# **5 Year Survival of NBA Rookies**

## **Muthukumar Kadhirvel**

## Topic

## 5 Year Survival of NBA Rookies - Predict whether a rookie NBA player will last 5 or more years in the league based on their rookie year statistics.

## Background/History

Rookies are professional athletes in their first year and I have been fascinated by how these talents are identified and become household names in a few years’ time. NBA draft is used to add rookies to their teams and sometimes become championship contenders. For any rookie player who are part of the first round of the NBA draft, contract is of 2 years with full guarantee and the 3rd and 4th years being options. Other players as part of their draft have their contract from 1 to 4 years with no guarantees. (Wright 2022).

## Business Problem

For a team to be successful they need to identify rookie players who will benefit them in the long term. This is one of the most important questions to be answered before every season and I am going to utilize Data Analytics to try and answer this problem. Using the dataset available that lists the rookie year statistics I am going to build a predictive model that would carry out the analysis on whether a rookie NBA player will last 5 or more years in the league as the average career length of an NBA player is 4.5 years. (Luke 2021).

## Data Explanation

I have downloaded the rookie dataset from the below Kaggle URL that has the rookie year statistics of players from 1980 to 2015 and the final column is the target variable of 0 (NO) or 1 (YES) that they made it 5 years and above in NBA.

<https://www.kaggle.com/datasets/mamadoudiallo/5-year-survival-of-nba-rookies-from-19802015>

The dataset has 1424 rows and 22 columns out of which one variable is the target variable. I will use this data to split into training and test data and develop models to predict the target variable of whether the rookie makes it to 5 years and above in the league.

* I will drop attribute 'Unnamed: 0' that would not be needed in the model prediction.
* There are no missing or NULL values and so no need to handle them accordingly.
* Since all variables are numeric and continuous there is no need to create dummy variables.
* Perform EDA on the dataset to visualize the attributes so that we see a trend in them.
* Created a correlation matrix between the attribute variables.

Figure 1: The below pie chart shows the % of rookies who lasted 5 years in the league and not.

Chart, pie chart

Description automatically generated

We can see that we have around 60% of the surveyed rookies lasted 5 years in the league.

Figure 2: The below histogram shows the years and the number of rookies.

Chart, bar chart

Description automatically generated

As we can see that 2009 had most number of rookies who lasted 5 years in the league with 33 with 1988 and 1989 being next with 32 rookies.

Figure 3: The below scatter plot shows the correlation between Field goals made per game and Field goals attempted per game.

Chart, scatter chart

Description automatically generated

As we can see that there is a highly positive correlation between Field goals made per game and Field goals attempted per game.

Figure 4: The below correlation matrix shows the relation between feature variables and target variable.

Chart

Description automatically generated

We can see from the Correlation Matrix that there are no clear variables that have a good correlation with target variable.

## Methods

* Split the data into training dataset and test data in the ratio of 80% to 20%.
* Train and test various prediction models such as Logistic Regression/KNN – K Nearest Neighbours/Support Vector Machine Model/Random Forest Classifier etc. to find the performance using Accuracy/Confusion Matrix etc. to arrive at the best model to be used.

## Analysis

Since ours is a binary classification problem, accuracy is the best metric to be used. Along with accuracy we will consider Confusion Matrix also as a metric to arrive at the best model.

Accuracy of Various Models

Logistic Regression 72.63

KNN 67.72

Support Vector Machine 67.72

Random Forest Classifier 72.98

Random Forest Classifier 69.47 (Scaled Data)

Random Forest Classifier model has the best accuracy and the best confusion matrix as shown below and so that can be used for further usage over the other models.

Chart

Description automatically generated

Figure 5: The below bar plot shows the top 10 Feature Importance scores against Target Variable.

Chart, bar chart

Description automatically generated

We can see that Games Played has the most feature importance and then comes Minutes Played.

## Conclusion

Since Random Forest Classifier has the best accuracy and the confusion matrix, we choose that one and go ahead with implementation.

## Assumptions/Limitations

Since the data does not consider injuries etc. which are a key factor in the rookie lasting 5 years in the league, our data would not be wholesome. But for data modelling purposes we can consider injuries as a non-impact factor.

## Challenges

None.

## Future Uses/Additional Applications

We can use this modelling technique with any other sports datasets as well if the question we are answering is a binary classification problem.

## Implementation Plan/Recommendations

Using the model, we have any team can utilize existing rookie players statistics and take a call whether to continue investing in them or move on to other rookies and concentrate on them. This way team’s resources are spent wisely on a long-term bet.

## Ethical Assessment

The dataset has rows pertaining to NBA rookies but the data that is present is not PII data and it even does not identify the player by their names. Hence there are no ethical issues with this data. Typically, datasets like these would have Race/Ethnicity etc. that could cause ethical issues when analysing the data or modelling but since there are no such attributes we look good here.

## References

Wright, S. *et al.* (2022) *NBA rookie contracts explained*, *Franchise Sports*. Available at: https://franchisesports.co.uk/nba-rookie-contracts-explained (Accessed: March 19, 2023).

Luke (2021) *What is the average career length of an NBA player?*, *Dunk or Three*. Available at: https://dunkorthree.com/nba-player-career-length/ (Accessed: March 19, 2023).

Daubs, K. (2021, November 25). *Top 10 NBA players with the most PPG during a rookie season*. Retrieved April 2, 2023, from https://fadeawayworld.net/nba/top-10-nba-players-with-the-most-ppg-during-a-rookie-season-wilt-chamberlain-was-a-scoring-god-since-his-first-season

## Appendix

Top 10 NBA Rookie players (Daubs, K, 2021)

## 