# EXPERIMENT REPORT

|  |  |
| --- | --- |
| **Student Name** | Justin Mah |
| **Project Name** | Assessment Task 1 |
| **Date** | 13 February 2022 |
| **Deliverables** | Notebook name:  mah\_justin-13290085-week1\_logistic-knn-random\_forest.ipynb  (Note the model is saved in the Group3 Github account under Jasleen Kaur’s Github Account. All group submissions are in that account)  Models used:   * Logistic Regression * KNN * Random Forest Classifier |

## **Experiment Background:**

## The business Objective

The objective for this assignment experiment is to predict a rookie player from the National Basketball Association (NBA) to determine whether the player will last a minimum 5 years in the NBA game basing on the statistics received from the dataset.

The prediction results will be used most likely by sports commentators and fans to guess how well the rookie player will perform in the future. Also, it is possible this prediction could be used to evaluate feedback to interested stakeholders of the rookie player for other decision-making purposes.

This experiment may not be accurate as there is the time factor we need to consider for each player. For example, if the player starts playing for that year is this all players or some. Which means there could be a bias since there is plenty of statistic data on some players and few on others.

Another factor we need to consider is the players time on the basketball court compared to players on the reserve bench or injuries that he may have during time away from the basketball court. The data we have does not factor time for each player to equally have their fair share demonstrating talent that could be recorded in this data.

If the experiment works well the impact will be substantial in how players are selected in the NBA. Sponsorship deals and gambling companies will use this information to enhance revenue and maximise winnings.

If this proof of concept is unsuccessful then the result would be inefficient players playing the big league but fall short of their teams’ expectations and sponsorship dealings. The interest of the game may falter which in turn lead to less spectators of one of the major sports in America. Incorrect results would place stakeholder in a deal of stress because of selected features provide wrong outcomes.

## Hypothesis

This is a classification problem as the goal is to see whether each NBA rookie is likely to be in the big league for at least 5 years in the game, using each the rookie statistics and scoring abilities in the past games that they have played seems to be a good dataset to experiment on and worth considering.

The test we would likely to test is which model can demonstrate the best in predicting NBA rookies for at least five years using the rookie’s performance and discovering which metric is the best to determine a rookie’s ability to stay with the NBA for the period of five years minimum.

I want to experiment whether a slam dunk vs a three-point shot or offensive or defensive position will be likely determinant for a players career longevity. Plus it could give a possible snapshot of understanding how NBA rookies are selected by analysing past data and the characteristics of past rookies who have stayed on the professional for more than 5 years.

## Experiment Objective

The outcome of this experiment is to show which model can perform the best in predicting players maintaining their careers in the NBA for 5 years minimum. We are also analysing whether the dataset itself is enough to maintain a suitable and reliable source of information for this experiment, the understand the pros and cons of each model’s performance. I am hoping that we can get a reasonable amount of accuracy and understand which features in the dataset provide false positive and misleading results.

There could be several possible scenarios from this experiment:

* The objective of this experiment may need to be reframe (re-evaluated) based on the data presented to provide a resolution.
* We may require additional information from other reputable sources to provide accuracy.
* The models used may not provide a conclusive result for the experiment

**2. Experiment Details**

Data Preparation

In this experiment retrieved from Kaggle we have two datasets train.csv and test.csv. The train data has 21 features, and 8000 records entries and test data have 20 features and 3799 records, note that there is one less feature due to the fact we do not have a target variable.

First, I checked the size of each dataset, first checking if the records are consistent, if there is missing (null) entries, what data types identified so I can determine whether it is classification or regression model.

From the objectives and reviewing the target variable it is a classification algorithm to be used.

Next step was to understand what the columns represent, as you can see the list below is the data glossary highlighting the definitions of the 22 features. Please note that no.1 “Id\_Old” does not exist on both the train and test datasets and no.22 “TARGET\_5Yrs” only exists in train dataset only.

