Predicting Outcomes of Regular Season NBA Games to Minimize Risk When Placing Wagers

IE 590: Big Data Risk Analytics for Engineering Management & Public Policy

Overleaf Link
Jakson Terpak
Benjamin Sons
Austen Horton

PURDUE
UNIVERSITY

Executive Summary

In recent years, the Sports Gambling Industry has seen a tremendous increase and is expected to grow more. After the decision from the United States Supreme Court that the Professional and Amateur Sports Protection Act of 1992 was unconstitutional, it has been left up to the States to decide their legislation on sports betting. With a dozen states providing the legal avenue for online sports betting and the number projected to grow, many have wondered if there is any way to predict sports statistics to assist in their betting. (Dorson, 2020)

With basketball being the number five sport based on betting volume, our goal was to create a model that would assist in predicting the outcome of NBA Games. As avid basketball fans, we came into this with some prior hypotheses that will be discussed later, however, we knew that we had two avenues to go about our predictions. First, we could utilize regression models to predict team points based on various statistics. Second, we could use a classification model to predict the outcome of a game based on similar statistics. Ultimately, we chose to do both and compare our findings. (Purdum, 2020)

Key words

Basketball; Classification; Regression; Modeling; NBA

Contents

List of Tables

List of Figures

1. Introduction

Our goal was to create a model that will best minimize the risk in Sports Gambling. Therefore, as Kaplan and Garrick recognize, it is vital that our model addresses the risk triplets: What can go wrong? What is the probability? What are the consequences (Kaplan and Garrick, 1981)?

In our model, what can go wrong is fairly straightforward as the bet can be lost by the team winning or losing. The consequences of this are also relatively straight-forward as one can either win money or lose money. That being said, our model aims to determine the probabilities of winning and losing. However, Sports betting is not as simple as betting on wins and losses, there are also other things that our model would accomplish. For example, a sports better could bet on a parlay which are combined bets into a single wager. In fact, legislation in some areas, such as Ontario, Canada require this. That being said, our model would provide win and loss probabilities that would then be paired with decision trees to determine whether a bet is worth taking. This also adds to the consequences by providing a quantitative dollar value.

Using Figure 1 as an example, we would use our best model to estimate the chance that a team wins and the chance that a team loses. In this example, there is a sixty-four percent chance the team wins and a thirty-six percent chance that a team loses. Due to this, the realized value if a bet was made on the team to win would be one hundred and five dollars where the realized value if a bet was made on the team losing would be thirty-six dollars. The realized value would be calculated by taking the summation of the probabilities multiplied by the payouts. In this example, the payout would be higher for the loss since payouts are often higher for riskier bets. That being said, our model would recommend betting on the team to win as that will minimize risk and still provide a payout.

This approach to minimize risk also takes into account other very important risk factors, such as uncertainty and the current state of knowledge. As discussed below, we will be formulating our models based on averages from a prior number of games. This assists in the prevention of over-fitting and also conceptually makes sense as many in-game statistics are not known until after the game is played.

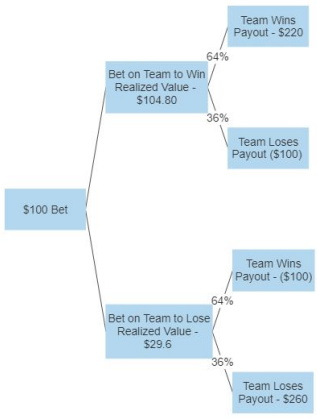

Fig. 1. Decision Tree Analysis for one team bet.

When approaching this problem, there are questions that arise such as what model will be best for our goals and what variables will be important in our model. For this project, there are two approaches that we are going to take. The first approach will be using regression models to predict points that a team will score and then also predicting points that the opponent will score. In this case, the winner would be determined by the team that scores more points. The second approach that will be used is classification models to predict the binary result of a win or loss of a team. Our team hypothesises that a Five Game Average Random Forest Classification Model will be best for predicting games and that the best variables in that prediction are going to be the Team Elo Score, the average Offensive Ranking in the last five games, and the average Defensive Ranking in the last five games. Our reasoning for this hypothesis is due to past research that has been done on this subject; in many cases, a Random Forest Classification Model does the best. Also, the Elo Rating is highly representative of many factors of the Team Box Score, as well as the Offensive and Defensive Rating. These variables will be discussed further in the following sections.

1.1 Prior Work

There has been extensive research on predictive modeling in the NBA. To begin our research we began by searching for other papers that had built predictive models based on NBA Statistics as we wanted to get an idea on the best models to use as well as any research gaps. This lead us to a few research reports that we thought were relevant. First, was a paper by Josh Weiner et. al where they developed a predictive model utilizing Elo Ratings and Average points from the prior ten games; ultimately, their best model was a Random Forest Classifier with an accuracy of around sixty-seven percent (Weiner, 2021). Their paper lead to our discovery of Elo Ratings and was an inspiration for using them in our model, however, we hypothesized that other variables were of more significance. Another realization in this paper that we thought was important was that single player stats are insignificant in team results. This makes sense as player scoring only contributes so much to the actual team score and there are many variable factors when it comes to single

players, such as injuries. For this reason, we realized that player statistics were fairly irrelevant to our goal and omitted them from our data so avoid this research gap.

Another article that we had researched was by Alaxander Fayad and his best model was a Support Vector Machine classification model with an accuracy of around seventy-three percent, however, he noted that his model was biased due to its use of monthly stats to predict game stats. In essence, he was using factors that would have not yet been known to predict the outcomes of the games. (Fayad, 2020). This was another research gap we learned of and decided that the best way to mitigate this over fitting was by utilizing average stats over past games. Fortunately, his article did lead to some research pertaining to what NBA statistics actually translate to wins. This article stated that some of the best statistics in predicting wins were offensive rating, defensive rating, and rebound differential.

Finally, we researched the NBA Four Factors by Dean Oliver, a well known NBA Statistician. These are Effective Field Goal Percentage, Turnover Percentage, Offensive Rebounding Percentage, Defensive Rebounding Percentage, Free Throw Percentage, Opponent Effective Field Goal Percentage, Opponent Turnover Percentage, and Opponent Free Throw Percentage. The reason it is actually eight factors is because it is four for each team (Jacobs, 2017). This research also assisted in making our hypotheses on the important variables, however, we ultimately decided that Offensive and Defensive rating would be better as they took these into account and also suppressed excessive correlation between variables.

After reading the above research, we had decided on our predictions and our goals even further. While our predictions were the same in that a Random Forest Classifier focusing on the three variables of the playing team's Elo Rating, offensive rating, and defensive rating would perform best in out of sample accuracy, we also learned of other factors that may possibly contribute in our models. We also had a goal of creating a model that would accurately predict NBA games around 67% of the time. Other research of ours, such as the article by Kimbugwe Nasser, shows that some models can achieve an accuracy upward of 70%. However, due to time constraints, our goal was more so focused on creating a simple model that is better than chance and challenges other research we had seen prior (Nasser, 2016).

2. Data and Methods

This section will explore the relationships present in the refined data set as well describe different regression and classification methods used to both explain past games and predict future games. We will discuss predictor variable interaction and selection, response variable selection and usage, and the methods in which the predictors and responses are utilized.

2.1 Data

In the modern-day era of the NBA, there are countless statistics that describe nearly every aspect of the game of basketball. How well a team shoots the ball can be characterized by 5 different shooting percentages. How often a team turns the ball over can be expressed in turnovers per game or turnover percentage or even turnovers per offensive possession. Oftentimes there are multiple statistics that seem to serve identical purposes but with small, nuanced differences. These differences can be used to account for each aspect of basketball from both an offensive and defensive perspective (team vs. oppt). A full explanation of each variable in our base data set is in the appendix, however, I will display the maximum set of variables used in any of our regressions below.

Table 1. Input data (each variable has two values per game, a team value and an opponent value).

