UNIVERSITY OF CALIFORNIA
Los Angeles

Statistical Analysis and Predictive Modeling in Basketball: Unveiling Key Variables for Championship Success

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics & Data Science

by

Garvyn Jay Chua

2023

ABSTRACT OF THE THESIS

Statistical Analysis and Predictive Modeling in Basketball: Unveiling Key Variables for Championship Success

by

Garvyn Jay Chua
Master of Applied Statistics & Data Science
University of California, Los Angeles, 2023
Professor Frederic R. Paik Schoenberg, Chair

The game of basketball has witnessed constant evolution, necessitating the use of statistical data for predicting winners. While the odds of winning a championship are traditionally 1 in 30 each year, strategic positioning in various statistical categories can surpass this baseline, regardless of a team's annual ranking. By analyzing data and identifying crucial variables, it becomes evident that basketball outcomes are not predetermined. This study employs modern data science methods to make predictions for future years or generations, emphasizing the importance of selecting accurate via cross-validation for the machine learning (ML) techniques. All data used in this study is sourced from Basketball-Reference. This study debunks the fallacy that a team solely relying on prominent three-point shooters guarantees championship success. It underscores the significance of other factors and variables in determining outcomes. The analysis emphasizes the importance of relying on objective statistical analysis rather than subjective perceptions. Factors like overtime play and defensive prowess in blocks significantly impact a team's likelihood of becoming an NBA champion, reinforcing the need for data-driven insights and accurate predictions.

The thesis of Garvyn Jay Chua is approved.

Michael Tsiang

Miles Satori Chen

Chad J. Hazlett

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Background

The game of basketball has always been a continuously evolving sport, leading to the utilization of various statistical data to potentially predict the winner. In reality, each year there is a 1 in 30 chance of winning the championship. However, if teams strategically position themselves in multiple statistical categories, their odds of success could surpass the baseline of 1 in 30, irrespective of their ranking in a given year. By analyzing statistical data and exploring key variables that offer the best odds, it becomes apparent that the outcome is not predetermined. Upon examining the data, it becomes necessary to employ modern "Data Science" methods to make predictions for future years or generations. The selection of the most accurate machine learning (ML) method by way of cross-validation, becomes crucial in this decision-making process. All data used in this study was obtained from Basketball-Reference.

Among the numerous changes in the game, one particular change stood out—the introduction of the "three-point" shot [**?**]. Stephen Curry of the Golden State Warriors played a pivotal role in revolutionizing the game by significantly increasing the average points scored per game since its implementation in the late 1970s. This revolution encompassed several rule changes throughout the game's recent history [**?**]. These changes included providing more space for shooters and a shift away from physicality, as the game moved from the rim to the extended three-point line. Machine Learning is predominantly employed in sports for

tasks like determining draft picks and projecting players' productivity and salaries. Therefore, the models used in this study were inspired by a previous research conducted at Bryant University [**?**].

1.2 Purpose

Simply stated, the motivation behind this report stems from a deep passion for both the game of basketball and my favorite academic subject. Combining these two interests has prompted me to delve into the realm of basketball statistics, ranging from basic metrics like points, rebounds, and assists to more advanced indicators such as win-shares and player efficiency rating. As an ardent fan, exploring these statistics has been truly captivating. It is worth noting that the statistical data used in this study aligns with commonly known categories, which should be considered common knowledge.

Interestingly, the results of this study may come as a surprise to some. Upon closely observing the game of basketball, it became apparent that the factors I initially believed would determine the championship winner might not be as decisive as expected. It cannot be predetermined that teams must excel in numerous statistical categories (which will be specified in the concluding section of the report).

The study that served as my inspiration was a 2012 research conducted at Carnegie Mellon titled "Predicting NBA Championship by Learning from Historical Data" by Jackie B. Yand and Ching-Heng Lu [**?**]. Since the study was more than a decade old, I aimed to add my own unique contributions while incorporating modern statistical methods used in Data Science today. This study involved performing advanced statistical analysis on well-known data, and through the implementation of Cross-Validation, achieved an accuracy rate of approximately 86.75%. My goal was to develop models that could surpass or reach the 90% threshold.

CHAPTER 2

Methodology

The data utilized in this study was gathered and extracted from Basketball-Reference.com. The entire study was conducted using the R programming language, employing tools and techniques taught within the Statistics Department at the University of California, Los Angeles (UCLA). With over 30 available variables for analysis, the focus was on exploring the general patterns and narratives that highlight the evolving nature of the game and the crucial characteristics of NBA champions. The selection of the best model at the conclusion of the study was based on achieving the highest accuracy and is recommended for future investigations.

The Exploratory Data Analysis phase of this case study involved the use of histograms, box plots, and scatter plots to visualize the variations within the data. The goal was to determine whether the data exhibited a "normal" distribution or any other type of variation. Ideally, a normal distribution is preferred for generating random outcomes in the Machine Learning section of the study. Assumptions were made regarding scoring and defensive categories to facilitate the prediction aspect of the study.

The prediction phase of the study employed six Machine Learning techniques: Logistic Regression, Random Forests, XGBoost, Principal Component Analysis (PCA), Support Vector Machines (SVM), and Neural Network. Most of these techniques were classification methods used to further infer the key characteristics of a future championship team. It is important to note that the statistics analyzed in this experiment primarily focused on team data rather than individual player data. For the supervised learning methods, cross-

validation was used to avoid over-fitting and to create the partitioned data sets per model for the most robust results. An outlier analysis was conducted to strengthen the understanding of the factors necessary for achieving the highest level in this sport.

CHAPTER 3

Exploratory Data Analysis (EDA)

A commonly held assumption regarding the determinants of a successful season, and particularly a championship, is the analysis of a team's average points per game. Since the primary objective in each game is to outscore the opposing team, it is pertinent to examine the frequency distribution of average points. Figure 3.1 illustrates a fairly normal distribution of average points dating back to the year 2000.

In Figure 3.2, the histogram displays a right-skewed distribution of three-pointers made, indicating that teams tend to average around 7 successful three-point shots per game. However, this can be considered an anomaly in today's game, as individual players can often attempt 7 three-pointers on their own.

Moving on to Figure 3.3, free throws exhibit a more normal distribution, with the majority falling within the range of 15 to 20 made free throws. This variation may be influenced by factors such as the frequency of foul calls made by referees over the years, whether there are excessive or insufficient calls. All three of these variables exhibit a rather normal distribution from an Anderson-Darling test with P-Values all less than 0.05.

