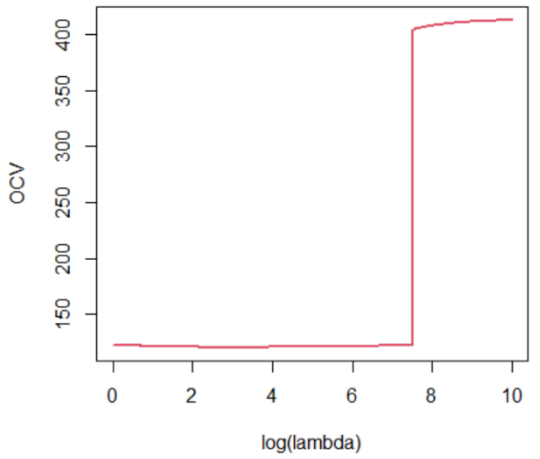


*Figure 26: Plot of the natural logarithm of the tuning parameter and its minimized value for ordinary cross validation.*

This non-parametric regression model for Adjusted Offensive Efficiency Rank significantly outperformed all our other models, correctly predicting 48 (71.64%) of the games. The mean squared errors for the GCV and OCV models were approximately 16662.0014 and 16664.2112 respectively and were the second lowest mean squared errors behind only our original LASSO regression model.

## Conclusions and Discussion

Table 5 summarizes the results obtained from our three parametric and twelve non-parametric models.

| Model | MSE | Correct | Lambda | Variance |
|---|---|---|---|---|
| LASSO | 15772.9494 | 43 | 0.5129 | |
| Linear | 19544.4437 | 43 | | |
| Ridge | 17383.9814 | 40 | 1.2589 | |
| Points For (GCV) | 18331.4679 | 44 | 4142.955 | 117.8161 |
| Points For (OCV) | 18331.4679 | 44 | 4142.955 | 117.8161 |
| Opponent Neutral Site Wins (GCV) | 21750.3951 | 39 | 28.4081 | 132.8350 |
| Opponent Neutral Site Wins (OCV) | 21805.2785 | 39 | 23.6079 | 132.8075 |
| Opponent Adjusted Defensive Efficiency Rank (GCV) | 22165.3489 | 37 | 1.8045 | 119.7930 |
| Opponent Adjusted Defensive Efficiency Rank (OCV) | 21983.3087 | 38 | 2.0967 | 119.8367 |
| Adjusted Offensive Efficiency Rank (GCV) | 16662.0014 | 48 | 13.4140 | 118.8281 |
| Adjusted Offensive Efficiency Rank (OCV) | 16664.2112 | 48 | 20.4199 | 118.9358 |
| Opponent Non-Conference Strength of Schedule (GCV) | 20212.0280 | 34 | 0.5002 | 133.8866 |
| Opponent Non-Conference Strength of Schedule (OCV) | 21171.4300 | 34 | 4226.6905 | 137.7207 |
| Opponent Adjusted Tempo Rank (GCV) | 19805.3004 | 32 | 3115.1171 | 127.8422 |
| Opponent Adjusted Tempo Rank GCV) | 19805.3004 | 32 | 3115.1171 | 127.8422 |

*Table 5: Summary of the prediction accuracy of our 15 estimated models.*

Interestingly a non-parametric model with only one independent variable outperformed all of our parametric models with 20 independent variables when predicting the winner of each game. We were not expecting this single variable model to perform so well and predict the winning teams this accurately. Intuitively one would assume that more variables would allow the model to create a more accurate representation of how each game would unfold, but those variables may have only been contributing noise to our predictions.

The flexibility allowed by the non-parametric models seems to allow them to follow trends in the data more accurately. As the researcher, if we can minimize the amount of "wiggle" in our model then we have the opportunity to create quite an accurate model with minimal variables.

Even though, our best non-parametric regressions outperformed the parametric regressions when predicting each games winner, they were unable to outperform them in mean squared error. LASSO regression, our first model, held on to its position with the lowest mean squared error. Our Adjusted Offensive Efficiency Rank models did finish a close second, though.

As an individual interested in predicting March Madness the choice of optimal model is dependent on the intended use of the model's predictions. If the user wishes to fill out a March Madness bracket, then the non-parametric regression model for Adjusted Offensive Efficiency Rank would be the best choice.

However, for an individual interested in betting on games, based upon points, for example will the total points scored in the game exceed 157.5 or will Gonzaga beat Baylor by more or less than 4.5 points. In this case the individual would prefer to use the LASSO regression model, since it has the lowest mean squared error and predicted points scored in each game the most accurately of the 15 models.

If we can only choose one of these models, the non-parametric regression model for Adjusted Offensive Efficiency Rank would be our choice as the best model. Although it did not produce the lowest mean squared error, it significantly outperformed all other models when predicting each game, which makes up for the small loss in score precision. This model seems to have the most balanced predictions and gives the user a best of both worlds scenario. This model successfully predicted upsets like #11 UCLA over #6 BYU, #12 Oregon State over #5 Tennessee, #11 Syracuse over #6 San Diego State, #10 Rutgers over #7 Clemson, #6 USC over #3 Kansas and #11 UCLA over #2 Alabama.

It was interesting to observe how the different non-parametric models were able to predict certain upsets. For example, the non-parametric regression model for Points For was able to predict extremely unlikely upsets like #13 Ohio over #4 Virginia, #14 Abilene Christian over #3 Texas, #15 Oral Roberts over #2 Ohio State and #15 Oral Roberts over #7 Florida. The issue with this model is that, even though it predicts these massive upsets, it inaccurately predicts quite a few of the games between teams that are similar in skill.

If we were able to combine some of these non-parametric regressions into one prediction model, we could potentially create a very accurate model. Many statisticians spend their lives working on these sorts of problems and trying to figure out how to create a balanced model that can predict the close games while also predicting the major upsets. We feel this was a successful first attempt at solving this problem and is a good start that we can continue to work from and hopefully find a solution to this problem in the future.

# Appendix

## Bibliography


Larry Wasserman. *All of Nonparametric Statistics.* Springer, 1999.

Tianyu Guan. *MATH 4P82: Nonparametric Statistics Notes* Brock University, 2021.

*NCAA Bracket for March Madness*. National Collegiate Athletic Association, www.ncaa.com/march-madness-live/bracket.

*Men's College BASKETBALL POLLS*. Fox Sports, www.foxsports.com/college-basketball/polls?season=2019&poll=1.

*2021 College Basketball STANDINGS: College Basketball*. Associated Press, collegebasketball.ap.org/standings/2021/d1.

*ESPN*, ESPN Internet Ventures, www.espn.com/mens-college-basketball/statistics/team/_/stat/scoring/sort/points.

*2021 Pomeroy College Basketball Ratings*, Ken Pomeroy, kenpom.com/.