Using Machine Learning Techniques to Predict Factors of Winning in the NBA

Author: Anthony Chan (500766830)

CIND820 DJ0- Big Data Analytics Project

Supervisor’s Name: Sedef Akinli Kocak

March 28, 2022

Table of Contents

[Abstract 3](#_Toc452457337)

[Literature Review 4](#_Toc452457338)

[Data Description 7](#_Toc452457340)

[Methodology 13](#_Toc452457340)

[Data Preparation 14](#_Toc452457340)

[Initial Results 15](#_Toc452457340)

[Models 16](#_Toc452457340)

[Are these results acceptable? 19](#_Toc452457340)

[Should we try a different approach? 19](#_Toc452457340)

[Final Results 21](#_Toc452457340)

[References 23](#_Toc452457340)

# Abstract

Data analytics have impacted the sporting industry greatly since we can now analyze and predict what a player’s potential can be. Many of these statistics and factors can be translated to what we can see developing in the game. A game where each team can optimize their strategies to win. In this capstone project, the theme will be revolting around predictive analysis and knowledge discovery about the evolution of shooting the basketball within the NBA (National Basketball Association). The dataset used for this project is composed of game statistics of NBA teams from the years, 2014 - 2018. Many factors can go into winning a championship in the NBA, for example the 3-Point Field Goal Percentage, Field Goal Percentage, Free Throw Percentage, Turnovers, Offensive and Defensive Rebounds, and many more statistics.

The model that I will be creating will use the data provided to simulate the how the players in the NBA are shooting in the modern era and compare them to an older generation of basketball. This will allow us to find out the best performing model and use it as a basis for our study. Next, the model will attempt to predict how players will adapt to this new playstyle based on trends found which will help create insight to what teams and coaches should prioritize. This project would be able to investigate the problems: How important are each variable within the sport? Can we determine which factors affect winning chances of the team? Can we accurately predict a team’s chances of winning in the NBA using these variables? For this project, I will be using classification techniques and predictive modelling methods such as Logistic Regression, Naïve Bayes, and XGBoost. I will be implementing my analysis with software in Python where I will use packages such as Pandas, Numpy, SKlearn, Matplotlib, and others to compute predictive modelling techniques and visualize our data.

In conclusion, this capstone project will dive into the variables that take place in every NBA basketball game and learn about how they impact the players around the organization.

## Literature Review

Teams that use data analytics to track and predict data gives them a huge advantage to their competitors. This literature review will discuss what data analyst have studied within the NBA and how teams have taken advantage of these circumstances to determine a winning strategy.

The NBA has been an organization since 1946 and has been growing and developing to be one of the biggest sports in the world. There have been many changes within the rules and how the game has been played. For example, the league did not introduce the three-point shot until 1979 and how the league has changed from Centers playing near the basket to Guards shooting far from the basket. The game of basketball will forever be evolving as players grow up watching their favourite basketball player and they develop their own skills.

Many teams have started using data analytics to determine the outcome of their players. An example for this would be in an article by Wharton (2017) regarding the commissioner of the league. As mentioned in Wharton (2017), teams have started to monitor their players during practice to measure fatigue and prepare their players for any resting time. This will allow players to continue to stay healthy within the season and avoid injuries which should result teams to have a better win percentage. Another interesting fact from this article is that teams are using analytics to plan strategies. An exampled explained in Wharton (2017), 42% of the time a player goes forward with his left foot. This can lead to coaches developing strategies to counter this movement from players. Teams have been using data acquired from prospects in high school and college to determine how the future of the franchise by drafting players with certain aspects and skillsets. These are some of the external factors that can determine how teams perform during the season.  
 This capstone project will use statistics from players currently in the league. In the study shown from Lieder (2018), they investigated certain factors related to matches to create a model for predicting the results of NBA games. The authors used machine learning techniques such as logistics regression and linear regression to build predictive models to evaluate the outcome of these games. The model resulted in a 70% accuracy rating which the author has noted that this is extremely high as there can be multiple variables and aspects to the game that cannot be predicted.