Statistic[1]	Response/Predictor	Units w/ Source
Points	Response	PTS
Result	Response	Win or Loss
Elo	Predictor	Formulaic Rating
Offensive Rating	Predictor	Formulaic Rating
Defensive Rating	Predictor	Formulaic Rating
Location	Predictor	Home or Away
Days Off	Predictor	# of days since last game
Assists	Predictor	AST
Turnovers	Predictor	TO
Turnover %	Predictor	TO%
Field Goals Made	Predictor	FGM
Free Throws Made	Predictor	FTM
2 Point FGs Made	Predictor	2PM
3 Point FGs Made	Predictor	3PM
Offensive Rebounds	Predictor	ORB
Off. Rebound Percentage	Predictor	ORB%
Total Rebounds	Predictor	TRB
Total Rebound Percentage	Predictor	TRB%
True Shooting Percentage	Predictor	TS%
Effective Field Goal Percentage	Predictor	EFG%
Assist to Turnover Ratio	Predictor	AST:TO
Steal to Turnover Ratio	Predictor	STL:TO
Steals	Predictor	STL
Steal Percentage	Predictor	STL%
Blocks	Predictor	BLK
Block Percentage	Predictor	BLK%
Personal Fouls	Predictor	PF
Defensive Rebounds	Predictor	DRB
Def. Rebound Percentage	Predictor	DRB%
Floor Impact Counter	Predictor	Formulaic Rating
Point Per Shot	Predictor	PPS
Assist Rate	Predictor	Formulaic Rate
Block Rate	Predictor	Formulaic Rate
Pace	Predictor	poss/game
Free Throw Rate	Predictor	Free Throw made per Field Goal Attempted height

[1]All sources are sited at the end of this paper.

1

2.2 Exploratory Data Analysis and Visualization

As mentioned previously, some NBA statistics have inherent correlation such as Field Goals Made and Field Goal Percentage. Both of these are positively correlated since Field Goal Percentage uses the number of Field Goals Made in its calculation. This is just one example of how some statistics are naturally intertwined. To find lesser-known instances

[1]In the raw imported data frame, each statistic has two values for each game, a team value and an opponent value.The raw data is found from (Rossotti, 2019)

of an individual predictor influencing the response variable(s) or other predictors, correlation plots; pairwise plots; and plots to test linearity must be employed. Furthermore, the insight gained from these visualizations is essential for identifying which regression or classification method should be used.

2.2.1 Correlation Plots

In order to better visualize the correlation between predictors and response variables, the data must be compartmentalized. An effective way to do so is to divide the data into offensive statistics and defensive statistics.

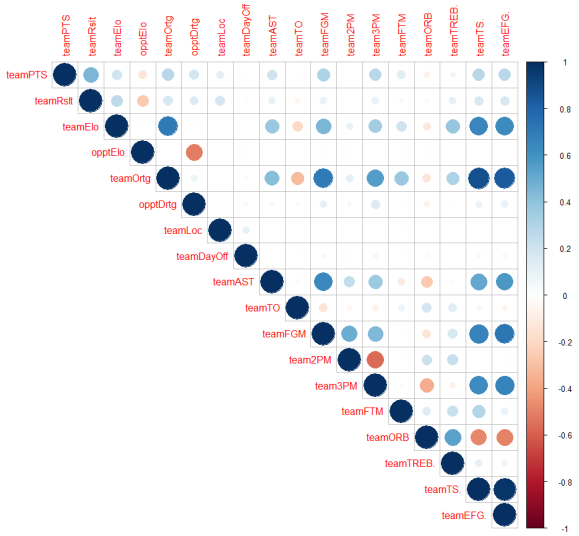

Fig. 2. Offensive Correlations

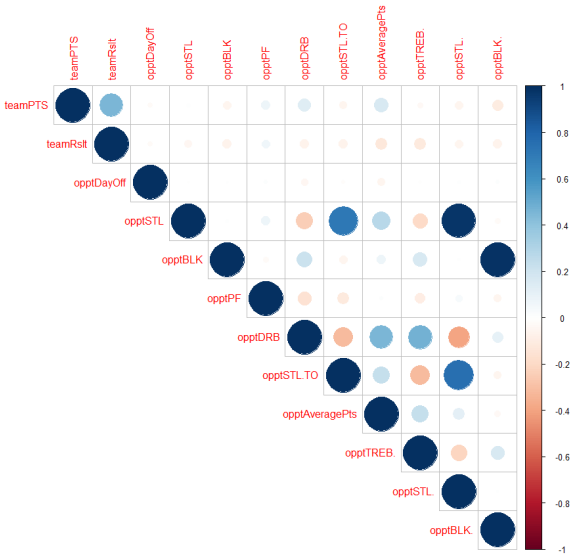

Fig. 3. Defensive Correlations

As seen in both Figure 2 and 3, there are some alarmingly high correlations. However, after taking an in depth look at each plot, the correlations start to make sense. For example, 'opptSTL' (number of opponent steals) and 'opptSTL.' (opponent steal percentage) have a correlation of 0.97. This is logical since the steal percentage uses the number of steals in its calculation. This can be said for every stat that has a regular quantity and a percentage value counterpart (DREB, OREB, BLK, etc.). Correlation plots are a great method to identify potentially significant variables which could make regression analysis easier in the long run. Predictors relative to offensive stats which give cause to investigate the variable for significance (as seen in Figure 2) are teamElo, teamOrtg, teamLoc, teamAST, teamFGM, team3PM, teamTS., and teamEFG.. All appear to have an impact on both teamRslt and teamPTS. Predictors relative to defensive stats which give cause to investigate the variable for significance (as seen in FIgure 3) are opptElo,opptBLK, opptDRB, and opptSTL.TO. While these defensive predictors mildly correlate to teamRslt, they have a stronger correlation with teamPTS.

2.2.2 Linearity Assessment

Based on existing knowledge and a practical sense for the game of basketball, we did not expect the data to be linear due to the short-term variability that the NBA is known to possess. However, we did suspect that linear regressions could be useful for modeling long-term trends in the NBA. An example of this is the increase in popularity of the 3-point line over the last decade, as seen in the level of significance '3PM' has within our resultant regressions seen later in this paper. Figures 4 and 5 below tests the data against the four assumptions of linear regressions.

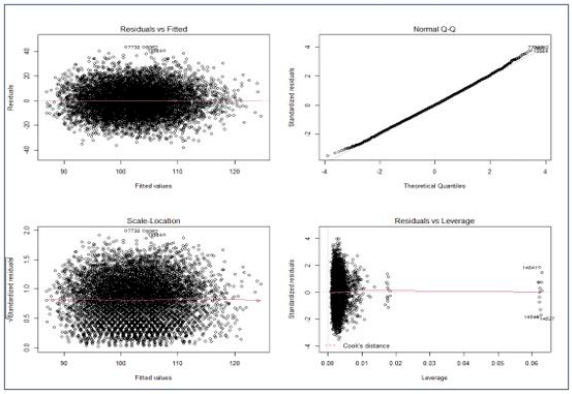

Fig. 4. Linear Model 1 Assumptions

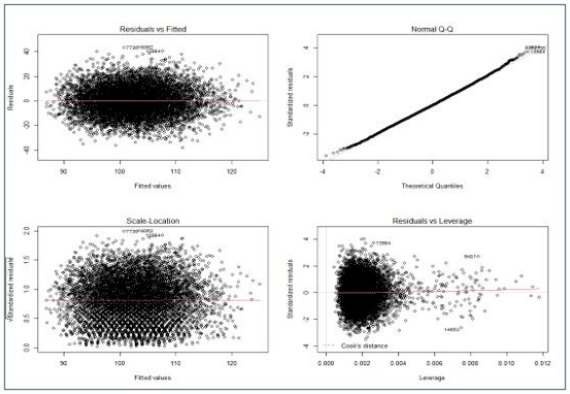

Fig. 5. Step Model 1 Assumption

The data seems nearly perfectly normal, which could cause our linear models to predict overall similar results. The graphs above prove that the data was more linear than previously expected but is still not perfectly linear as seen in the "Residuals vs. Fitted Graph" in the top-left quadrant of each plot set. The use of pairwise plots between the predictors and response variable can be helpful in identifying which predictors are responsible for the lack of linearity. A general additive model can then be used to include, exclude, spline, make exponential, or keep linear different variables based on fit and accuracy metrics. See Figure 16 in the appendix for the full matrix of scatter plots used to identify non-linear relationships.