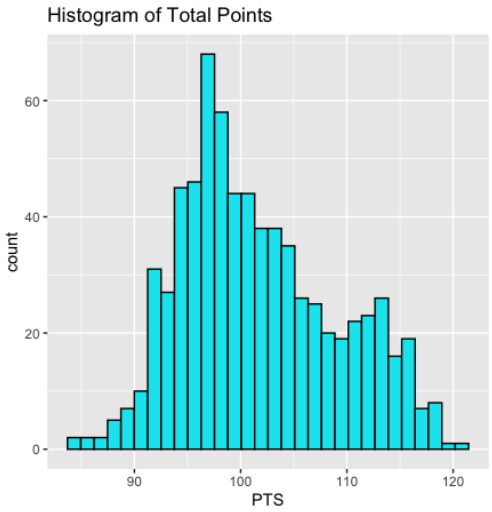

Figure 3.1: Histogram of PTS

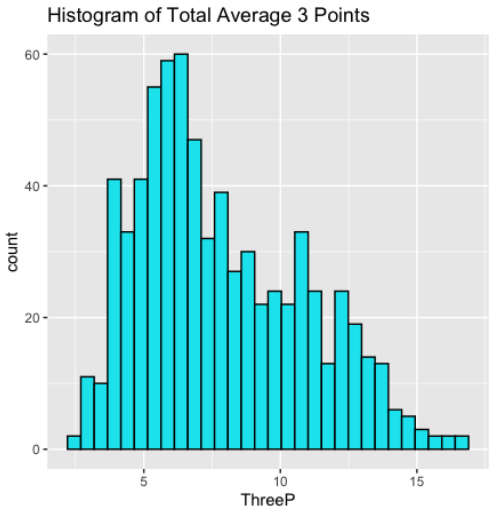

Figure 3.2: Histogram of 3PTS

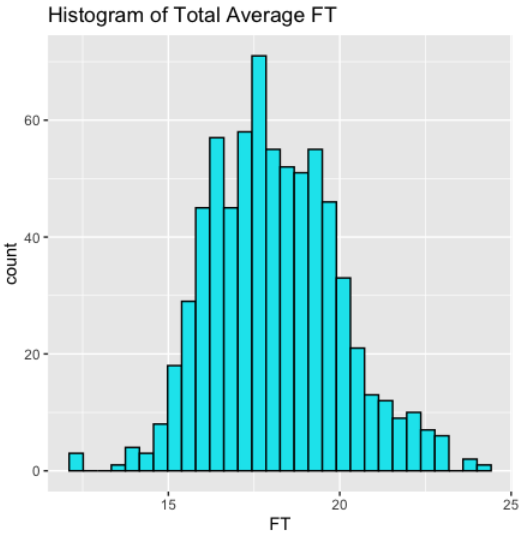

Figure 3.3: Histogram of FT

Delving further into offensive categories commonly believed to define the characteristics of a championship team, namely Average Points, Average Three-Pointers Made, and Average Free Throws Made, an analysis was conducted across NBA seasons in the 21st century. Figure 3.4 presents a scatter plot with a trend line depicting the average points scored over the past 23 years. Similarly, Figure 3.5 illustrates the trend of three-pointers made during

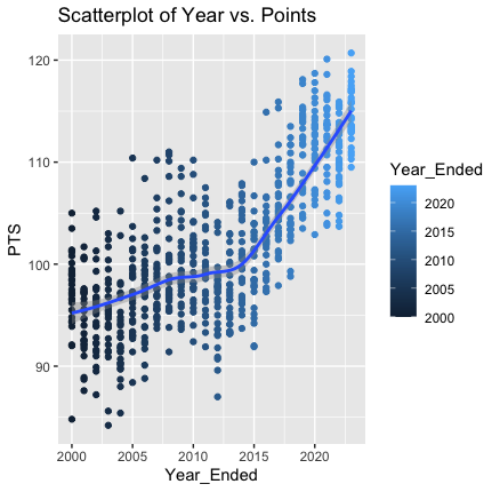

Figure 3.4: Scatterplot of Yr vs. PTS

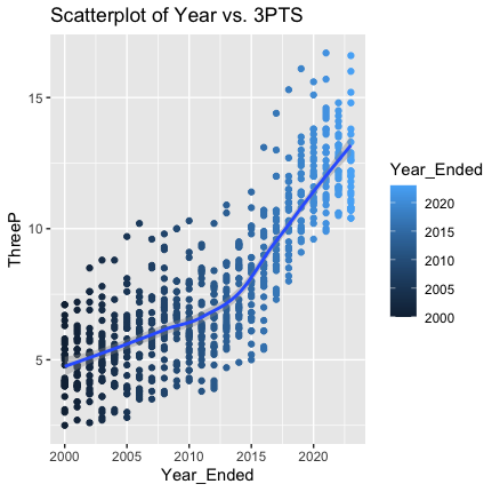

Figure 3.5: Scatterplot of Yr vs. 3PT

the same period. Both graphs exhibit a positive trend, indicating that as time progresses, average points and three-pointers are on an upward trajectory, potentially contributing to a team's success in reaching the pinnacle of achievement.

Conversely, Figure 3.6 demonstrates a decline in free throws, solidifying the notion that the game is shifting towards a perimeter-centric style of play, with less emphasis on contact generated from three-point attempts and overall points derived from them. The scatterplots all have a 'LOWESS' or 'a 'Locally Weighted curved line Scatter plot Smoothing relationship", which is a type of regression analysis that identifies the trends within the data non-linearly. In Figure 3.6, depicts a steep decline from 2008 - 2018 in free throws made and a steady increase seems to appear; largely due to rule changes to shooter's space and no "hand-checking" on an opponent.

Conducting a thorough analysis, we begin by assessing the relationship between offensive metrics and the likelihood of winning a championship. For this purpose, a binary classification was employed, assigning '0' to teams that did not win a championship and '1' to those that emerged victorious. In Figures 3.7, 3.8, and 3.9, we present the corresponding box plots for selected offensive metrics.

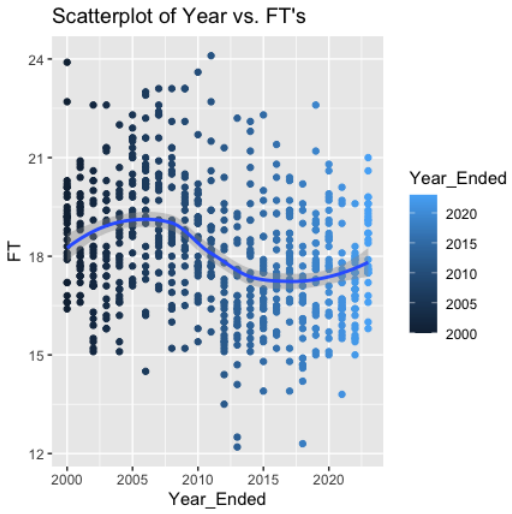

Figure 3.6: Scatterplot of Yr vs. FT

Figure 3.7 highlights the disparity in points scored, with the box plot revealing that championship-winning teams tend to average over 100 points per game throughout an entire season. On the other hand, Figure 3.8 demonstrates that the upper quartile of championship teams typically make more than 10 three-pointers, while their non-championship counterparts generally have an average of fewer than 10 three-pointers made. It suggests that three-point shooting proficiency plays a significant role in determining a team's championship potential.

Figure 3.9, however, presents an inconclusive box plot concerning free throws made, indicating that there isn't a substantial difference between non-championship contenders and actual winners in this category. It suggests that teams may rely on other offensive or possibly defensive aspects to secure their championship victories.

Overall, these box plots provide initial insights into the offensive metrics associated with championship success, emphasizing the importance of scoring high points and excelling in three-point shooting, while indicating that the impact of free throws may be less significant.