Miljkovic (2010) studied machine learning methodologies such as k-nearest neighbor, decision trees, linear regression, and Naive Bayes where they would attempt to predict the outcome of NBA games. The dataset that they used were games from the years 2009 – 2010 and some examples of the variables used were free throws, three points, field goals, blocks, home games and away games, fouls, win percentage, and loss percentage. Their results using these machine learning models had an impressive 67% predictive accuracy. Like the study done in Lieder (2018), many variables that cannot be quantified affected the accuracy of the models.

Mikołajec (2013) conducted a study where they wanted to determine what teams in the NBA were doing that lead them to succeed. In the study, they worked with 52 variables such as Win percentage, Offensive efficiency, 3rd Quarter Points Per Game, Average Steals and Fouls. They determined that they would use the winner of the current season as their dependent variable and used the methods Pearson’s coefficient and regression analysis with the correlation matrix to identify optimal combinations of variables. Mikołajec (2013) noted that studies done in other basketball leagues such as the European Basketball League showed that the factors 3 points attempts, field goals percentage, free throws made, free throw percentage, defensive rebounds, and turnovers impacted a team’s chances of winning the most.

Another study done, Oliver (2005) had the factors field goal percentage, turnovers, offensive rebounds, and free throws made affected the results of the game. The reason for these 4 statistics to be so impactful is how they lead to advantages to the opposing team. For example, turnovers gave the opposing team more possessions and chances to score a basket. Rebounds allow the offensive team to have a second chance at scoring. Additionally, causing a significant number of free throws to a team can affect how the coach makes substitutions and plays.

As shown from multiple articles, this type of study has been done before as analytics in the sporting industry is growing exponentially. However, this literature review will also highlight similarities and differences between this project and others. I will be focusing on different research questions as I would like to take on the importance of certain variables and I will also be using different seasons of the NBA to differentiate my study against others. Based on the articles I have reviewed; I will be using XGBoost to assist with classification methods as it is a decision tree-based machine learning algorithm. Logistic regression and Naïve Bayes have been a standard model used in majority of the studies shown from the articles. For that reason, I will be using those predictive modelling methods to understand the relationship between variables and be able to compute predictions. Since this study will be based on the same material, I will be able to this capstone project with studies that have been done in the past to further prove or disprove my research questions.

#### Data Description

The dataset used consists of 9840 rows of games played in the NBA which consists of 40 variables. However, since many rules were not implemented in the early years of the NBA.

In Table 1, a brief description of the variables used is shown.