|  |  |  |
| --- | --- | --- |
| **Data Glossary** | | |
| **No** | **Column Names** | **Comments** |
| 1 | Id\_old | Previous Player Identifier (Not on datasets) |
| 2 | Id | Player Identifier |
| 3 | GP | Games Played |
| 4 | MIN | Minutes Played |
| 5 | PTS | Points Per Game |
| 6 | FGM | Field Goals Made |
| 7 | FGA | Field Goals Attempts |
| 8 | FG% | Field Goals Percent |
| 9 | 3P Made | 3-Points Made |
| 10 | 3PA | 3-Points Attempts |
| 11 | 3P% | 3-Points Percent |
| 12 | FTM | Free Throw Made |
| 13 | FTA | Free Throw Attempts |
| 14 | FT% | Free Throw Percent |
| 15 | OREB | Offensive Rebounds |
| 16 | DREB | Defensive Rebounds |
| 17 | REB | Rebounds |
| 18 | AST | Assists |
| 19 | STL | Steals |
| 20 | BLK | Blocks |
| 21 | TOV | Turnovers |
| 22 | TARGET\_5Yrs | Outcome: 1 if career length >= 5 years, 0 otherwise |

Furthermore, in the Train dataset I have separated the target variable “TARGET\_5Yrs” from the training dataset so I can run the train and validation before using the prediction and compare the performance against the target variable.

## Feature Engineering

There does not seem to need for data cleaning at this stage, the next step was to focus on the train data because this is where we are going to build our machine classification algorithms on. Below I ran a correlation plot to understand the strength of each comparing features as the data all numerical and no categorical values were found.

*Correlation Plot of Train data*

Chart, treemap chart

Description automatically generated

We identified that “id” should be index or removed from the feature as it is a primary key identifier for the rookie players. When comparing features against TARGET\_5Yrs the predictor variables: 3P Made, 3PA, 3P%, FT%, AST and BLK have very poor association with the target variable.

I have created a list called *feature\_selection2* for these to remove and experiment in the models, however they did not improve the model’s performance by using ROC\_AUC score as our performance metric which we shortly explain later.

*feature\_selection2* (features dropped) =3P Made, 3PA, 3P%, FT%, AST and BLK

## Modelling

Reviewing the training data, I decided to select a number of models:

* Logistic Regression
* K Means Classifier
* Random Forest Classification

For Logistic regression and K means before a start modelling each feature has a different metric used to calculate the players performance. In the interest of keep it fair I used standard scaler the data to prevent other features from dominating others which in turn affect the model to learn.

I train test split the training data 80/20 with random state of 42 for the training dataset, for the logistic regression I have included ‘liblinear’ in the model for hyperparameter tuning (L1 and L2 penalty).

I used all the predictor variables to experiment and see what performance I get for each of the model.

For logistic regression I tried using all predictors and dropping some selected predictors (*feature\_selection2*) to evaluate the ROC\_AUC score however they show no improvement of the model performance.

Given time I want to improve and run other models like AdaBoost to see which performs better. In addition understanding more on types of regularisation techniques.

## **Experiment Results**

## Technical Performance

We used AUC-ROC performance metric for our model as this experiment is a binary classification problem. It measures how well the model behaves against the true positive vs false positive.

Here are the results of each model:

**ROC-AUC Score**

Logistic Regression 0.7166091726838646

Logistic Regression (selected features) 0.7139728863496108

K Means Classifier 0.6262516672307049

Random Forest Classifier 0.515322160793

Surprisingly Logistic Regression is the better model and at this early stage I will use other regularisation techniques and further feature engineering analysis to determine the root cause of poor performance of the models.

## Business Impact

At this moment this is an experiment however if given the go ahead the performance of the logistic regression is 71% probable in choosing NBA rookies who would likely to be in the organisation for minimum 5 years.

This is still work in progress and hopefully improve of the performance outcome for the following weeks. I do suspect their will be high chances of multicollinearity because for one feature to work like points scored there is a chance of correlation with time in game and free throw attempts for example.

## Encountered issues

I am unable to improve the performance of each model for the time being, I will need to review the selected features, hyperparameters and regularisation techniques to see any improvements in the score metric.

There is problem I believe that is not represented in this data, that is some players do not perform straight away however they perform better during later years which this experiment cannot provide. It is only in this moment of time through a represented timeframe that we predict on the chances of rookies will stay in the NBA for 5 years or more.

## **Future Experiment**

## Key Learning

The outcome learnt from the experiment is initial techniques used to understand the quality of the dataset. Using machine learning techniques in python and sklearn package and what options that could enhance the performance of the model.

## Suggestions and Recommendations

I will need to do further research on regularisation if it will improve the model in addition I want to experiment with other models such as Ada Boost, XGBoost, Decision trees and other classification models.