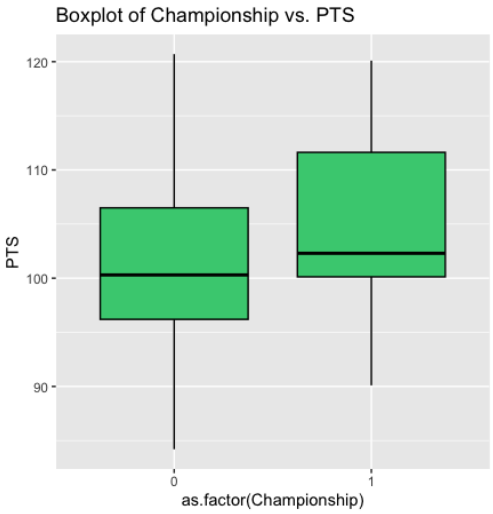

Figure 3.7: Boxplot of Win vs PTS

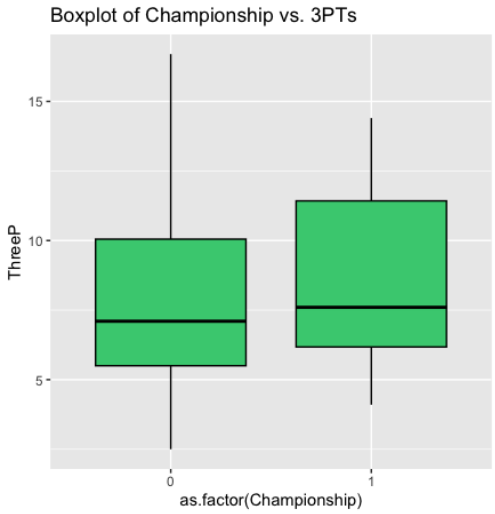

Figure 3.8: Boxplot of Win vs 3PTS

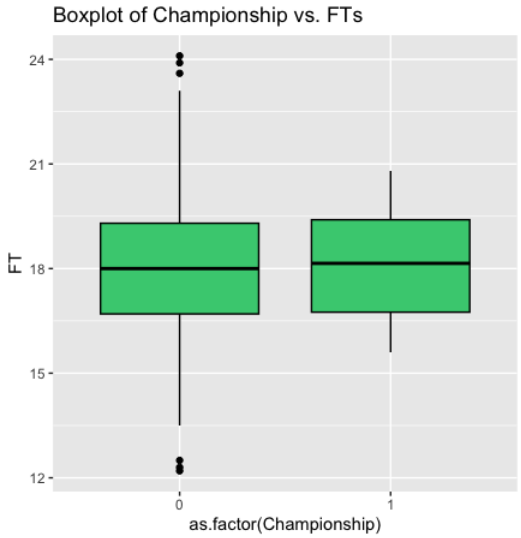

Figure 3.9: Boxplot of Win vs FTs

CHAPTER 4

Machine Learning Methods and Models

4.1 Applications of Logistic Regression in NBA Data

In Logistic Regression, we use the following equation below:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$$

The equation provided encompasses several components. Firstly, $P(Y = 1|X)$ represents the probability of the dependent variable Y being equal to 1, considering the values of the independent variables $X_1, X_2, \ldots, X_n$. Secondly, the coefficients or parameters $\beta_0, \beta_1, \ldots, \beta_n$ are estimated by the logistic regression model and contribute to the calculation of this probability. Lastly, $X_1, X_2, \ldots, X_n$ refer to the independent variables or features included in the equation. It is important to note that this equation specifically pertains to binary logistic regression, where the dependent variable Y takes on two potential values, such as 0 and 1. In cases involving multinomial logistic regression or other variations, the equation may need to be modified to suit the specific context and requirements of the model.

In Table 4.1 below, the initial model is presented, featuring some variables commonly used to assess a team's performance in the present day. The response variable is the championship outcome, while the remaining variables were selected based on their ability to be easily understood and interpreted by a broader audience from the available pool of 28 variables. Half of the variables fall under the scoring category, while others reflect various characteristics of the game itself, such as minutes played, playoff participation (0/1), turnovers, and defensive

aspects like steals and blocks. This initial logistic regression model serves as a starting point for further refining the final model.

Variable	Estimate	Std. Error	z value	Pr(z)
(Intercept)	134.40	1058.0	0.13	0.90
Playoffs	17.67	1055.00	0.02	0.99
FT	-0.11	0.15	-0.72	0.47
Minutes	-0.68	0.34	-2.00	0.0453*
FG	0.66	4.79	0.14	0.89
ThreeP	-0.69	4.79	-0.15	0.89
TwoP	-0.64	4.78	-0.13	0.89
TRB	0.11	0.15	0.74	0.46
AST	0.09	0.16	0.56	0.57
STL	-0.01	0.30	-0.02	0.99
BLK	0.63	0.30	2.07	0.0381*
TOV	0.09	0.29	0.32	0.75

Table 4.1: Table of Initial Logistic Regression Model Output

Table 4.2 presents a refined version of the model depicted in Table 4.1. The final logistic regression model includes only two variables, in addition to the constant response variable (Championship (0/1)). As observed in the previous table, this iteration reignites the discussion surrounding the correlation between scoring and a team's ability to win a championship. Notably, both minutes and blocks exhibit an α value of 0.5, which surpasses the designated significance level (P-values). This phenomenon will be further explored and utilized in subsequent analyses and machine learning models within the report.

Given that the response variable is binary, it is essential to examine various metrics to evaluate the logistic regression model's performance. Two important metrics to consider

Variable	Estimate	Std. Error	z value	Pr(z)
(Intercept)	146.92	72.89	2.02	0.04
Minutes	-0.64	0.30	-2.12	0.03
BLK	0.88	0.27	3.23	0.00

Table 4.2: Table of Final Logistic Regression Model Output

are sensitivity and specificity. Sensitivity represents the probability of the model accurately predicting a positive outcome, while specificity measures the probability of correctly predicting a negative outcome. To assess both of these metrics simultaneously, the AUC (Area Under the Curve) is introduced. A higher AUC value, ranging from 0 to 1, indicates better predictive performance.

In Figure 4.1, a reliable predictor for distinguishing true positive and true negative outcomes is depicted. The ROC (Receiver Operating Characteristic) curve serves as one of many metrics available for evaluating binary classification in logistic regression techniques.

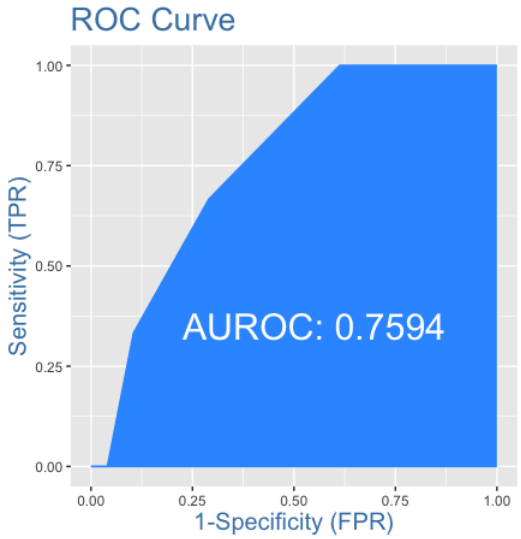

Figure 4.1: Logistic Regression (AUC)

4.2 Implementing Random Forests for Prediction

In Random Forest, we have the following equation below:

$$F(X) = \frac{1}{N} \sum_{i=1}^{N} f(X, \Theta_i)$$

In the given scenario, we are working with a Random Forest model that consists of several key components. The function $F(X)$ represents the prediction produced by the Random Forest when provided with a specific input X. The variable N indicates the total number of trees within the forest. Moreover, $f(X, \Theta_i)$ denotes the prediction generated by the ith tree in the forest, with Θ_i representing the parameters associated with that particular tree. These elements collectively contribute to the prediction process of the Random Forest model. By aggregating predictions from multiple trees, the model can yield robust and accurate outcomes.