2.3 Methods

When running regressions, it is difficult to identify from an initial glance which type will fit the data best even with the utilization of the correlation plots. Early in the data analysis work, the team was faced with a problem. How can we regress future games (validation set of games) if we are inputting all of the statistics for those games and only withholding teamPTS and teamRslt? The regression can figure out the result every time based on 2PM, 3PM, and FTM and accurately predict each game because it was already given the statistics for the game. To counteract this, the team decided to average the previous 'n' number of

games, creating matrices that record the appropriate statistics of each team from several different time frames. There were some statistics that would not make sense to average, such as Elo Rating, and the number of days off prior to that game, so we did not average those statistics. By varying 'n', we can account for the volatility of NBA games and see which team(s) are experiencing hot or cold streaks. It also gives our models more predictive power since it can account for short-term trends in the data, increasing the ability of models to identify upsets more often. In order to train and assess the performance of each model, we used supervised learning in the form of 10-fold cross-validation and portioning of data into learning and validation sets. As seen in Figure 1, there are two response variables. Having access to two response variables gives us the opportunity to use different modeling methods, since some models work better with continuous response variables and some require categorical data. Since the variable "teamPTS" is continuous, it will be the response variable for all linear regressions. In contrast, "teamRslt" is categorical (1 = win, 0 = loss) and will serve as the response variable for all classification models. To understand game results better, we will break down the causes of winning into two different factors: team offensive prowess and opponent defensive prowess. This breakdown is important since a team could hypothetically record great offensive statistics per game but still lose because their defense is unable to sufficiently limit the opponent's scoring ability. Once again, this increases the predictive power of our models by accounting for the opponent's recent performance, as opposed to only focusing on the historical performance of one team in any given matchup.

2.4 Regression - Multiple Linear Regression

A multiple linear regression attempts to explain the behavior of a response variable by weighting different predictors to account for the impact that each independent variable has on the dependent variable. Each parameter is tuned with the goal of minimizing the residual sum of squares. We made use of the Gaussian method when using a continuous response variable and the Binomial method when using a categorical response variable. Below is the general equation used in Multiple Linear Regressions.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \epsilon$$

Fig. 6. Multiple Linear Regression Equation

2.5 Regression - Stepwise

In an attempt to decrease complexity without sacrificing model performance, a stepwise regression is useful for sifting through variables that may not be significantly contributing to the response variable behavior. Both forward and backward variable selection will be used to remove noisy data and unnecessary variables that may be causing over or underfitting. Forward variable selection starts with the null model, testing and adding predictors

with the goal of minimizing the residual sum of squares. Backward variable selection does the opposite, taking the model with every predictor and systematically testing variables and removing those with significance outside of the appropriate threshold. Step-wise regressions help to manage the bias-variance tradeoff of the model.

2.6 Regression - Lasso & Ridge

The use of a lasso or ridge regression is justified when a data set contains variables that are highly multi-collinear. The lasso method executes variable selection automatically and utilizes shrinkage of variables towards a central point. The ridge method is particularly helpful in cases with limited observations and an ample number of predictors since it adds necessary bias to regression estimates to reduce standard error. See Figure 7 and 8 for the equations utilized by the lasso and ridge methods.

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Fig. 7. Lasso Regression Equation

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

Fig. 8. Ridge Regression Equation

2.7 Regression - GAM

General additive models (GAMs) attempt to better fit local parts of linear model that show non-linear characteristics. Each predictor is represented by a function as opposed to a coefficient to map specific parts of the data more accurately. Functions used include exponential, linear smoothing, splining, and exclusion, all of which make the linear model more flexible and locally sensitive to the data.

$$g(E[y]) = \beta_0 + \sum_{j} f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$

Fig. 9. GAM Regression Equation

2.8 Classification - Logistic Regression

As Logistic Regression is a Parametric Model, the variables that go into it are very important. Due to this, we thought it would be best to do variable selection to reduce the noise within our data. The two methods we used were Variable Importance & Recursive Feature Elimination. Both of these methods utilized repeated ten-fold cross-validation to determine the most significant variables for all of our datasets. For the Variable Importance Selections, all of our datasets had the same significant variables: teamElo, opptElo, teamFIC, teamOrtg, teamEFG. The RFE Selections yielded different significant variables. The dataset using five game averages had teamElo, opptElo, opptFIC, teamOrtg, and teamFIC as significant variables. The dataset using fifteen game averages had teamElo, opptElo, opptFIC, teamFIC, and teamEFG as significant variables. Finally, the dataset using twenty-five game averages had teamElo, opptElo, opptFIC, teamFIC, and teamEFG as significant variables. We then created subsets of the five game, fifteen game, and twenty-five game averages for both Variable Importance and RFE using the significant variables and split the data into training and test sets where eighty percent of the data was training data and the remaining twenty percent was test data.

All of the training datasets were then used to fit logistic classification models through ten-fold cross validation where the best model was chosen using the estimated in sample accuracy. These were then tested on the testing data to estimate the out of sample accuracy. A couple of key observations that were made were that teamElo and opptElo were highly significant in every model and opptFIC was very significant in the Five Game Average RFE Model. Another observation was that the Variable Inflation Factors of all the variables in each model stayed below ten. This assisted in the prevention of overfitting the data as well as keeping the logistic regression assumption of observational independence and negligible multicollinearity.

2.9 Classification - Trees

As Classification Trees are non-parametric models, we did not utilize any variable selection prior besides the removal of redundant and highly correlated variables. Similar to above, we did split the data into eighty percent training data and twenty percent testing data. For each dataset pertaining to the number of games utilized for the average, we then fit a Classification Tree Model. When fitting the models, we also used ten-fold cross-validation as a training parameter and this chose the best complexity parameter based on the estimated in sample accuracy for all of our models. The complexity parameter vs. the accuracy of the of the Five Game Classification Model is graphically represented in Figure 17 of the appendix. The tree structure can be seen in appendix Figure 18 of the appendix. The variables that were utilized in the Nodes of this model were teamElo, opptElo, teamAveragePts, teamFIC, teamOrtg, and opptEFG and the optimal complexity parameter was 0.001176471.

2.10 Classification - Random Forests

Similar to the variables utilized in the Classification Tree Models, we did not do any selection past the removal of redundant and highly correlated variables. Also similar to above, data was split into eighty percent training data and twenty percent testing data. When fitting the Random Forest Models, two packages were used: Caret and Random Forest. To fit the Random Forest Models using the Caret Package, for each dataset pertaining to the selected average number of games, we utilized ten-fold cross-validation as a parameter on the training data.This found the model with the optimal number of randomly selected predictors based on the estimated in-sample accuracy. To fit the Models using the Random Forest Package, we would keep the generated trees and also set the maximum number of trees to 500. The most important variables in this model are teamElo, opptElo, teamOrtg,teamFIC, and teamEFG. Refer to Figure 19 for the variable importance graphic. The estimated error also decreases as the amount of trees increases and the optimal number of randomly selected predictors for each split was four. This is shown in the Appendix Figure 20.

