EXPERIMENT REPORT

Student Name	Priyanka Srinivasa
Project Name	MDSI ADSI Assignment 1_Part C
Date	21-02-2021
Deliverables	srinivasa_priyanka_13684182_week3_xgb_hyperopt.ipynb model name: Xgboost_hyperopt https://github.com/roger-yu-ds/assignment_1/tree/priya

---

1. EXPERIMENT BACKGROUND

Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.

1.a. Business Objective

The goal of the project is to predict if a rookie player in the National Basketball Association (NBA) will remain at least 5 years in the league. Measuring the success of rookie players and learn about the future possibilities on how long these young talents will last is an important question in sports analytics as they help the business side of sports to secure a competitive edge.

By predicting the career performance of a rookie player with the help of their performance statistics, the team management can make better decisions which improves their business by providing the organization an opportunity to win a championship

1.b. Hypothesis

In Week_1 experiment, a Logistic Regression model was used to predict the career length of NBA rookie players.

In Week_2 experiment, a Support Vector Machine (SVM) model was used to check if it can outperform the Logistic Regression model. The AUROC scores of both the models are given below.

Model	AUROC score - validation dataset	AUROC score - Kaggle
Week1 - Logistic Regression	0.72707	0.71034
Week2 - Support Vector Machine	0.73166	0.70595

Table 1

In Week_3 experiment, Xgboost model was used to examine if it can perform better than the previous Machine learning models (Logistic regression and Support Vector Machine) and provide better results in predicting If a rookie player will last at least 5 years in the league based on his performance statistics.

1.c. Experiment Objective	The outcome expected from this experiment is to find out if the end result agrees or differs from the hypothesis. Xgboost is known to reduce bias, has multiple hyperparameters for cross-validation, regularization, performs better with unbalanced data and they give more weight to the observations that are misclassified.

---

2. EXPERIMENT DETAILS

Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.

2.a. Data Preparation

- **Removed Id_old / Id** – these columns were removed as it captures the uniqueness of the players.
- **Missing / NULL values** – the data was examined for missing / NULL values to ensure the completeness of the data.
- **Handling Duplicates** – there were no duplicate values present in the data.
- **Class Imbalance** – Checked for class imbalance as it can influence the accuracy of the model.
- **Target variable** – the 'Target_5YRS' feature was assigned to a new variable and removed from the main dataset.
- **Splitting dataset** – the training data was split into 80% training data and 20% validation dataset using sklearn's train_test_split function with using a 'random state'=8. While splitting the dataset, 'stratify' = 'target' was specified in order to have the same proportion of class labels in the training and test subsets.

2.b. Feature Engineering

Feature engineered variables created.

Featured Variable	Formula	Description
ATR (Assists to Turnover ratio)	df['ATR'] = df['AST']/df['TOV']	To check the player's control on the ball.
FTR (Free Throw Rate)	df['FTR'] = df['FTA'] / df['FGA']	Indicated offensive efficiency.
TOV% (Turnover Percentage)	df['TOV%'] = 100 * df['TOV'] / (df['FGA'] + 0.44 * df['FTA'] + df['TOV'])	Estimated turnovers per 100 plays.

Table 2

An assumption was made that the above featured variables might improve the performance of the model and provide better predictions with a higher accuracy and AUROC score. However, 'ATR' and 'TOV%' variables had low correlation with the target variable and were dropped at a later stage. 'FTR' had a better correlation and was included in the model which resulted in providing a better AUROC score.

2.c. Modelling

In week_3 experiment, Xgboost was used to determine if an NBA rookie will last in the league for at least 5 years. Xgboost has options to fine tune the model using different choices of hyperparameters and cross-validation techniques.

Initially a basic Xgboost model without hyperparameter tuning was applied to the training and validation datasets. The results are as shown in table

Metric	value
Accuracy	0.82
Precision	0 - 0.37 / 1 - 0.84
Recall	0 - 0.08 / 1 - 0.97
F1	0 - 0.14 / 1 - 0.90
AUROC	0.63

Table 3

In order to further improve the model performance, hyperparameter optimization was performed using the python library 'hyperopt'. Hyperopt provided the best parameters for the model. The best parameter values are,

Best: 'colsample_bytree': 0.4
'learning_rate': 0.05
'max_depth': 7
'min_child_weight': 3.0
'subsample': 0.9

The model was then trained using the above parameters which provided a better **accuracy** of **0.83** and **AUROC** score of **0.67**.