Table 4.3 provides insights from multiple runs, indicating that the optimal number of trees is 114. Initially, with $N = 500$, the lowest Mean Squared Error (MSE) was achieved at a value of three. Mean Squared Error[?] is the average of the squared differences between the predicted values and the actual values. Minimizing this error helps reduce risk and enhances the reliability of our prediction models. The Root Mean Square Error (RMSE), which is the square root of the MSE, is at a value of 0.18. Furthermore, the 27 random predictors represent the optimal number of predictors required to achieve the optimal RMSE, as illustrated later in the report.

Model	Trees	Number of Trees MSE	RMSE	Random Predictors
Random Forest	114	3	0.18	27

Table 4.3: Table of Optimal Values for Random Forests

Figure 4.2 provides a clear illustration of the optimization process. Initially, the model is

set with $N = 500$, but the graph demonstrates that the optimal performance can be achieved with $N = 114$. At the beginning, the graph exhibits significant variability, but as the model progresses, it begins to stabilize. Although running the model for a longer duration may result in even lower errors, it is advisable to stop at $N = 114$ to balance computational time and efficiency.

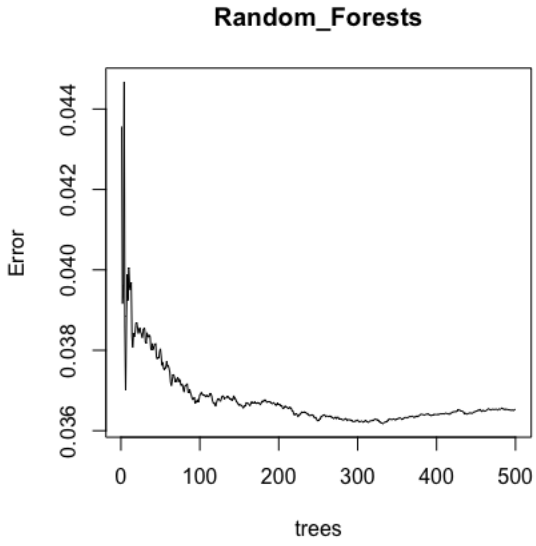

Figure 4.2: Random Forests: 'N' Trees

Figure 4.3 presents the variable importance graph obtained from an R[**?**] package. This graph is based on a conditional approach for Random Forests (RF) and highlights the importance of variables derived from the logistic regression binary classification model. The graph indicates that the variable "Minutes" falls within the range of two to three, while "Blocks" falls within the range of five to six. The variable importance graph emphasizes that among the small group of predictors, "Blocks," which represents a defensive statistic, holds greater importance than "Minutes," which measures playing time.

In Figure 4.4, the RF graph illustrates the Out-Of-Bag (OOB) error on the y-axis, which represents the prediction error of the Random Forests model. The x-axis, labeled as 'm-try,' corresponds to the number of nodes. In this particular model, the lowest OOB error or the

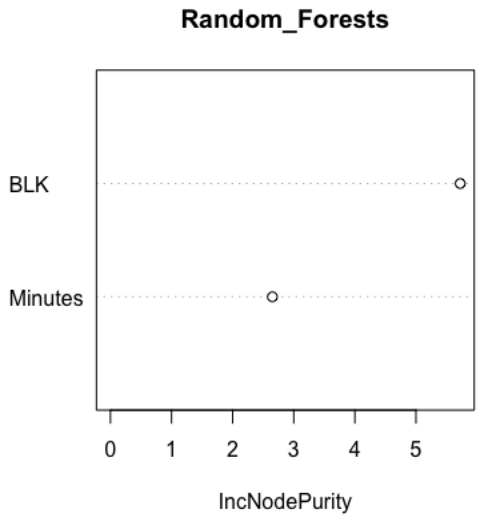

Figure 4.3: Variable importance for RF

highest predictive power is observed at 27 nodes. This finding demonstrates that Random Forests (RF) is a highly effective approach for predicting a team's chances of winning the championship, surpassing the performance of the logistic regression model.

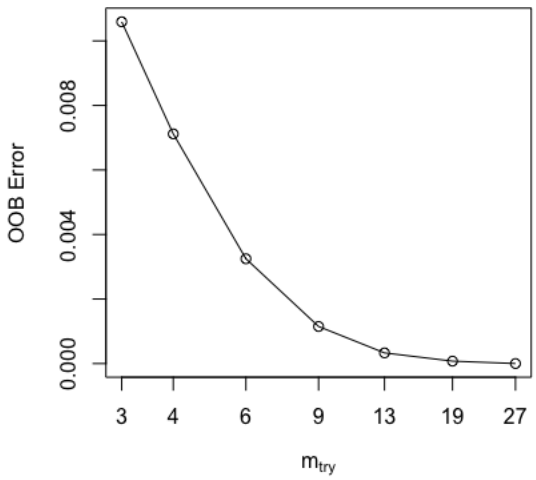

Figure 4.4: Optimal Number of Nodes for RF

4.3 Utilizing XGBoost for further Exploration

In Extreme Gradient Boosting, we have the following mathematical notation:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_{i,(t-1)} + \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

In the context of Extreme Gradient Boosting (XGBoost), which is an ensemble learning method that combines multiple weak learners (decision trees) to create a highly accurate predictive model, several key elements are defined. The predicted value for the ith instance is represented as $\hat{y}_i$. The term f_k denotes the kth weak learner within the ensemble, where k ranges from 1 to K to indicate the total number of weak learners. x_i corresponds to the ith instance in the dataset. Additionally, $\hat{y}_{i,(t-1)}$ signifies the prediction obtained from the previous iteration. The symbol $\mathcal{F}$ encompasses the space that includes all possible weak learners. It is important to note that the provided formula demonstrates the additive nature of XGBoost, where the prediction for each instance is obtained by summing the predictions from the previous iteration ($\hat{y}_{i,(t-1)}$) and the predictions made by the weak learners in the current iteration (f_k). Through this iterative process, the model progressively refines its predictions, thereby enhancing its predictive capabilities.

In Table 4.4, XGBoost incorporates multiple metrics to assess its accuracy beyond calculating the simple accuracy score (which will be discussed later). Both the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) have the same value of 0.5192. MAE is the sum of non-negative errors between the predictions and actual values, while MSE can amplify larger residuals due to the squaring operation, potentially giving more weight to larger errors. In this case, since both errors have the same value, we cannot determine which one is a better predictor. After an initial 80 rounds of boosting or iterations, the optimal Root Mean Squared Error (RMSE) was achieved at 62 rounds.

In Figure 4.5, the XGBoost plot illustrates the decision splits for each node. This supervised learning method, which involves using a training and test set, presents the scores

Model	Mean Sq. Error	Mean Abs. Error	RMSE	Number of Rounds
XGBoost	0.5192	0.5192	0.7206	62

Table 4.4: Table of Values for XGBoost

and associated conditional statements. When a value satisfies a given condition, it proceeds to the corresponding node and eventually reaches the leaf section of the graph. This plot takes into account the complexity of the different trees, and higher values indicate greater importance and better performance. In this particular case, the combination of FG% - TOV - STL has the highest positive value, indicating its strong predictive power, while the combination of FG% - STL - 3PT% has the lowest value, which is negative and suggests weaker predictive capability.