2.11 Classification - Support Vector Machines

Our Support Vector Machine models utilized the same variables as the Classification Trees and Random Forest Classifiers. The same split of data was also used with eighty percent being training data and twenty percent being testing data. For each of our sets of averages, we created two SVM Models: a tuned model and an untuned model. To tune our SVM Models, we utilized the tune function in the e1071 package and set ranges for our Cost and Gamma Parameters in an effort to consolidate computational power and time. This tuning gave us the best model based on our estimated in-sample accuracy and this was then used to fit the tuned models.

2.12 Bias-Variance Trade-off

Techniques used to balance bias and variance included accounting for over-fitting by reducing number of predictors with minimal accuracy loss; limiting surface statistical redundancies by removing idiomatic variables in order to decrease model complexity; and diversifying regression types. Additionally, by identifying the proper number of games to average as mentioned in section 2.3, our models will account for more recent trends of various teams and account for the volatility throughout the season.

3. Results

3.1 Model Performance

The following section will analyse and summarize the results obtained throughout all the models that our team has tested and built throughout the course of this project. In this project we employed two main styles of prediction in our models. The first type of model we used were regression models. In our regression models we were essentially trying to

predict the points scored from each team within a given NBA game. From the points predicted for each team, we then used this to determine who won the game, naturally the team with the predicted higher points was the projected winner of the game. The second types of models we built were classification models. All of the models in this type were predicting a binary result of either a team winning or losing. As mentioned in section 2.3 the models we used for classification were Logistic, Classification Trees, Random Forest and Support Vector Machines. The metric we used to measure the success of the classification models was determining the number of matches it predicted the correct winner to obtain the accuracy of the model. The reason we decided to implement both regression models and classification models was to provide more variety in determining the best model to move forward with.

3.1.1 Regression Model Performance

Our team implemented a plethora of regression models to predict the scores of each game and we then used the prediction of the scores from the regression to determine the winner of the match (i.e. the team with the projected higher score will be the winner). Referring to the table below, we can see both the accuracy of the predicted winner of the match, as well as the in sample and out of sample r^2 and RMSE on the predicted points. From the table below, based on the model with the highest out of sample accuracy, we see the top performing regression model is the 15 Game Average GAM Model.

Table 2. Regression Model Accuracy

Model	In Sample Accuracy	Out Sample Accuracy	in Sample r^2	out of sample r^2	In Sample RMSE	Out of Sample RMSE
Null Model	50.32%	48.71%	-	-	12.26	12.09
5 Game Linear Model 1	66.29%	65.47%	0.1882	0.1683	11.03	11.13
15 Game Linear Model 1	66.98%	65.88%	0.2122	0.1715	10.87	11.09
25 Game Linear Model 1	66.99%	65.40%	0.1857	0.1684	10.90	10.99
5 Game Step Model 1	66.29%	65.43%	0.1875	0.1677	11.03	11.13
15 Game Step Model 1	66.74%	65.51%	0.2118	0.1712	10.88	11.09
25 Game Step Model 1	67.20%	65.15%	0.2064	0.2083	10.90	11.00
5 Game Lasso Model	66.14%	65.43%	0.1874	0.1676	11.03	11.13
15 Game Lasso Model	66.81%	65.51%	0.2121	0.1707	10.88	11.09
25 Game Lasso Model	66.70%	65.30%	0.2067	0.2057	10.90	11.02
5 Game Ridge Model	66.87%	65.18%	0.1861	0.1659	11.04	11.14
15 Game Ridge Model	66.67%	65.63%	0.2060	0.1661	10.92	11.12
25 Game Ridge Model	66.67%	65.20%	0.2024	0.2005	10.93	11.05
5 Game GAM Model	66.34%	65.36%	0.1949	0.1705	10.99	11.11
15 Game GAM Model	**66.88%**	**66.05%**	**0.2173**	**0.1709**	**10.84**	**11.09**
25 Game GAM Model	66.67%	65.20%	0.2144	0.2099	10.85	10.99

3.1.2 Classification Model Performance

For predicting whether a team should win or lose a specific game, our team implemented a variety of classification models. This section will touch on the accuracy and performance of our classification models. For a more detailed explanation on the manipulation of how each of the models were constructed, please refer to section 2.3 above. When trying to predict outcomes, it is important to implement a variety of different models to be able to compare the effectiveness and accuracy of each model. This allows us to be sure

that we are getting the most accurate predictions on the outcomes of NBA games. In the table below we have the in sample and out of sample accuracy for each of the models that we built. Since these are classification models, it does not make sense to use metrics such as R squared and RMSE as accuracy measurements. From the table below, we review the out of sample accuracy column to determine which model is the best classification model. As we can see, the 25 Game Average Logistic Importance Model had an out of sample accuracy of 66.99% which is the highest performing classification model that we built. Not only is this the best classification model, but it is also the best overall model that we generated.

Table 3. Classification Model Accuracy

Model	In Sample Accuracy	Out Sample Accuracy
5 Game Average Importance Logistic	65.65%	66.91%
15 Game Average Importance Logistic	66.14%	64.02%
25 Game Average Importance Logistic	**65.65%**	**66.99%**
5 Game Average RFE Logistic	65.54%	66.84%
15 Game Average RFE Logistic	65.77%	64.81%
25 Game Average RFE Logistic	65.54%	66.03%
5 Game Average Classification Tree	63.67%	64.7%
15 Game Average Classification Tree	65.51%	63.6%
25 Game Average Classification Tree	66.69%	62.38%
5 Game Average Random Forest - CARET	63.94%	63.29%
15 Game Average Random Forest - CARET	63.79%	64.56%
25 Game Average Random Forest - CARET	63.75%	63.35%
5 Game Average Random Forest - RF Package	62.66%	64.45%
15 Game Average Random Forest - RF Package	63.73%	62.85%
25 Game Average Random Forest - RF Package	63.47%	63.26%
5 Game Support Vector Machines untuned SVM	68.17%	65.24 %
15 Game Support Vector Machines untuned SVM	68.42%	65.65%
25 Game Support Vector Machines untuned SVM	68.39%	64.98 %
5 Game Support Vector Machines tuned SVM	96.24%	62.71 %
15 Game Support Vector Machines tuned SVM	94.11%	62.43%
25 Game Support Vector Machines tuned SVM	96.24%	55.69 %

From the table above, we see the performance of every classification model that we tested on our data. The row that is emphasized in bold above illustrates our best model. The best model is the 25 Game Average Importance Logistic Model. This model was chosen as the best model because it has the highest out of sample accuracy of 66.99%. This means we expect this model to predict the correct outcome of an NBA game approximately 66.99% of the time. Looking at the in sample predictions for the 3 SVM models, the accuracy results are in the 90th percentiles, however the out of sample accuracy contains the lowest values out of all the models that we tested. This is likely due to the model over fitting to the training data, and in turn performing much more poorly on the out of sample data. Referring to Figure 10 below, for the in sample plot, we can see that our best model predicted a team would lose 3,925 times and the team did lose (correct prediction), we also see our model predicted a team would win 2,059 times when the team actually lost (incorrect prediction). Next our model predicted the team would lose 2,054 times when they actually won (incorrect prediction). Lastly, our model predicted a win 3,920 times when the team actually won (correct prediction). The out of sample accuracy plot in Figure 10 can be interpreted in the same way.