Table 1 – Variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Description | Datatype | Type of Variable | Mean |
| Team | NBA Team | Object | Categorical | N/A |
| Game | Which game of the season | Int64 | Numerical | N/A |
| Date | Date of Game | Datetime64 | Categorical | N/A |
| Home | Game played at Home or Away | Object | Categorical | N/A |
| Opponent | Who they played against | Object | Categorical | N/A |
| WINorLOSS | Was the game a win or a loss | Object | Categorical | N/A |
| TeamPoints | How many points were scored | Int64 | Numerical | 103.652337 |
| OpponentPoints | How many points the opponent scored | Int64 | Numerical | 103.652337 |
| FieldGoals | Number of shots made | Int64 | Numerical | 38.602439 |
| FieldGoalsAttempted | Number of shots taken | Int64 | Numerical | 84.902439 |
| FieldGoals% | Percentage of Field Goals | Float64 | Numerical | 0.455689 |
| 3PointShots | Number of 3 point shots made | Int64 | Numerical | 9.126829 |
| 3PointShotsAttempted | Number of 3 point shots taken | Int64 | Numerical | 25.623679 |
| 3PointShots% | Percentage of 3 Point Shots | Float64 | Numerical | 0.354321 |
| FreeThrows | Number of Free Throws made | Int64 | Numerical | 17.320630 |
| FreeThrowsAttempted | Number of Free Throws taken | Int64 | Numerical | 22.749390 |
| FreeThrows% | Percentage of Field Goals | Float64 | Numerical | 0.762395 |
| OffRebounds | Offensive Rebounds | Int64 | Numerical | 10.287602 |
| TotalRebounds | Total Rebounds | Int64 | Numerical | 43.520630 |
| Assists | Number of Assists | Int64 | Numerical | 22.546545 |
| Steals | Number of Steals | Int64 | Numerical | 7.750508 |
| Blocks | Number of Blocks | Int64 | Numerical | 4.827642 |
| Turnovers | Number of Turnovers | Int64 | Numerical | 13.638618 |
| TotalFouls | Number of Fouls | Int64 | Numerical | 20.058537 |
| Opp.FieldGoals | Number of shots made by the opponent | Int64 | Numerical | 38.602439 |
| Opp.FieldGoalsAttempted | Number of shots taken by the opponent | Int64 | Numerical | 84.902439 |
| Opp.FieldGoals% | Percentage of Field Goals by the opponent | Float64 | Numerical | 0.455689 |
| Opp.3PointShots | Number of 3 point shots made by the opponent | Int64 | Numerical | 9.126829 |
| Opp.3PointShotsAttempted | Number of 3 point shots taken by the opponent | Int64 | Numerical | 25.623679 |
| Opp.3PointShots% | Percentage of 3 Point Shots by the opponent | Float64 | Numerical | 0.354321 |
| Opp.FreeThrows | Number of Free Throws made by the opponent | Int64 | Numerical | 17.320630 |
| Opp.FreeThrowsAttempted | Number of Free Throws taken by the opponent | Int64 | Numerical | 22.749390 |
| Opp.FreeThrows% | Percentage of Field Goals by the opponent | Float64 | Numerical | 0.762395 |
| Opp.OffRebounds | Offensive Rebounds by the opponent | Int64 | Numerical | 10.287602 |
| Opp.TotalRebounds | Total Rebounds by the opponent | Int64 | Numerical | 43.520630 |
| Opp.Assists | Number of Assists by the opponent | Int64 | Numerical | 22.546545 |
| Opp.Steals | Number of Steals by the opponent | Int64 | Numerical | 7.750508 |
| Opp.Blocks | Number of Blocks by the opponent | Int64 | Numerical | 4.827642 |
| Opp.Turnovers | Number of Turnovers by the opponent | Int64 | Numerical | 13.638618 |
| Opp.TotalFouls | Number of Fouls by the opponent | Int64 | Numerical | 20.058537 |

Table

Description automatically generated Table

Description automatically generated Table

Description automatically generated with medium confidence

Figure 1 – Correlation between variables

In Figure 1 is a quick correlation between each the variables.

#### Methodology

Figure 2 shows a flow chart of the methodology that will be done throughout this capstone project.

A picture containing text, screenshot, businesscard

Description automatically generatedFigure 2

#### Data Preperatation

Graphical user interface, text, application, email

Description automatically generated

A picture containing text

Description automatically generated

#### Initial Results

A picture containing scatter chart

Description automatically generated

Figure 3. Correlation Matrix

Some observations can be seen from the correlation matrix in Figure 3. A problem that we might see in the figure is that correlation does not imply causation. However, this matrix will give us a general idea of the results we might see from the regression models. We can see that the variables TeamPoints, FieldGoals, FieldGoal %, 3PointShots, 3PointShot%, FreeThrows, Assists, TotalRebounds, Blocks correlate with the variable WINorLoss the most.

|  |  |
| --- | --- |
| Variable | Correlation to Wins |
| TeamPoints | 0.46 |
| FieldGoals | 0.38 |
| FieldGoals% | 0.45 |
| 3PointShots | 0.24 |
| 3PointShots% | 0.32 |
| FreeThrows | 0.14 |
| Assists | 0.31 |
| TotalRebounds | 0.26 |
| Blocks | 0.17 |

Figure 4. Correlation description

For team points, it is obvious as the more points you score, the higher chance of winning the game is. Same be said about field goal and free throw statistics, the more field goals or free throws made, the more points scored which results in a higher winning percentage. 3 pointers are worth the most in terms of scoring in the NBA, if teams can shoot at a higher rate, this should result in more wins. The defense variables such as rebounding and blocking have a high correlation with the WINorLOSS variable because these are actions that teams do to stop the opponent from scoring, thus leading to less points scored by the opponent.

#### Models

Logistic Regression

Logistic Regression is our first classification model that we will be using, and it is one of the simplest and commonly used machine learning algorithm. It describes and estimates the relationship between one dependent binary variable and independent variables.