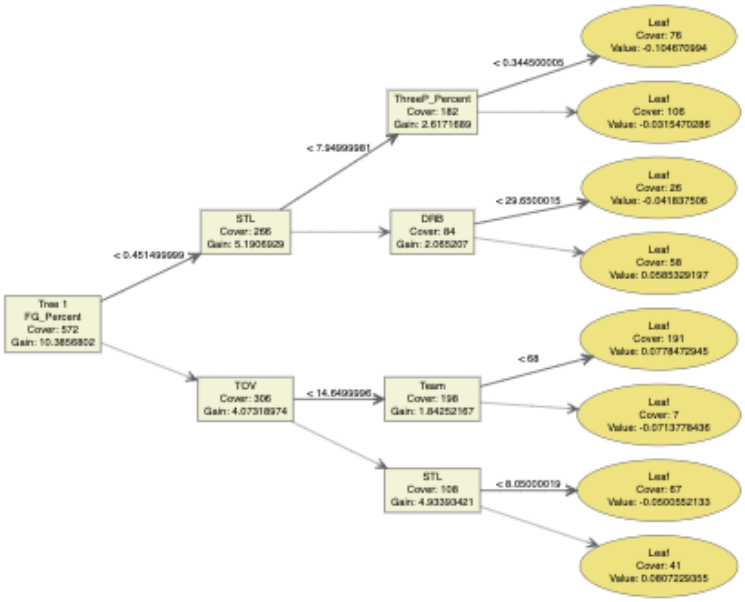

Figure 4.5: XGBoost Tree Plot

4.4 Principal Component Analysis on NBA Data

In the Principal Component Analysis Model, the mathematical derivation and notation is as follows:

$$\text{PCA}(X) = XW$$

The given equation defines the components of Principal Component Analysis (PCA) as follows: - X represents the input data matrix with dimensions $m \times n$, where m signifies the number of samples and n represents the number of features. - W denotes the transformation matrix, also known as the eigenvectors or loadings, with dimensions $n \times k$, where k indicates the desired number of principal components. - PCA(X) represents the transformed data matrix with dimensions $m \times k$, where each row corresponds to a sample, and each column corresponds to a principal component.

It is important to note that the provided PCA formula assumes a preprocessing step in which the mean of each feature has been subtracted from the data. This processed data matrix is denoted as $\tilde{X}$. Therefore, the complete formula for PCA, including mean centering, can be expressed as:

$$\text{PCA}(X) = \tilde{X}W$$

In Table 4.5, regarding PCA, after running the descriptive statistics in R, we observe a very small standard deviation, indicating a low variance in the data. The component number associated with these optimal and low values is 18, accounting for 100% of the cumulative proportion of the data.

Model	Std. Dev	Prop of Variance	Component Number	Cumulative Prop
PCA	4.655766e-04	6.537468e-08	18	1

Table 4.5: Table of Values for Principal Component Analysis

In Figure 4.6, we can observe the correlation matrix for the PCA. This matrix plays a crucial role in identifying the relationships between components and variables, aside from their direct 1:1 correlations. The diagonal of the matrix represents the correlation of each variable with itself, which is always equal to 1. The color scheme used in the matrix highlights the strength of the relationships: red indicates a potentially high correlation, while blue represents a weaker correlation. It is worth noting that there is a significant correlation observed in the bottom left corner of the matrix. This correlation is expected since field goals (FG) are directly influenced by attempts, as well as the number of two-point and three-point shots made and attempted.

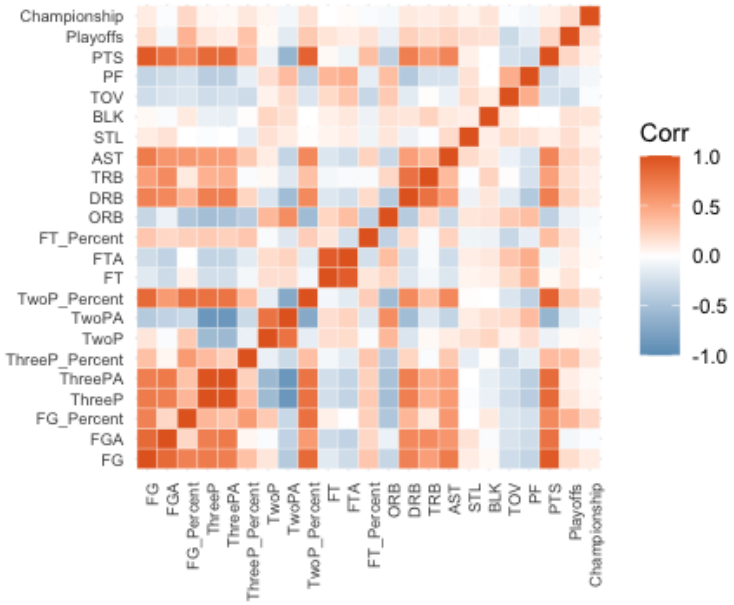

Figure 4.6: Correlation Matrix for PCA

The Scree plot, depicted in Figure 4.7, illustrates the variance explained by each of the ten principal components. The y-axis represents the percentage of variance explained. The plot reveals a distinct "elbow" point, indicating a significant drop in the variance explained

after the second component. This suggests that for further analysis and interpretation, it is sufficient to retain only the first two components.

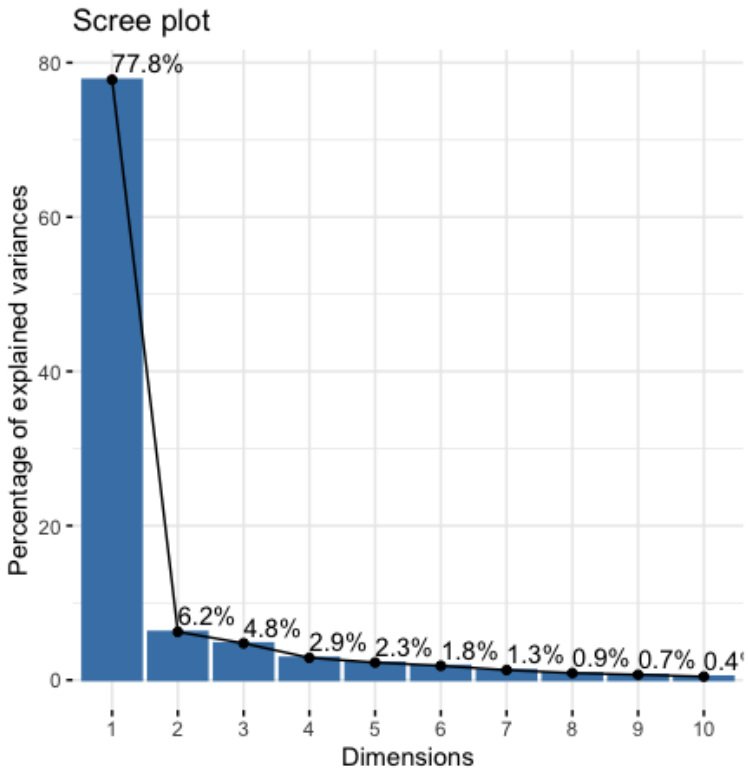

Figure 4.7: Scree Plot of PCA

In Figure 4.8, the Cosine Squared Pareto Chart for Principal Component Analysis is presented. This chart provides a visual representation of the variables ranked from highest to lowest importance for a given observation. The order of the variable vectors reveals that three-pointers made and attempted are the two most influential factors in winning a championship. This observation aligns with the dominance of the Golden State Warriors in recent years, as they have consistently excelled in three-point shooting, thanks to the exceptional performances of players like Stephen Curry and Klay Thompson.

Figure 4.9 showcases the components from the perspective of championship-winning teams. This graph employs a color gradient, with black representing higher importance and a stronger relationship.

