Classification using Logistic Regression, Gaussian Naive Bayes and Neural Networks for Prediction

Introduction

This report employed the use of 3 different machine algorithms Logistic Regression, Gaussian Naive Bayes and the construction of Neural Networks in the classification of the NBA rookie data. The machine learning algorithms were used to predict if a player would last 5 years in the NBA. The data contains several features used to evaluate rookie performance with the target variable being Target_5Yrs with 1: if career length >= 5 yrs or 0: if career length < 5 yrs. Find below a summary of steps used for the analysis.

Exploratory Data Analysis

Python was used for our data analysis and several libraries relevant to the analysis were imported.

The data was loaded, read and explored using several functions like data.head(), data.info(), data.describe() to get adequate summary and understanding of the data.

The data. shape function was used to get the dimensions of the data. This revealed that the data had 1340 rows and 21 columns.

Data Cleaning

To ensure data accuracy and consistency, the following data cleaning processes were done.

The ‘Name’ column was dropped as it wasn’t useful to our analysis.

Null values were also checked for, and it was identified in the ‘3 Point Percent’ column and replaced with 0.

Checked for outliers using the Interquartile range and visualised with box plots of the variables.

The independent variables were transformed using the label encoder.

Standardization of the data was done using Standard Scaler. This involves transforming the features of a dataset so that they have a mean of 0 and a standard deviation of 1. This is because different features of the dataset have different scales, and it is to prevent features with larger magnitudes from dominating the algorithm.

Data Analysis

Feature Selection: The features to be used for the model were selected with ‘target_5Yrs’ being the y since it is the target variable.

Training the Model: The dataset was divided into training and test sets using a 1:3 ratio. The training set was 1/3 of the test sets. The model was then fit into the training set. The 20/80 ratio was also experimented with in training the model, but the results show that the 1/3 ratio gave the best model.

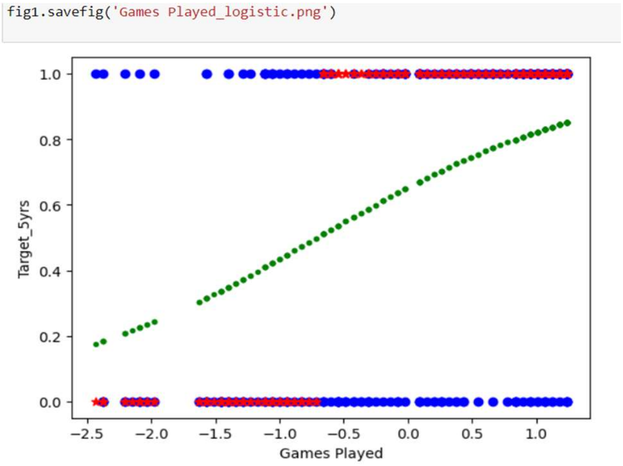
Several hidden layers were experimented with in constructing the neural network, but as shown in the results, 3 hidden layers of 10,50 and 20 gave the best results. The logistic (sigmoid) activation was used because this is suitable for binary classification. Other activation gave less accurate results and more mislabelled points. Reducing the train/test size to 20: 80 also gave less accurate results.

Results and Discussion.

The table below shows a summary of the results for the different analysis carried out using the 3 machine learning algorithms. Each model was evaluated for its effectiveness, and the most effective models are highlighted in the results shown below. Using all the independent variables in the data set gave the most effective result using logistic Regression and when constructing a neural network. This means that all the features are very important to predict if a player would be able to stay in the NBA for 5 years. From the results, games played and blocks are also 2 important features that would help make this prediction as prediction using these variables were 70% accurate using the 3 machine learning algorithms. According to Mikołajec, Maszczyk and Zając (2013), some of the game indicators that determines a players' performance in the NBA include offensive rebounds, steals, and points per game. When these variables were used, it was just moderately accurate. 68%(logistic regression, 69%(GNB)

