# Predict the career length of a player in the NBA – Task 3

## 1. Introduction

The National Basketball Association (NBA) is of the popular league in North America. The analysis was built models and found a best model to predict if a player will last five years in the NBA. The career experience is based on different factors for any players such as 'Field Goal Made', 'Field Goal Attempt', 'Free Throw Attempt', etc. This analysis was taken the target variable as 'Target_5Yrs' and it's a binary one. (1) expresses the meaning of the target variable in the project.

## 2. Methodology

The analysis was used one of the best three classifications algorithms which are Logistic Regression (LR), Gaussian Naïve Bayes (NB) and Neural Network Classification (NN) to build the best model. For logistic regression, when we have categorical output, we fit logistic function like sigmoid to the data instead of fitting straight line. Naïve Bayes algorithm is one of the fastest algorithms and it's based on the Bayes theory. Neural Network is a different algorithm from other two that were used in this analysis and its workflow is very similar to human brain workflow.

Accuracy and F1-score were used to calculate the model performance for further clarifications.

## 3. Results and Discussion

First and foremost, the analysis was tried to capture the patterns of the dataset using exploratory data analysis and necessary data preprocessing methods. Initially, the analysis was tried to capture the missing values, data distribution using density plots and correlation between each variable using heatmap and correlation with the response variable using bar chart of correlation matrix. There are some missing values in a variable and removed them. Basically, it was selected only two independent variables based on the correlation which have the highest relationship between the variables with target variable. The initial model was built using 'Field Goals Made' and 'Games Played' variables with target variable. LR, NB and NN models were respectively given 71%, 72% and 68% as model accuracy or correct prediction rate [ Task 3.1 ].

Figure 1, Figure 2 and Figure 3 represent the initial model visualizations of LR, NB and NN respectively.

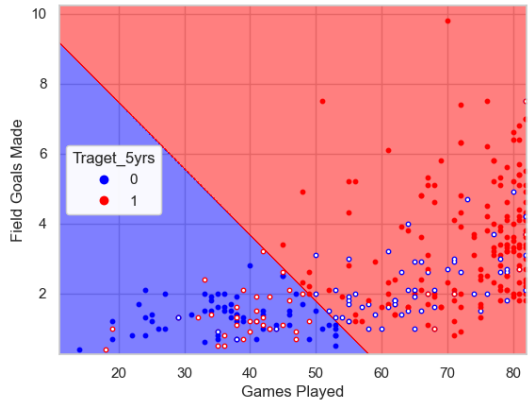


Figure 1: Initial model for Logistic Regression

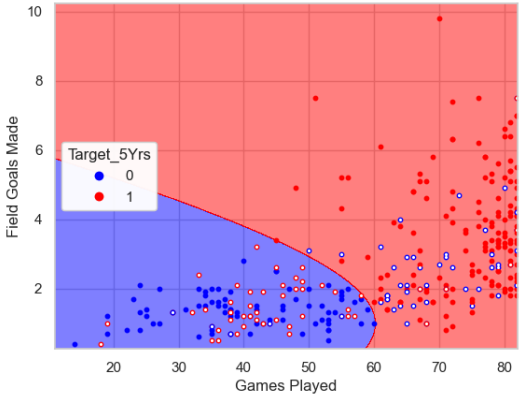


Figure 2: Initial Model for Naive Bayes

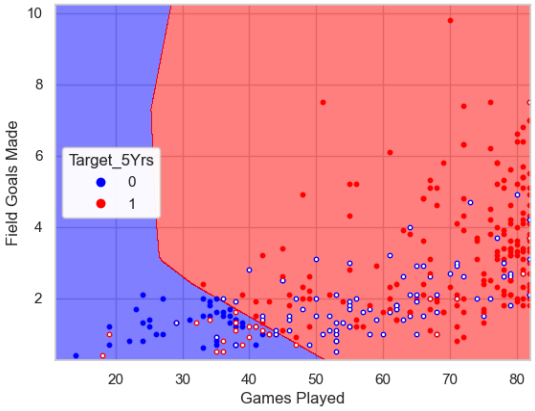

Figure 3: Initial Model for Neural Network

Secondly, the model was performed with all the independent variables under the same preprocessing techniques that were used in initial model and second initial model was given 70%, 66% and 71% accuracy respectively for LR, NB and NN algorithms [ Task 3.1 ]. Next, more data preprocessing techniques which are duplicate value removing, checking the data distribution, balancing the classes of target variable using Synthetic Minority Oversampling Technique (SMOTE), checking the correlation using heatmap, scaling the data using standardize (Std) and normalize (Nor) methods and experimenting the results with outlier removing and outlier imputation from median and not changing the outliers were added rather than initial model [Task 3.2 ].

Finally, the analysis was performed using the hyperparametric tuning with five folds under all predefined data preprocessing techniques without using unscaled data to get the best parameters to improve the model performance [ Task 3.3 ].

Table 1 represents overall accuracies after doing different type of experiments using preprocessing and tuning methods.

# 4. Conclusion and Recommendations

TABLE 1 : Results After Only Using the Data Preprocessing Techniques

| | Only using data preprocessing techniques | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | With Outliers | | | Imputing; Median | | | Remove Outliers | | |
| | Std | Nor | Unscaled | Std | Nor | Unscaled | Std | Nor | Unscaled |
| LR | 73% | 72% | 72% | 70% | 69% | 69% | 74% | 74% | 74% |
| NB | 69% | 69% | 69% | 69% | 69% | 69% | 70% | 70% | 70% |
| NN | 69% | 72% | 72% | 69% | 72% | 67% | 70% | 73% | 71% |

TABLE 2 : Results After Hyperparametric Tuning

| | Using Hyperparametric Tuning with Data preprocessing | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Imputing Median | | | Keep Outliers | | | Remove Outliers | | |
| | Std | Nor | Unscaled | Std | Nor | Unscaled | Std | Nor | Unscaled |
| LR | 73% | 71% | | 71% | 70% | | 72% | 70% | |
| NB | 70% | 69% | | 67% | 67% | | 69% | 71% | |
| NN | 71% | 72% | | 72% | 71% | | 70% | 71% | |

According to the final results we can conclude that the best accuracy which is the best prediction rate of a player will last five years in NBA. According to the TABLE 1 and TABLE 2, those models were explored by the analysis and 74% is best prediction rate of a player will last five years in NBA. Logistic Regression algorithm will give best predictions after doing relevant preprocessing steps according to TABLE 1.