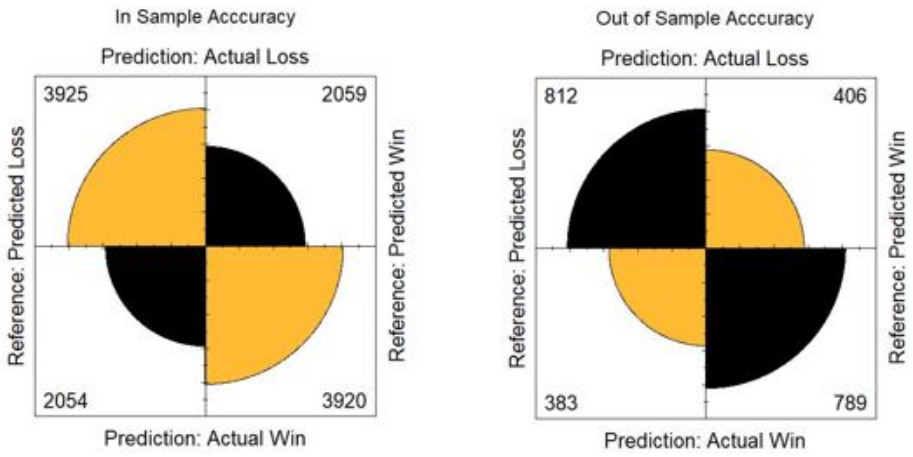

Fig. 10. Actuals vs Predicted For 25 Game Average Importance Logistic

3.2 Inferencing

The variable that our final model attempts to predict is **TeamRslt**. This is a classification model, so it is determining a binary result. To reiterate, TeamRslt is a classification variable that takes on one of two potential values, TeamRslt equals 1, when the team in question loses the game, and 2 when the team wins the match in question. Therefore this variable is Binary. Based on the methods discussed above during variable selection, it was determined that our best predictors were the following: **teamElo**, **opptElo**, **teamOrtg**, **teamFIC**, **teamEFG**. As mentioned in section 2.8 our final model implored the use of Variable Importance to determine which variables were the best for predicting the outcome in our Logistic Classification. Let us examine graphically how each of these predictor variables correlates on a general level to the variable teamRslt.

3.2.1 teamElo vs. teamRslt

Since teamElo Rating is a calculation that measures the strength of a team, one would expect to see a higher Elo Rating correlated to a higher chance of winning. We can refer to CD plot in the appendix that shows exactly this relationship.

As shown in the top right corner of Fig. 11, we can see the higher the teamElo Rating, the higher the density of 2 values we get as TeamRslt and 2 corresponds to winning. Therefore our hypothesis is accurate that a higher Elo Rating corresponds to more wins.

3.2.2 opptElo vs. teamRslt

Similar to the previous section, if your opponent has a high Elo Rating, you would expect that they are strong, and you are more likely to lose. We can confirm this relationship with another density plot. As we can see in the top middle of Fig. 11, our hypothesis is true, the density of teamRslt being equal to 1 grows as the opponent Elo Rating increases.

3.2.3 teamOrtg vs. teamRslt

A higher Team Offensive Rating typically corresponds to a team scoring more points as usually the more points your team scores, the more likely you are to win. This is again shown with the conditional density plot as shown in the bottom middle of Fig. 11.

3.2.4 teamFIC vs. teamRslt

We can prove that teamFIC and teamRslt are positively correlated with the density graph in the bottom left of Fig. 11. This shows how accurate and useful formulaic statics in the NBA can be at predicting outcomes.

3.2.5 teamEFG vs. teamRslt

TeamEFG. stands for Effective Field Goal Percentage. We expect effective field goals to be positively correlated with teamRslt. And this is proven to be correct from the top right of Fig. 11.

4. Future Steps

There are a variety of aspects to the this project that we as a group hope to explore moving forward if we choose to pursue the outcomes of this project beyond the scope of IE590. This section attempts to describe the goals that we would hope to accomplish when moving forward with this project. One incredibly interesting aspect of our project is that we are currently in the middle of the 2021 NBA season. This means that every single day we get new games that we can try to predict using our model, and test it in 'real time' Moving forward, we would set up our model to run and predict the games of every NBA game per day, and based on the results on a day to day basis we would continue to adapt and update our model. We ran our best predictive model, the 25 Game Average Importance Logistic, for the games played on **April 25 2021**. There were 7 games played on April 25, We ran each of the games through our predictive model. Refer to Table 4 below to see the results of what our model predicted vs the actual outcome of the game.

Table 4. April 25 2021 NBA Game Predictions

Team	Opponent	Predicted Winner	Certainty	Actual Winner	Correct?
Sacramento Kings	G.S. Warriors	G.S. Warriors	64.25%	G.S. Warriors	**YES**
Indiana Pacers	Orlando Magic	Indiana Pacers	75.1%	Indiana Pacers	**YES**
Milwaukee Bucks	Atlanta Hawks	Miluakee Bucks	56.6%	Atlanta Hawks	**NO**
Cleveland Cavaliers	Wash. Wizards	Wash. Wizards	61.5%	Wash. Wizards	**YES**
Memphis Grizzlies	Trail Blazers	Memphis Grizzlies	66.9%	Memphis Grizzlies	**YES**
Phoenix Suns	Brooklyn Nets	Phoenix Suns	49.9%	Brooklyn Nets	**NO**
Charlotte Hornets	Boston Celtics	Charlotte Hornets	57.0%	Charlotte Hornets	**YES**

From the table above, it is shown that we got 5 out of 7 predictions correct. This is around 70% and relatively close to the accuracy of our final selected model. As a next step,

our team would set up an automated script to run everyday. The script would automatically scrape the data from the NBA Stats website and run the model with those stats. It would then save the results of the prediction into a file to be compared to the actual results.

We currently have tested and built our models using the average stats across a teams previous 5, 15, or 25 games. A potential next step would be to add weights to the averages across previous games. For example the previous 5 games stats could be weighted with 50% importance, the previous 15 games could be weighted with 30% and the previous 25 games could be weighted with importance of 20%. This would take a lot of testing and analysis to get the proportions correct. This also could help with accounting for a teams large losing or winning streak, while also accounting for a teams long term performance on the season.

The last thing our team is planning to work on moving forward is to implement analysis based on our model predictions and given odds to determine the ideal wagers to place that will maximize profits when placing a bet. To do this hopefully we can get current odds per game, use those odds to develop a function that takes our predicted result, compares it to Vegas odds and returns whether or not a wager is reasonably safe.

5. Conclusion

In conclusion, Sports betting is a multi billion dollar industry across the United States where gamblers and statistician try to predict outcomes of events with the most statistical accuracy as possible. Our report has taken an in depth look at the NBA data to identify relationships that are most important and significant in predicting the outcomes of NBA games. We have accurately outlined the different models that we tested and identified our most significant model to be the 25 Game Average Importance Logistic model. As a team, we were able to identify and build upon previous similar analysis done by other professionals in the field and as a result our model is currently predicting with just about **66.99%** accuracy. Although there is still work to be done moving forward, are team has made excellent progress and we are comfortable and confident in the legitimacy and accuracy of the analysis presented within this report. In the end, our team was able to meet our goal of creating a predictive model that would predict NBA games with approximately 67% accuracy. Although we did not out perform many of the studies that are currently published today, our final model had an accuracy in the high 60's percentile, which seems to be the benchmark average for these types of projects.

6. Individual Contributions

6.1 Jakson Terpak

- Section 3,4, and 5 of report
- using and tracking results of models on current (2021) NBA matches (Ongoing).
- Script for calculating averages stats across X Previous games
- NULL model, Visualizations for multiple models and results
- various contributions to project powerpoint

6.2 Benjamin Sons

- Section 2 of the report
- Script for running all regressions that use points as the response variable
- Initial data research, variable selection, and visualizations

6.3 Austen Horton

Contributions from Austen include the development of Classification Models: Logistic Regression, Classification Tree, Random Forest, Support Vector Machine. Contributions also include the Executive Summary, Introduction Section, and Data Method Section for models specified.

References

Dorson, J. R. (2020). What is paspa, the federal ban on sports betting? https://sportshandle.com/what-is-paspa-sports-betting-ban-professional-amateur-sports/.

Fayad, A. (2020). Building my first machine learning model— nba prediction algorithm. https://towardsdatascience.com/building-my-first-machine-learning-model-nba-prediction-algorithm-dee5c5bc4cc1.

Fischer-Baum, N. S. . R. (2015). How we calculate nba elo ratings. https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/.