---

3. EXPERIMENT RESULTS

Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

3.a. Technical Performance

AUROC (Area Under ROC) is the performance metric used to assess the model. The scores of the validation, and test set were proportional.

AUROC for validation dataset – **0.67828**
AUROC for test dataset on Kaggle – **0.69684**

The classification report for Xgboost model after hyperparameter tuning is as shown in Figure.1 which is an improvement compared to the Xgboost model without hyperparameter tuning.

	precision	recall	f1-score	support
0	0.39	0.03	0.05	266
1	0.84	0.99	0.91	1334
accuracy			0.83	1600
macro avg	0.61	0.51	0.48	1600
weighted avg	0.76	0.83	0.76	1600

Figure 1

The AUROC score obtained on Kaggle was lesser compared to the week_1 and week_2 experiments using Logistic Regression model and SVM model respectively. Thus, Xgboost model is not considered to be the most suitable choice for this scenario.

3.b. Business Impact

Accuracy of the results helps the sports analysts and NBA franchises focus on the players who are more capable of staying longer in the league and invest on improving the performance of those players. Given the results of this experiment, the opportunities for new businesses might become lesser and lower the wealth of the economy if the model is predicting a potential rookie player to not last in the league for 5 years. This results in getting rid of an excellent player by looking at the prediction even before the player had a chance to perform and prove his potential.

A confusion matrix was plotted for better understanding. As seen in the Figure.2, the model predicts only 7 players will remain the NBA league for at least 5 years. This can be misleading to the NBA sponsors and also lose potential players who can perform better in the future, as the results predict that 1323 players will not last in the NBA league for 5 years.

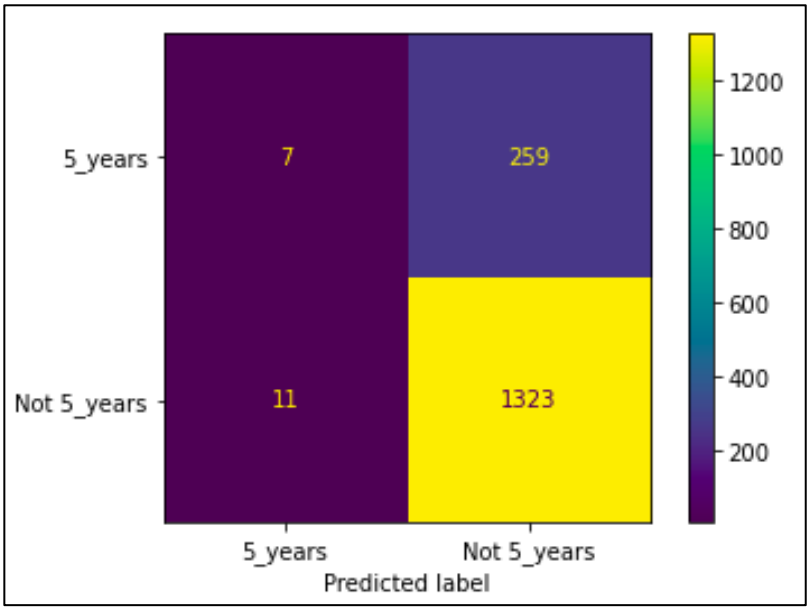

Figure 2

3.c. Encountered Issues

Issues	Solution
Imbalanced data	**Solved**: while splitting the dataset, the parameter 'stratify' = 'target' was specified to have the same proportion of class labels in the training and test subsets
Hyperparameter tuning	**Solved**: Hyperopt was used to get the best parameters to train the model. The results obtained were better than the Xgboost model which was without specifying the hyperparameters.

Table 4

4. FUTURE EXPERIMENT	
Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.	
4.a. Key Learning	Outcome / Insights gained from this experiment Xgboost has a wide range of options to tune the model to get better results. Creating feature engineered variables can improve the model performance to a certain extent. However, some of the feature engineered variables could also make the models perform poor. Hence it is important to select the right features to get good results.
4.b. Suggestions / Recommendations	• Since Logistic Regression with SMOTE provided the best results compared to SVM and Xgboost models, I would like to explore more options to check if I can improve the results of Logistic regression by using more feature engineered variables and try to eliminate the less significant variables. • My teammate used a voting classifier model and I would like to try that in the next experiment, as this week I wanted to experiment with Xgboost since I assumed it would perform better given its ability to improve the performance of the model by giving more weight to weak variables. • If the experiment achieved the required outcome, the solution can be deployed into production using Docker.