First, we will determine our dependent and independent variables. We will use the variables that correlate the most with the WINorLOSS variable as our independent variable and the WINorLOSS variable as our dependent variable. We will split the data into train and test datasets: X\_train, X\_test, y\_train, and y\_test and we will run the train and test datasets through the logistic regression model and plot a confusion matrix to determine Accuracy, Precision, and Recall.

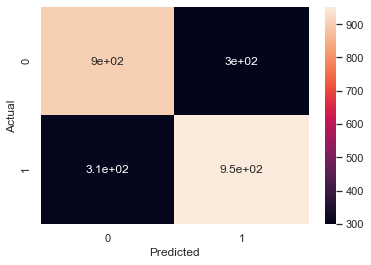


Figure 5. Confusion Matrix

XGBoost

XGBoost will be our second classification model and it is an implementation of gradient boosted decision trees. This classification model is designed to be very efficient and flexible. This is because they use decision trees to create a model that predicts the label by evaluating a tree and estimating the minimum number of questions needed. We use decision trees for classification to predict a category, or regression so that we can predict a continuous value. In our model, we ran with the default parameters to test the accuracy of the model.

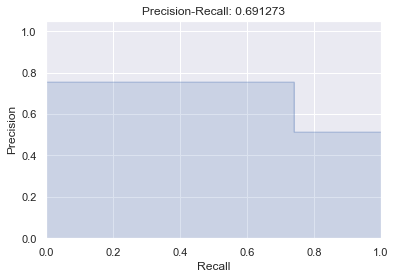


Figure 6. Precision-Recall graph

Naïve Bayes

Our third classification model, Naive Bayes, is built on Bayesian classification methods. These rely on Bayes's theorem, which is an equation describing the relationship of conditional probabilities of statistical quantities. Naïve bayes is a simple and easy classification model to build and is a great model to run for large datasets. For this capstone project, we will be using Gaussian Naïve Bayes to classify our dataset.

Accuracy, Precision, Recall, F1

Accuracy, Precision, Recall, F1 are very important in determining whether our model can be trusted. Accuracy is a measure where it is a ratio of correctly predicted observation to the total observations. However, high accuracy does not mean that our model is best, and we would have to look at other parameters to evaluate the true performance of your model. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations and recall is the ratio of correctly predicted positive observations to the all observations in actual class. F1 Score is the weighted average of Precision and Recall which means that it also takes both false positives and false negatives into account. These values coincide with each other to help us give an understanding of our models and if we would need to make any adjustments to them. From the graph below, each of these values for each model is shown and our Logistic Regression model seems to be the most accurate in determining a win.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 |
| Logistic Regression | 75.4% | 76% | 75.7% | 75.88% |
| XGBoost | 74.39% | 75% | 74% | 75% |
| Naïve Bayes | 71.7% | 73.96% | 69.88% | 71.87% |

#### Are these results acceptable?

In our literature review, an accuracy rating of 70% and above is an acceptable result. Not every game is played the same and the players are not constant in each game and there will always be factors that cannot be quantified and unpredictable due to the nature of this sport. For example, if the team's best player has been injured for the season, their results for that season will not only impact the chances of winning but also the variables for the entire season.

#### Should we try a different approach?

Before concluding any models, check if our WINorLOSS variable is balanced or not. The reason we check for balance is because having a balanced dataset for a model would generate higher accuracy models. Therefore, it is important to have a balanced data set for a classification model. From our code, we have determined that the variable WINorLOSS is balanced. If it was imbalanced, we would have to balance WINorLOSS and run the models again when it is balanced.

To try another approach, we can change how we split our data. We can run the same models with a different train-test split and calculate our findings again. The most common split percentages include:

With Train: 80%, Test: 20%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 |
| Logistic Regression | 75.05% | 75.32% | 75.7% | 75.51% |
| XGBoost | 73.98% | 73% | 74% | 74% |
| Naïve Bayes | 71.39% | 73.24% | 69.99% | 71.58% |

With Train: 67%, Test: 33%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 |
| Logistic Regression | 74.66% | 76.03% | 73.8% | 74.9% |
| XGBoost | 74.66% | 76% | 74% | 75% |
| Naïve Bayes | 71.7% | 74.24% | 69.71% | 71.9% |

After changing how we train our dataset, we can see that the XGBoost model is giving similar results to our Logistic Regression model.