	Logistic Regression		GNB		Neural Network		
Features	Accuracy	No of mislabelled points	Accuracy	No of mislabelled points	Hidden layer	Accuracy	No of mislabelled points
All features	0.71	128/447	0.67	146/447	10,50,20	0.73	122/447
Points per Game, Assists, Rebounds, Field Goal Percent, 3-Point Percent, Free Throw Percent, Minutes played, Steals, Turnovers	0.7	135/447	0.68	142/447	10,50,20	0.69	138/447
Offensive rebound, Points per game,	0.68	142/447	0.69	140/447	10,50,20	0.68	141/447

Steals							
Games played	0.68	142/447	0.68	142/447	10,50,20	0.68	141/447
Defensive rebounds	0.66	153/447	0.67	146/447	10,50,20	0.65	155/447
Free throw percent, Assist	0.67	148/447	0.59	182/447	10,50,20	0.60	178/447
Games Played and Blocks	0.70	136/447	0.71	131/447	10,50,20	0.71	131/447



```
fig1.savefig('Games Played_logistic.png')
```

Fig 1: Prediction with Games Played using logistic regression

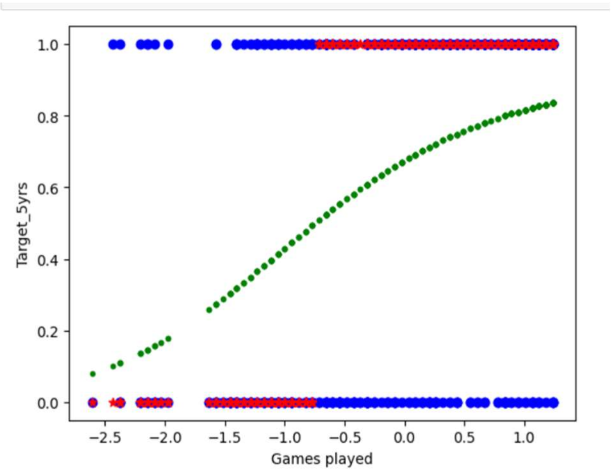
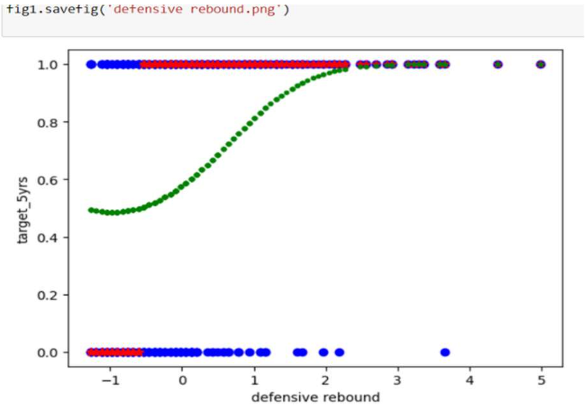


Fig 2: Prediction with Games Played using   GNB



```
fig1.savefig('defensive rebound.png')
```

Fig 3: Prediction with Defensive rebound Logistic regression

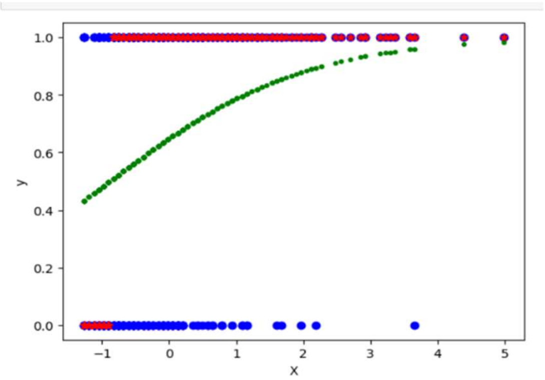


Fig 4: Defensive rebound using

Using GNB.

Conclusion

Using all features gave us relatively good accuracy using the 3 machine learning algorithms. This shows that all the features are important to predict if a player would stay in the NBA for more than 5 years.

References.

Mikołajec, K., Maszczyk, A. and Zając, T. (2013) 'Game indicators determining sports performance in the NBA', Journal of human kinetics, 37(1), pp. 145–151. doi: 10.2478/hukin-2013-0035.