# EXPERIMENT REPORT

|  |  |
| --- | --- |
| **Student Name** | Justin Mah |
| **Project Name** | Assessment Task 2 |
| **Date** | 20 February 2022 |
| **Deliverables** | Notebook name:  mah\_justin-13290085-week2.ipynb  (Note the model is saved in the Group3 Github account under Jasleen Kaur’s Github Account. All group submissions are in that account)  Group Github: https://protect-au.mimecast.com/s/BqnUCYW8n2I9Pjo3uG4IsR?domain=github.com  Github: <https://github.com/justinmuts/adv_dsi_at_1/notebooks.git>  Models used:   * Cross validation on: * Decision Tree Classifier, * K Nearest Neighbours, * Logistic Regression * XGBoost * Logistic Regression using SMOTE oversampling technique using: * SMOTE * Borderline SMOTE * SVM SMOTE and * ADASYN (Adaptive Synthetic) * Grid search cross validation Logistic Regression * K Nearest Neighbours Classifier using SMOTE with Principal Components Analysis |

## **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

## Week 1

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.

## Week 2

In this experiment for week 2 I wanted to concentrate on the imbalanced dataset for the train data. The reason behind this was because during the first initial run of the experiment using Logistic Regression, K Nearest Neighbours and Random Forest classifier, I was not getting much improvement on the performance AUC ROC metric.

The number of options that I came across was either we do a manual feature selection by examining the correlation plots between each variable and identifying which variable has the closest association with the target variable.

Or we use imbalancing data techniques which is this week’s focus by looking at using SMOTE oversampling by artificially inflate on the minority data to have equal balance.

Also using number of hyper parameters tuning such as Grid Search, Cross Validation and possibly random search. By doing so on the training data before running the model will make sure we remove any biases on the model when it comes to predicting the test data.

## 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.

The outcome of this experiment for this week 2

* Using different techniques in solving the imbalance data to improve the performance of each classification model and using several different types of models to test and identify the best model by using the AUC ROC metric.
* Test some hyperparameters in the models to tweak the performance of new and existing model

The possible scenario for this week would be a huge bump on the model performance in better predicting NBA players who would last up to 5 years in the NBA.

Otherwise, I could be expecting no change or worse performance than last week results. I have a guess that it is understanding the background of the definition of each variable and further test which variables should be dropped to improve predictions on players longevity in the NBA careers.

**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.

**Week 2**

I do believe there is a high chance of multicollinearity between each predictor variables which is not giving us much performance as we expected. It is possible I need to perform test like principal components analysis to reduce and select which important feature for our model predictions.

Table

Description automatically generatedTraining Data (target variable “y”)

The image above shows that we have an imbalance training dataset that could skewered cause the model to be biased. Hence this is the reason we should look at machine learning techniques to overcome them in the hope of better model improvement.

## Feature Engineering

In week 2 experiment we are going to investigate how to improve the AUC\_ROC score for the existing models by performing a series of hyperparameter tuning and resolving imbalance data to generate higher balance accuracy models and balanced detection rate.

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

In week 1 I decided to use these three classification models

* Logistic Regression
* K Means Classifier
* Random Forest Classification

For week 2 I will be using of cross validation to confirm issues of overfitting and bias in evaluating model performance, Synthetic Minority Oversampling Technique (SMOTE) for increasing the number of cases in our dataset to balanced training data for modelling and improve AUC ROC score performance.

The models I am experimenting on are:

* Cross validation on:
* Decision Tree Classifier,
* K Nearest Neighbours,
* Logistic Regression
* XGBoost
* Logistic Regression using SMOTE oversampling technique using:
* SMOTE
* Borderline SMOTE
* SVM SMOTE and
* ADASYN (Adaptive Synthetic)
* Grid search cross validation Logistic Regression
* K Nearest Neighbours Classifier using SMOTE with Principal Components Analysis

I chose cross validation as I want to examine quickly which model performed the best and worst and identify the patterns of possible overfitting and bias especially with lower auc scores for each of the models in cross validation.

I used repeated stratified k fold to ensure that each fold of dataset has the same proportion of observations with a given label, with parameters of 10 splits, repeating 3 times with a random state of 42.