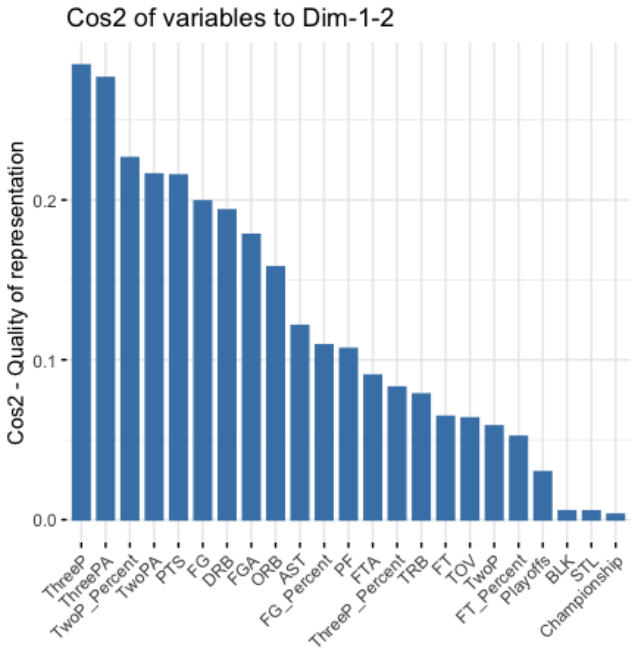

Figure 4.8: Cosine Sq. Pareto Chart

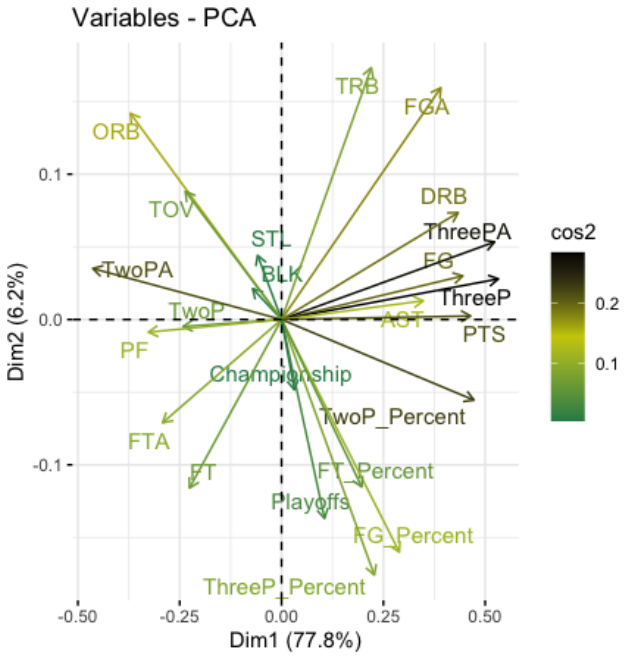

Figure 4.9: Quality of Representation for PCA

4.5 Support Vector Machines for Classification

The Support Vector Machine (SVM) classifier is designed to discover a hyperplane that can effectively divide the training data into two distinct classes. The decision function of SVM can be expressed as follows:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right)$$

In the context of binary classification, several components are defined as follows:

- $f(x)$ represents the predicted class label for a new input sample x. - α_i denotes the Lagrange multiplier associated with the i-th training sample. - y_i corresponds to the class label of the i-th training sample. - $K(x_i, x)$ is the kernel function used to calculate the similarity between the training sample x_i and the new sample x. - b represents the bias term.

The choice of the kernel function $K(x_i, x)$ depends on the specific problem being addressed. Common options include the linear kernel, polynomial kernel, Gaussian (RBF) kernel, and sigmoid kernel. It is important to note that the presented formulation assumes a binary classification problem. For multi-class classification, techniques like one-vs-one or one-vs-rest can be employed to effectively handle the task.

In Table 4.6, after performing the support vector machine analysis, we obtained a remarkably high accuracy of 96.64%. As mentioned in the report, each of the different models presented in the analysis has an accuracy value, except for PCA, as it is not inherently a classification method. The high confidence interval, with the lower bound at 0.9474 and the upper bound at 0.98, indicates a significantly low p-value. Since SVM is a supervised learning technique, cross validation was used to determine the test set for the recording of the accuracy.

The following plots demonstrate different relationships between championship character-

Model	Accuracy	Confidence Interval	P-Value
SVM	0.9664	(0.9474, 0.98)	6.151e-05

Table 4.6: Table of Values for Support Vector Machines

istics and various x and y variables. The x variables remain consistent across all four plots, representing the years.

In Figure 4.10, we observe a positive linear relationship between field goals and championship outcomes. The red 'X' marks represent teams that won the championship, and they tend to fall in the middle or upper range of the field goals distribution. Similarly, Figure 4.11 shows a positive linear relationship between championship success and three-pointers made. Teams that won the championship tend to have higher values in this category as well.

Figure 4.12 focuses on the distribution of minutes played. Although it was a major component in previous machine learning models, it is not a strong predictor of championship success. The distribution of minutes for teams predicted to win the championship appears to be highly random.

Moving on to Figure 4.13, we shift our attention to a defensive category: Blocks. Blocks prove to be a significant predictor, as teams aiming for championship success need to be in at least the upper quartile of the blocks distribution. This defensive statistic plays a crucial role in achieving the highest level of success in basketball: winning the championship.

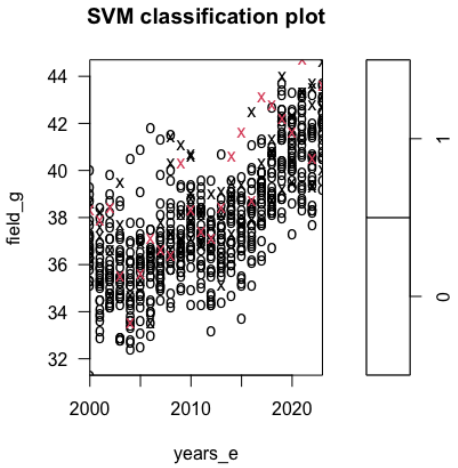

Figure 4.10: SVM plot of FG + Yr

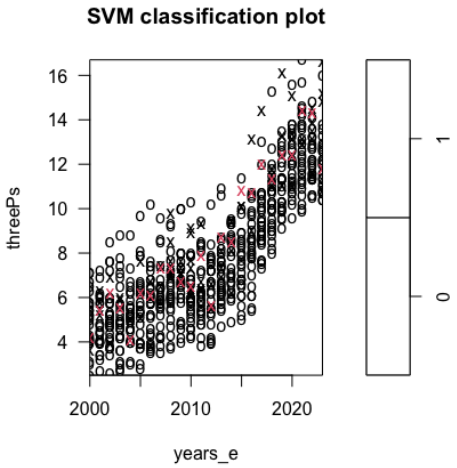

Figure 4.11: SVM plot of 3s + Yr

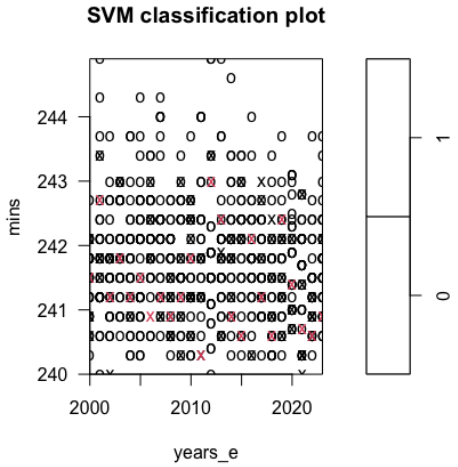

Figure 4.12: SVM plot of Mins + Yr

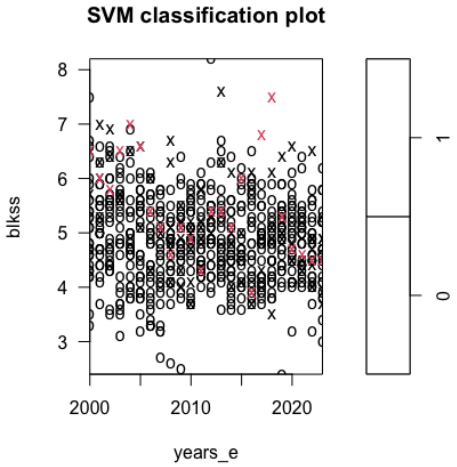

Figure 4.13: SVM plot of BLKs + Yr

4.6 Deep Learning: Neural Networks

In a neural network, we can identify several layers: the input layer, one or more hidden layers, and the output layer. Each of these layers consists of multiple neurons, also known as nodes. The calculation involved in determining the output of a neuron within the neural network can be described using the following formula:

$$a_j^l = \sigma(\sum_{k=1}^{n} w_{jk}^l a_k^{l-1} + b_j^l)$$

In the context of a neural network, we define several components as follows:

- a_j^l represents the activation or output of neuron j within layer l. - w_{jk}^l denotes the weight associated with the connection between neuron k in the previous layer $(l-1)$ and neuron j in the current layer l. - a_k^{l-1} signifies the activation of neuron k in the preceding layer $(l-1)$. - b_j^l indicates the bias term specific to neuron j in layer l. - σ denotes the activation function applied to the weighted sum of inputs and biases.

Neural networks can employ various activation functions, such as the sigmoid function ($\sigma(x) = \frac{1}{1+e^{-x}}$), the hyperbolic tangent function ($\sigma(x) = \tanh(x)$), or the rectified linear unit (ReLU) function ($\sigma(x) = \max(0, x)$).

This formula allows us to describe the computation that takes place within an individual neuron in a neural network. To describe the entire network, this formula is applied to each neuron in each layer, enabling the propagation of inputs forward through the network until the output layer is reached.

In Table 4.7, two separate runs with different variables are displayed. Model 1, using four variables, achieves an error rate of 15.55, takes 35,420 steps, and achieves a high accuracy of 94.41%. The accuracy comes from cross validation, in which there are partitioned data coming from the training and test sets. On the other hand, Model 2, with just two carefully selected variables (Minutes and Blocks), has a lower error rate of 14.57, a smaller number

of steps at 18, and achieves the same accuracy of 94.41%. This suggests that the simpler Model 2 is more optimal, yielding lower error and complexity. Figure 4.14 illustrates the performance of the four-variable model, while Figure 4.15 showcases the optimal model for analysis.

Model	Number of Variables	Error	Steps	Accuracy
Model 1	4	15.55	35420	94.41
Model 2	2	14.57	18	94.41

Table 4.7: Table of Values for Neural Network

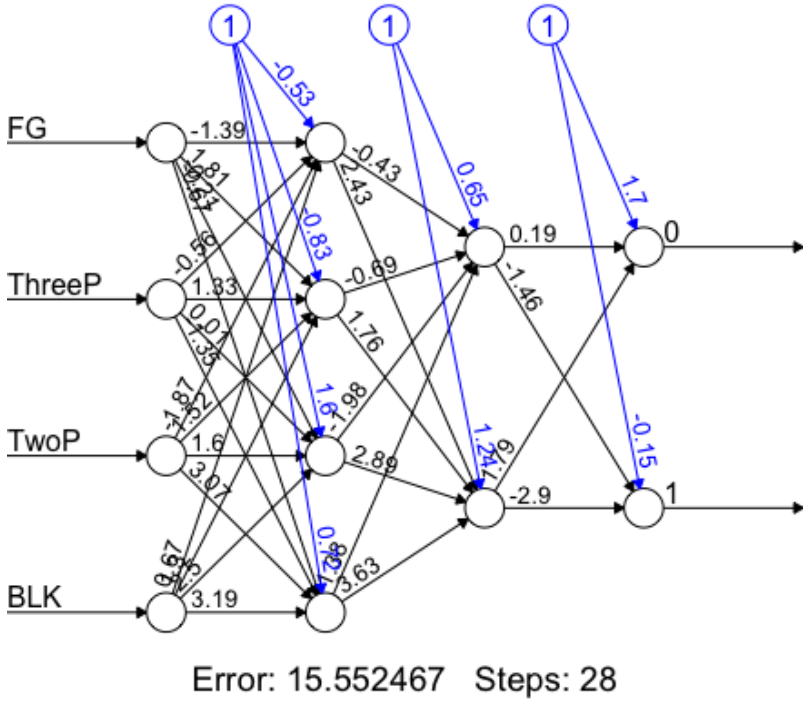

Figure 4.14: Plot of NN Model 1

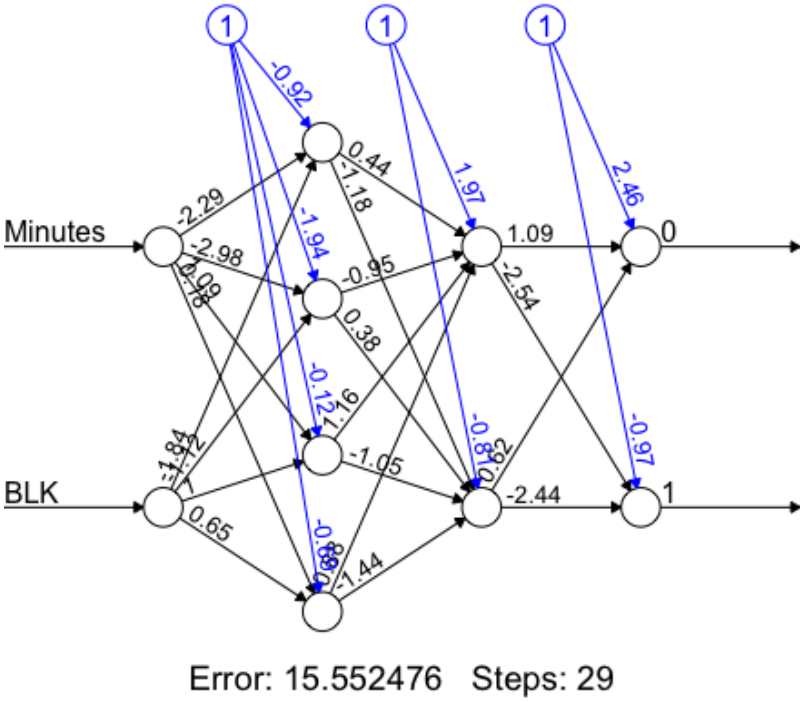

Figure 4.15: Plot of NN Model 2

CHAPTER 5

The Results and Analysis

We have conducted six different machine learning models and recorded their analyses. Table 5.1 presents the accuracies obtained for each of the models and their corresponding values. To avoid overfitting for the cases of the several of supervised learning techniques used in this case study such as: Logistic Regression, SVM, and Random Forests we have used partitioned data for both the training and testing to evaluate the accuracies. (Principal Component Analysis (PCA) is excluded from this classification analysis since it is not a classification method. PCA was used in this study to assess the significance of variables and their consistency across other models. Interestingly, the study confirmed the initial hypothesis that three-pointers made is an important variable. Additionally, through random generation and data splitting for all other machine learning models, it was found that minutes and blocks are the most influential variables. Consequently, it was crucial to maintain consistency in the number and selection of variables in order to draw robust conclusions from the data.