Formal, A. (2012). Understanding the nba: Explaining advanced offensive stats and metrics. https://bleacherreport.com/articles/1039116-understanding-the-nba-explaining-advanced-offensive-stats-and-metrics.

Hughes, G. (2014). Does pace matter in the nba? https://bleacherreport.com/articles/2209761-does-pace-matter-in-the-nba.

Jacobs, J. (2017). Squared statistics: Understanding basketball analytics. https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/.

Kaplan, S. and Garrick, B. J. (1981). On the quantitative definition of risk. *Risk analysis*, 1(1).

Nasser, K. (2016). Predicting the outcome of nba playoffs based on the maximum entropy principle.

Pinnacle (2018). The benefits of using floor impact counter in the nba. https://www.pinnacle.com/en/betting-articles/Basketball/floor-impact-counter-explanation/34L28L6QWDDUP8UT.

Purdum, D. (2020). Sports betting's growth in u.s. 'extraordinary'. https://www.espn.com/chalk/story/$_{/}id/29174799/sports-betting-growth-us-extraordinary$.

Rossotti, P. (2019). Nba enhanced box score and standings (2012 - 2018). https://www.kaggle.com/pablote/nba-enhanced-stats.

Unknown (2020a). Basketball rebounds. https://www.rookieroad.com/basketball/stats/rebounds/offensive-rebounds-in-basketball.

Unknown (2020b). Basketball steal. https://www.rookieroad.com/basketball/101/steals/.

Weiner, J. (2021). Predicting the outcome of nba games with machine learning. https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20.

Winkler, E. (2020). What is a turnover in basketball? https://dunkorthree.com/turnover-basketball/.

Zhang, L. (2019). What is a field goal in basketball? https://dunkorthree.com/field-goal-basketball/.