I want to compare with the new models like XGBoost and Decision tree classifier against the previous models used last week to see how well they fare.

The next experiment is to use Logistic Regression model with SMOTE oversampling given the fact there was an imbalance dataset of players 79% of were successful staying on with the NBA for at least 5 years while the remaining 21% rookie players did not make in the NBA.

Logistic regression performed well against other classification models, and this is the reason I wanted to give this model an additional boost to its performance auc roc metric.

I was not familiar of each of the SMOTE techniques, and I wanted to find out how well they fare and why they performed in that fashion.

I used SMOTE combats imbalanced data by randomly create oversampling the minority data in the training dataset, I want to make sure the data is balanced as possible to allow the model to prevent any bias in its classification decisions.

The next technique is Borderline-SMOTE which only makes synthetic data along the decision boundary between the two classes. But its weaknesses are that it brings more attention to these extreme observations as they commonly exist away from the classes centre.

SVM-SMOTE focuses on creating oversampling the minority class instances near borderlines but using SVM (support vector machines) to help establish boundary between classes.

The artificial data will be randomly created along the lines joining each minority class support vector close to its nearest neighbours. This technique avoids class overlap and seek attention on where the data is separated.

ADASYN focuses in producing oversample data on hard to learn areas you to sample more undesirable data for the model. It is adaptive but at the expense of accuracy and generates synthetic data where majority class are located which could produce false positives.

I wanted to experiment Grid search cross validation Logistic Regression to identify the best parameters for this model and returns the best performance score. I added penalties to the model with L1, L2 and elastic net regularisation techniques.

Finally, the last late experiment is K Nearest Neighbours Classifier using SMOTE with Principal Components Analysis. In this test I used PCA to reduce the dimensionality of the predictor variables after smote to balance the training data before KNN model get a better performance metric.

I did not have time to run AdaBoost model but given time I may see if it is worth the

## **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 from week 1:

**ROC-AUC Score**

Logistic Regression 0.7166091726838646

Logistic Regression (selected features) 0.7139728863496108

K Means Classifier 0.6262516672307049

Random Forest Classifier 0.515322160793

Here are the results of each model from week 2:

**ROC-AUC Score**

Decision Tree Classifier Cross validation 0.541

K Nearest Neighbours Cross validation 0.588

Logistic Regression Cross validation 0.703

XGBoost Cross validation 0.655

Logistic Regression using SMOTE 0.678

Logistic Regression using Borderline SMOTE 0.681

Logistic Regression using SVM SMOTE 0.678

Logistic Regression using ADASYN 0.678

Grid search cross validation Logistic Regression 0.500779

K Nearest Neighbours Classifier using SMOTE 0.5855

with PCA

For week 2 again logistic regression is no surprise that it performed well however I rather put my belief towards models with SMOTE because the data has shown that there is imbalanced data. A bias towards NBA rookie players 79% staying in the game for 5 years and over. For a data scientist it is not a good way to perform model with a skewed dataset.

It seems Logistic Regression using Borderline SMOTE was the best performing category for this week not worse than last week with the highest AUC ROC score of 0.716 compared to 0.681.

I believe the unperforming scores is probably more to do with the selection of predictors in addition I believe that we may need to eliminate certain records of players with a negative scores and re-evaluate is the negative score got to do with team performance while the player is out of the game or individual performance. I believe data cleaning needs to be looked into.

## Business Impact

The dataset is clearly biased and the models have shown a poor outcome of predicting players at this moment. We have demonstrated that we needed to perform feature engineering to make the data we predicting as fair to predict the probability of players with high accuracy as possible. I understand this weeks performance was not as expected but it was necessary to prevent bias predictions or wild predictions by chance.

This is still work in progress and hopefully improve of the performance outcome for the following weeks. I do suspect there 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 more data wrangling on selected features, parameters, and regularisation techniques to see any improvements in the score metric.

There could be problem as some rookie players do not perform well straight away however they perform better during later in the years that may affect the outcome of the model.

## **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 understand more on tweaking the parameters of each models and hopefully have other models to experiment.