Another approach I tried was changing the XGBoost parameters. I lowered the learning rate, max depth, and colsample\_bytree of the model to make the boosting process more conservative and attempt to not overfit our dataset. Our subsample was set to 0.8 instead of uniform to look for better results. With these parameter changes, the model outputted and accuracy of 73.94% and a mean absolute error of 0.26.

Cross Validation

Cross-validation is a statistical method used to evaluate machine learning models and how the model is expected to perform when used to make predictions on data not used during the training of the model. I decided to run a ten-fold cross validation on each of our models using our original training.

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Standard Deviation |
| Logistic Regression | 0.769 | 0.015 |
| XGBoost | 0.758 | 0.011 |
| Naïve Bayes | 0.725 | 0.015 |

#### Final Results

This project has explored the topic of machine learning in Python programming environment using various libraries such as pandas, numpy, sklearn, seaborn and classifiers such as Logistic regression, XGBoost and Naïve Bayes. The objective was to see if we can use machine learning techniques to determine if a team would win or lose based off the variables of the game. This was accomplished through exploratory data analysis, data cleaning, and multiple classification models.

From our correlation matrix, the main variables to focus on were: TeamPoints, FieldGoals, FieldGoal %, 3PointShots, 3PointShot%, FreeThrows, Assists, TotalRebounds, and Blocks. These were the variables that were found to have the most correlation to winning in the NBA. As shown from our literature reviews, most of these variables were important as it was in other NBA datasets that focus on different years. From the models that were created in this project, Logistic Regression has shown to be the most accurate in predicting a win with an accuracy of 75.4%. Normally, we would see a more accurate result from XGBoost due to its decision tree-based classification. However, this was not the case in this study and a reason could be because of overfitting the dataset.

From this study, we have answered our abstract questions and hypotheses where we have determined which factors lead to winning in the NBA and how impactful they are to the game. This project also shows that we can accurately predict the outcome of the game with an accuracy of at least 70%.

However, based on what we have completed in this research study, our results can still be improved. In the game of basketball, there are many external factors that go into the game that cannot be quantified with numbers. This can lead to weaknesses and limitations of our model due to the unpredictable nature of basketball. An example of this would be the players and coach of the team. Furthermore, the data that we have gathered covers on 4 years of the NBA and with the game changing every year, we cannot always accurately predict who will win. Hence, in near future if I were to do this study again, I would attempt to do more studies on each variable and factor of the NBA and run different types of classification models.

#### References

Khan, E. (2017, October 3). *Advanced NBA stats for dummies: How to understand the new hoops math*. Bleacher Report. Retrieved February 2, 2022, from https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. Annals of Data Science, 6(1), 103-116.

Wharton (2017, June 1). *The NBA's Adam Silver: How Analytics is transforming basketball*. Knowledge@Wharton. Retrieved February 8, 2022, from https://knowledge.wharton.upenn.edu/article/nbas-adam-silver-analytics-transforming-basketball/

Grant , H. (2015, July 23). How the game has changed. Retrieved February 8, 2022, from https://www.horacegrant.com/blog/how-the-game-has-changed/

Lieder, N. (2018, July 25). *Can machine-learning methods predict the outcome of an NBA game?* SSRN. Retrieved February 9, 2022, from https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3208101

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010, September). The use of data mining for basketball matches outcomes prediction. In IEEE 8th international symposium on intelligent systems and informatics (pp. 309-312). IEEE.

Mikołajec, K., Maszczyk, A., & Zając, T. (2013). Game indicators determining sports performance in the NBA. Journal of human kinetics, 37, 145.

Csátaljay, Gábor & James, Nic & Hughes, Mike & Dancs, Henriette. (2013). Effects of defensive pressure on basketball shooting performance. International Journal of Performance Analysis in Sport. 13. 10.1080/24748668.2013.11868673.

Oliver, D. (2005). What wins basketball games, a review of „Basketball on paper: Rules and tools for performance analysis”. Polomac Books, 26-85.

Wen, R. (2017, September 17). NBA Data. figshare. Retrieved January 23, 2022, from

<https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018>

Github - <https://github.com/anthonyychan/cind820>