Appendix A: Supplemental Materials

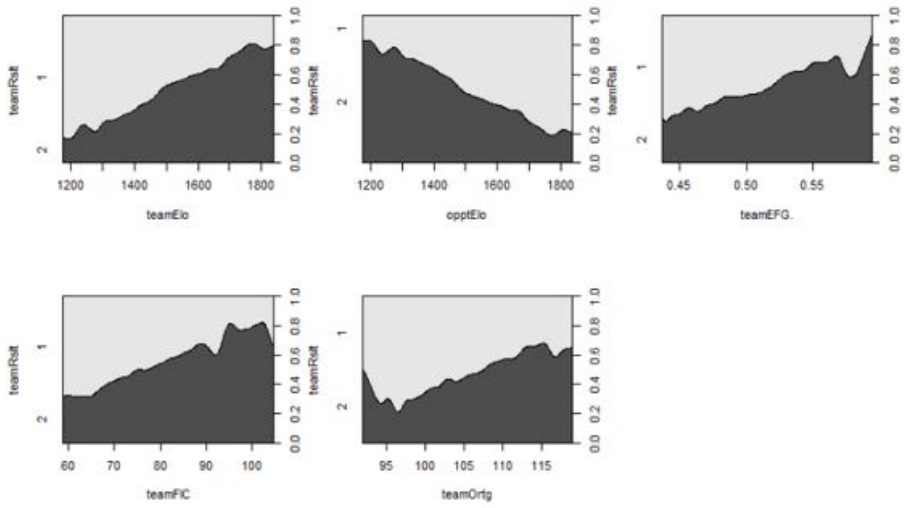

Fig. 11. Conditional Density Plots

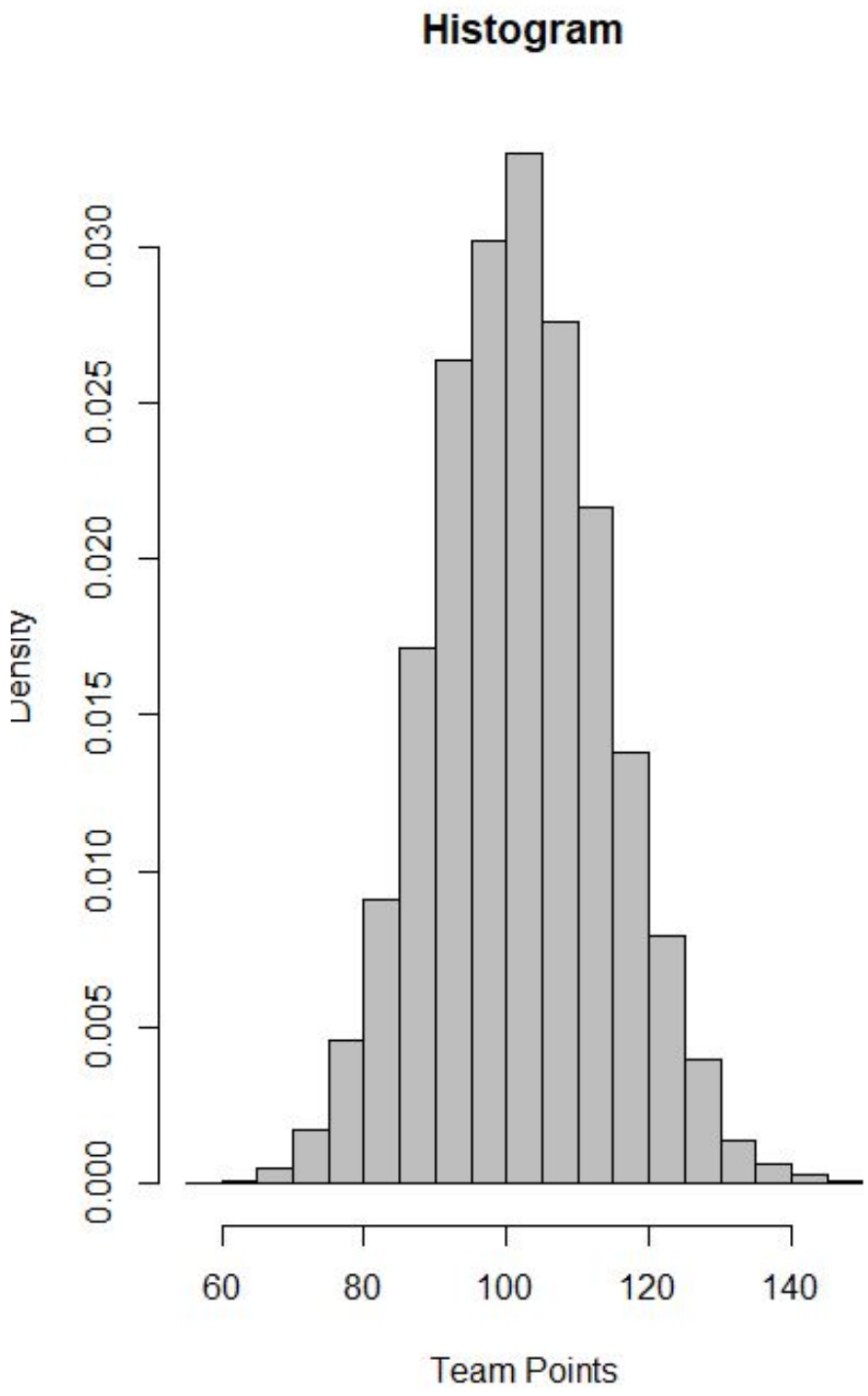

Fig. 12. Histogram Illustrating Density Of Points

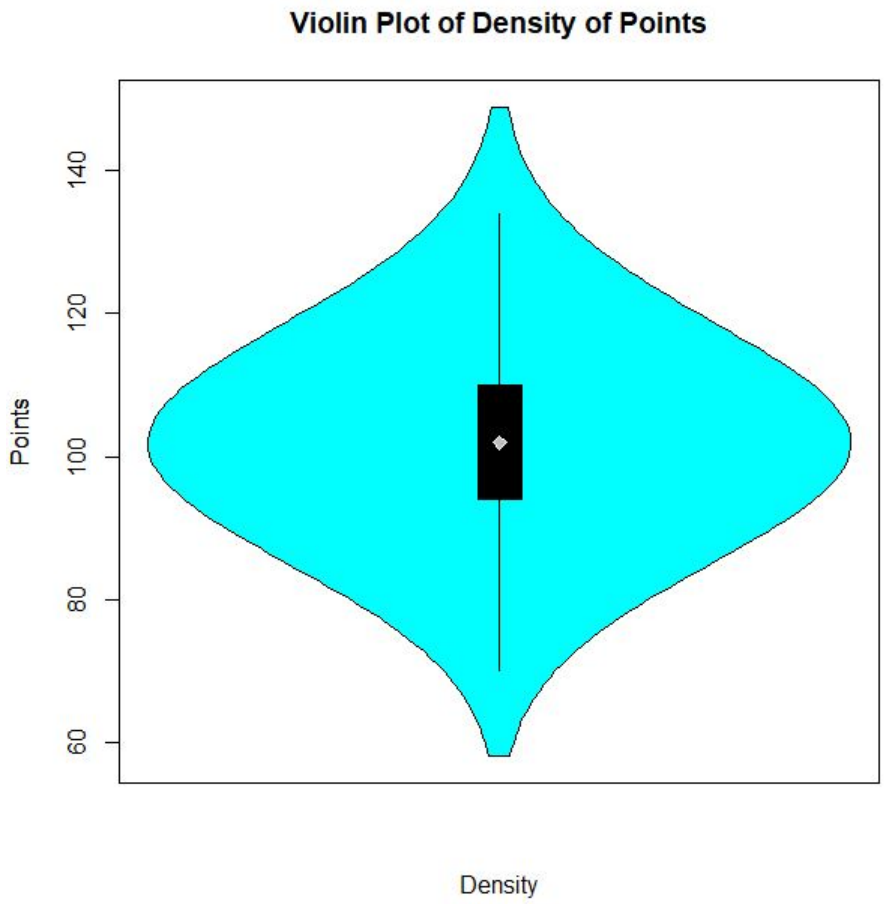

Fig. 13. Violin plot Illustrating Density Of Points

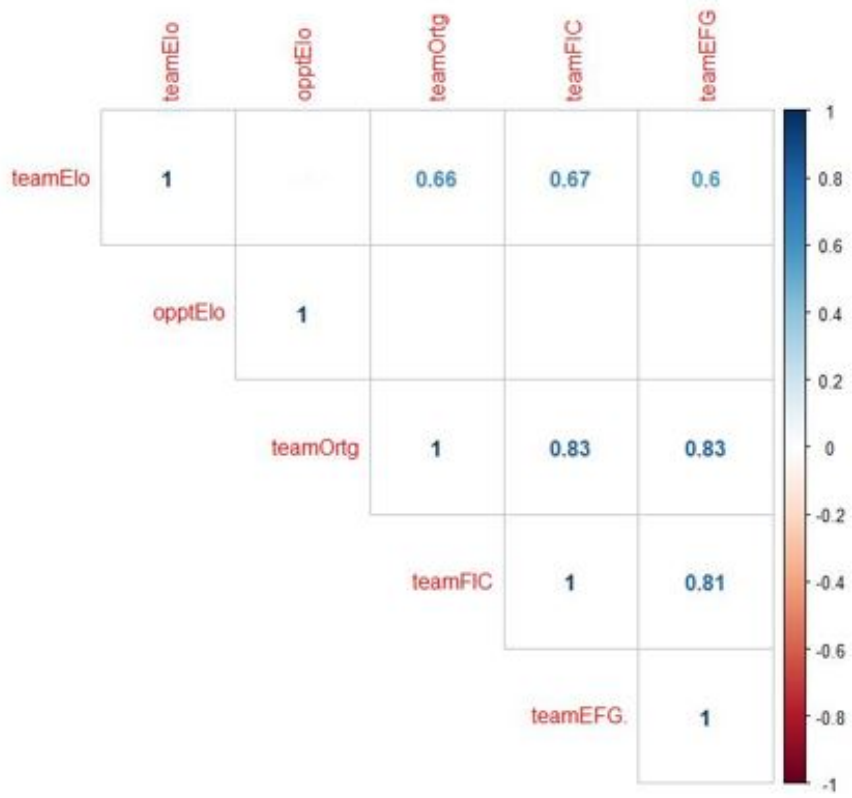

Fig. 14. Correlation Plot of Variables used in the final Model

Variable	Inflation Factor
TeamElo	1.811467
OpptElo	1.017732
TeamOrtg	4.395658
TeamFIC	3.850462
TeamEFG	3.563278