What was truly remarkable is that our initial model exhibited the best predictive performance. This aligns with the notion that simplicity often outperforms complexity, as evidenced by its impressive accuracy of 0.973 in this case study. On the other hand, Extreme Gradient Boosting or XGBoost had the lowest accuracy value of 0.878, although it still performed reasonably well. This discrepancy may arise from the diminishing impact of weaker learners on the model's predictive capabilities. While XGBoost is widely recognized and utilized for prediction tasks, it did not yield optimal results in this specific case study.

Logistic Regression binary classification, on the other hand, demonstrated the highest

Model	Accuracy
Logistic Regression	0.973
Random Forests	0.965
XGBoost	0.878
SVM	0.966
PCA	N/A
Neural Network	0.944

Table 5.1: Table of Accuracies

accuracy. To further enhance the model's fidelity, incorporating additional widely used variables such as attempts and other advanced statistical categories could potentially yield even better results. However, for the time being, the current model is deemed sufficient.

CHAPTER 6

Conclusion

It appears that minutes and blocks are the key variables in this study when it comes to predicting the success of a team in the National Basketball Association (NBA) and their chances of winning the championship. Whether using a simpler machine learning model like Logistic Regression or a more complex one like Neural Network, these variables play a significant role in shaping the predictions and telling a comprehensive story. It is a fallacy to assume that the rise of prominent three-point shooters alone can guarantee a championship-caliber team, as this study reveals the importance of other factors.

By employing theoretical frameworks and mathematical notations, the iterative nature of all the models becomes apparent. Utilizing the R programming language for both Exploratory Data Analysis (EDA) and Machine Learning allowed for real-time fine-tuning of functions to achieve optimal results. To further enrich the study, incorporating more advanced statistics and applying these models to other sports would be valuable. The National Hockey League (NHL) could serve as an interesting comparison due to similarities in schedules, offensive/defensive statistics, and pace of play.

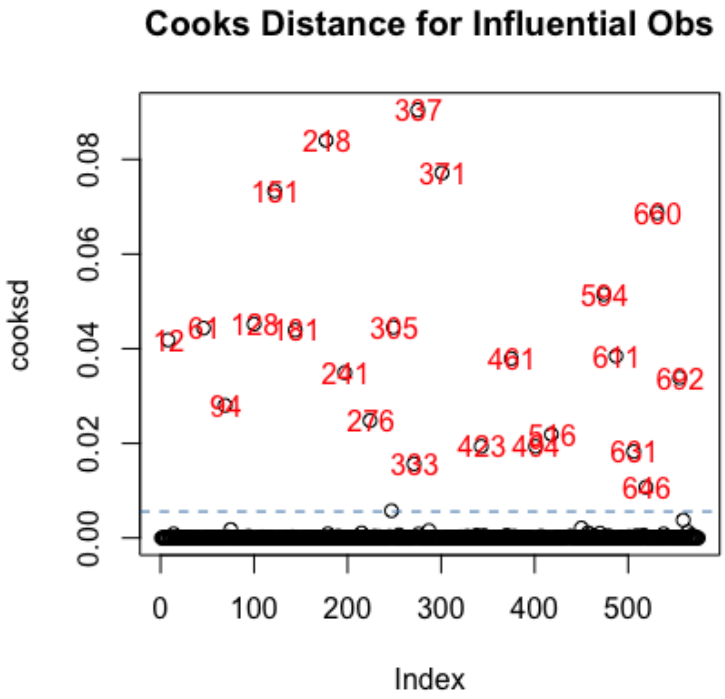

Figure 6.1: Plot of NN Model 2

To enhance the accuracy of the data and develop more robust machine learning models, it would have been beneficial to explore the concept of measuring the "Cook's Distance" to identify influential points. Figure 6.1 displays the outliers or influential points, and their presence may have biased the analysis in various ways. A clear limitation would be to eliminate these outliers and re-run the analysis. It is intriguing to observe the abundance of outliers; however, it is important to note that this model relies solely on the logistic regression model, and outliers can vary between different models. Alternatively, these values could be substituted with either the mean or median values if they are not chosen to be removed altogether.

Number	Team Name + Year
12	Denver Nuggets* 2023
61	Milwaukee Bucks* 2021
94	Los Angeles Clippers* 2020
128	Toronto Raptors* 2019
151	Golden State Warriors* 2018
181	Golden State Warriors* 2017
218	Cleveland Cavaliers* 2016
241	Golden State Warriors* 2015
276	San Antonio Spurs* 2014
303	Oklahoma City Thunder* 2013
305	Miami Heat* 2013
333	Oklahoma City Thunder* 2012
337	Miami Heat* 2012
371	Dallas Mavericks* 2011
423	Los Angeles Lakers* 2009
461	Boston Celtics* 2008
494	San Antonio Spurs* 2007
516	Miami Heat* 2006
594	Detroit Pistons* 2004
611	San Antonio Spurs* 2003
631	Los Angeles Lakers* 2002
646	Detroit Pistons* 2002
660	Los Angeles Lakers* 2001
692	Los Angeles Lakers* 2000

Table 6.1: Table of Outliers for Analysis

The outliers in this analysis comprised exceptional teams, with approximately 19 out of the 24 outliers being NBA Champions, as indicated in Table 6.1. This suggests that, in terms of statistics, it is crucial to excel in one or more categories in order to have a high chance of becoming a champion. The study revealed that teams that played more games in the regular season and had defensive leaders in blocks had a greater likelihood of emerging as victors compared to other teams in the league.

It is important to acknowledge that relying solely on raw statistics and subjective perceptions can often lead to divergent conclusions. Factors such as playing more overtime games or having defensive leaders in blocks can significantly influence the probability of winning an NBA championship. Therefore, it is essential to rely on objective statistical analysis for accurate predictions and valuable insights. Through exploratory data analysis, predictive modeling, and concluding with an outlier analysis, it is truly remarkable to uncover the factors that a Data Science team for an NBA organization should consider in their decision-making process.

REFERENCES

[1] L. Freitas, "Shot distribution in the nba: did we see when 3-point shots became popular?" *German Journal of Exercise and Sport Research*, vol. 51, no. 2, pp. 237–240, 2021.

[2] M. Nourayi *et al.*, "Strategically driven rule changes in nba: Causes and consequences," *The Sport Journal*, vol. 22, pp. 1–11, 2019.

[3] E. S. Jones, "Predicting outcomes of nba basketball games," Ph.D. dissertation, North Dakota State University, 2016.

[4] J. B. Yang and C.-H. Lu, "Predicting nba championship by learning from history data," *Proceedings of artificial intelligence and machine learning for engineering design*, 2012.

[5] D. Wallach and B. Goffinet, "Mean squared error of prediction as a criterion for evaluating and comparing system models," *Ecological modelling*, vol. 44, no. 3-4, pp. 299–306, 1989.

[6] A. Hapfelmeier, T. Hothorn, K. Ulm, and C. Strobl, "A new variable importance measure for random forests with missing data," *Statistics and Computing*, vol. 24, pp. 21–34, 2014.

[7] G. Y. Kanyongo, "Determining the correct number of components to extract from a principal components analysis: a monte carlo study of the accuracy of the scree plot," *Journal of modern applied statistical methods*, vol. 4, no. 1, p. 13, 2005.

[8] J. B. Miller and A. Sanjurjo, "Is it a fallacy to believe in the hot hand in the nba three-point contest?" *European Economic Review*, vol. 138, p. 103771, 2021.