Fig. 15. The Variable inflation factors of the final selected model

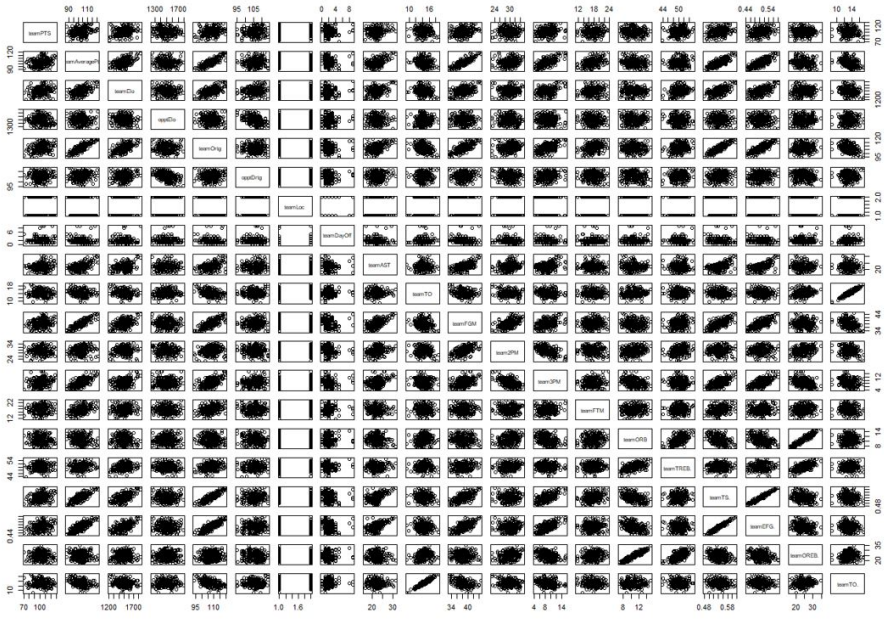

Fig. 16. Pairs Plot Showing relationships between variables of Initial Dataframe

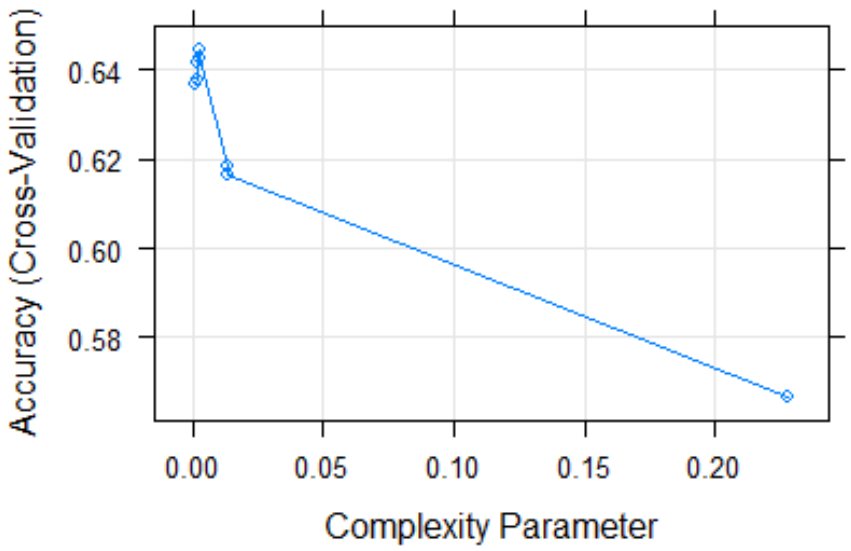

Fig. 17. Complexity Parameter for the 5 Game Average Cart Model

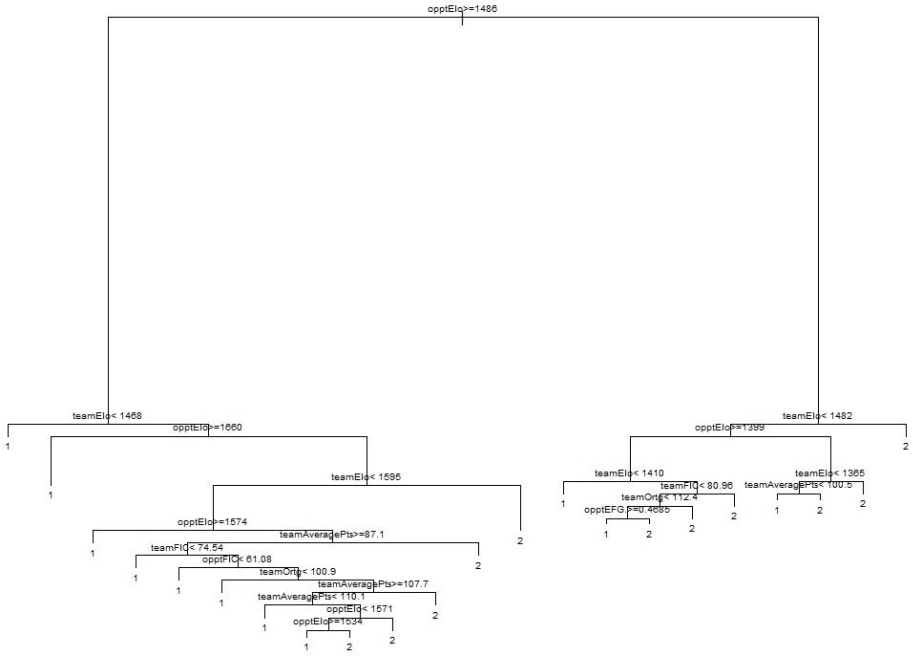

Fig. 18. Tree Model Diagram for 5 Game Average Cart Model

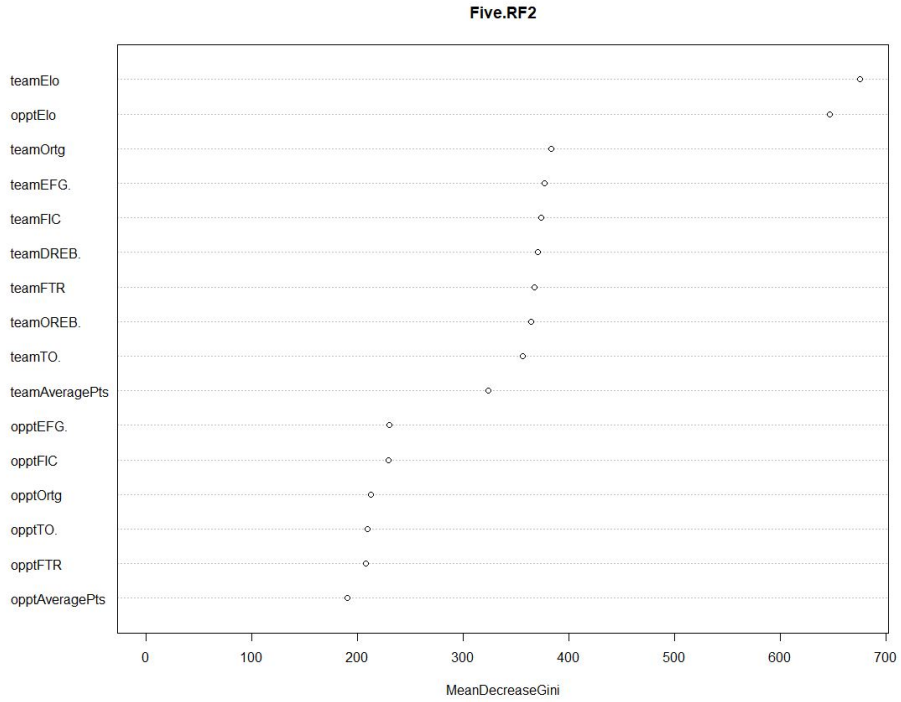

Fig. 19. Importance Variable ranking For 5 Game Average Random Forest

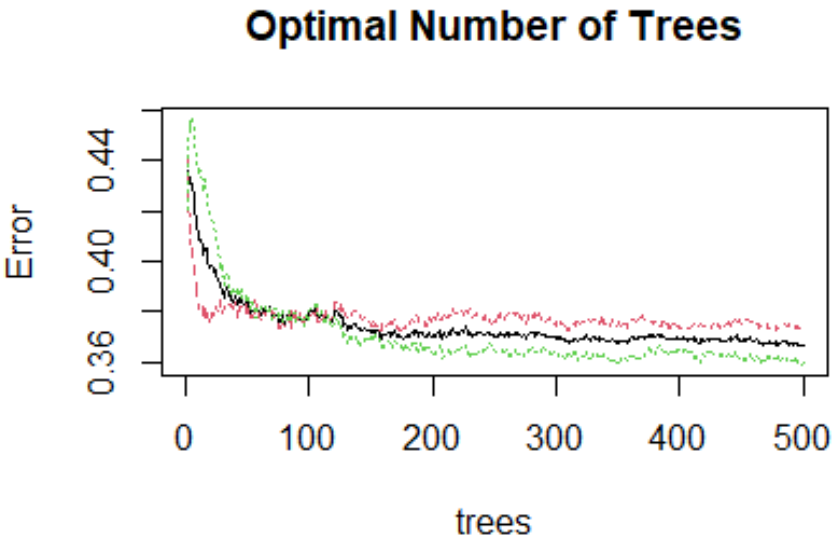

Fig. 20. Optimal Number of Trees in For 5 Game Average Random Forest

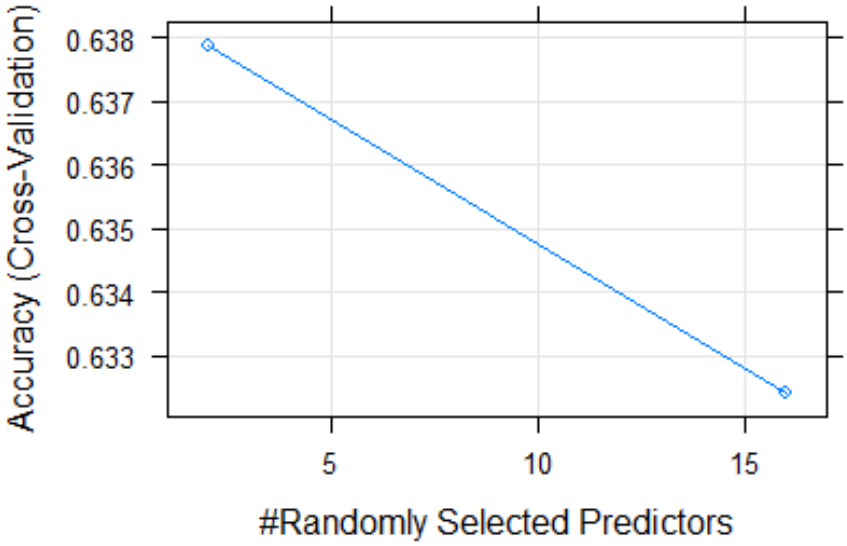

Fig. 21. Accuracy vs Number of Predictors in 5 Game Average Random Forest