EXPERIMENT REPORT

Student Name	MAHJABEEN MOHIUDDIN(St_Id:24610507)
Project Name	NBA Draft (Part – D)
Date	7-09-2023
Deliverables	<NoteBook> Mohiuddin_Mahjabeen_24610507_week4_ MLPClassifier_ NBA_draft.ipynb **<Model>:**MLP Classifier with Custom Function **Custom Package Link:** !pip install uts_mahe_binaryclassifier_24610507 **<Github_link>:** https://github.com/MAHJABEENMOHIUDDIN/Adv_ML_Project_Community_Prediction_Competition/tree/master

---

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

Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?

The NBA Draft-Kaggle Competition

- The annual basketball event named "NBA Draft" is organised in the US to select the best team players from American colleges as well as international professional leagues to join their roster.

	• The organisers decided to select the team players based on their previous performances. • The selection of each player depends on the auc-roc score of the predicted models, which gives the probability of a player getting selected in the NBA Draft. • If the model gives the au-roc score more than 90%, then it will help the business select the best players, whereas if the score is below or equal to the threshold value, then it means the model's performance is poor.
1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, • **Hypothesis**: The supervised machine learning binary classification model will help the business by producing an accuracy and au-roc score closer to 1 for selecting the best players for the NBA draft event. • analysing how well the model performs the prediction. The accurate predictive model helps the business select the best player for the NBA Draft based on their previous performances. This historical result helps the models predict the best outcome.
1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. • **Expected outcome:** The binary classification model, after fitting the data into class and custom functions, will produce an accurate predictive model that will help the business gain profit by selecting the best players. The players who could acquire a high scores will make the basketball match interesting, and it can also attract many audiences. • The goal of the project is to evaluate the probability of a player being selected for the team by fitting the data to train a model. • **Scenarios:** 1. There is a chance that data will be learned from all the classes.

	2. There is a chance of models becoming biased. 3. There is a chance that the model may overfit or underfit that data. 4. There is a chance that the model may provide an au-roc value close to 1 and result in a predictive model.

---

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	• Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments. **Data Processing:** • Loading the csv files of the training and testing datasets into the train dataframe and test dataframe. • Finding out the dimensions of the training and testing dataframes • Copying the dataframes into new variables named cleaned_train_df and cleaned_test_df before cleaning the data. • Initially there were 185037 nan values in the selected features and these nan values are replaced using the custom functions such as median_null, mean_null and replace_null_with_Zero • The nan values in the features 'Rec_Rank', 'ast_tov', 'rimmade', 'rimmade_rimmiss' and 'midnade_midmiss' are replaced by median value of each respective columns using custom function "**median_null**". The remaining nan values are 128717. • The nan values in the columns such as 'Rec_Rank', 'ast_tov', 'rimmade', 'rimmade_rimmiss', 'midmade_midmiss' are replaced by the mean value of each column by using custom function "**mean_null**" and remaining nan values in the dataframe are 5824. • The remaining nan values are replaced by 0 using the custom function "**replace_null_with_Zero**".

	• Duplicated columns are analysed using the function **"duplicated()"** and found there are 16 duplicated rows in the training dataset and are dropped using drop.duplicated() function. • **Important Step for future:** It is important to copy the original data before starting the cleaning process because there is a risk of any portion of the data getting deleted, which will give inaccurate predictions. • **Steps excluded:** Ordinal encoding and one-hot encoding are not useful for the data available in train and test datasets. The features on which encoding can be performed are identifiers, and others are irrelevant for prediction. To train the data, identifiers are not used to avoid inaccurate predictions. The Feature engineering process to create a new feature is ignored as it was creating some errors in the code.
2.b. Feature EngineeringM	Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments. **Dropping of features:** • There are a few features that are irrelevant to calculating the probability of a player's performance, and some of them are also identifiers. Those features are: name of conference (conf), student's year of study (yr), height of student (ht), player number (num), points over replacement per adjusted game (porpag), adjusted offensive efficiency (adjoe), type of metrics displayed (type), season's year (year), defensive rating (drtg), and teams. • **Dropping duplicate rows from test and train dataframes:** There are 16 duplicate rows in the training dataframe and 4 duplicate rows

in the testing data frame; training dataset duplicates have been removed using the Pandas duplicate() function.

StandardScaler:

- StandardScaler is a function of sciket learn, it had help the data to scale and center the numeric variables.
- Scaling the data to resize the distribution of values to make the mean of the observed values 0 and the standard deviation 1.

SMOTE sampler:

There were 56091 observations in the dataset, the processing time was too long, and there were many limitations of RAM.

- **Sampling:** Sampling the training dataframe using the SMOTE sampler with hyperparameters such as "sampling_strategy='auto', random_state=42, k_neighbors=5, n_jobs=None.
- The SMote sampler helps in balancing the unbalanced classes by combining the features of the target cases with the features of their neighbours.

Splitting features and target :

The column drafted is a target variable in the dataset and it is assigned to y, where as the remaining variables are considered as features and assigned to X using a custom function **drop_target().**

- **Splitting the data frame using a custom function:** The training data frame is split into training and validation sets using the custom function **"random_split_sets"**. Where training data is 80% and validation data is 20%.
- Using a validation set to evaluate your model on unseen data to increases the generalising capability of the model.

	The custom class and functions generated: ▪ NullBinaryClassifier class ▪ Drop_nan_values. ▪ Replace_null_with_Zero ▪ Mean_null ▪ Drop_target ▪ Random_split_sets ▪ Save_sets ▪ Load_sets ▪ Fit_assess_classifer ▪ Assess_classifier_set ▪ Print_classifier_scores **Important features for the future:** The features such as number of minutes played by the players and the number of games played (GM) and the performance of the player in the school are important, as this can help in analysing the capabilities of a player.
2.c. Modelling	Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments. • **Exploratory Analysis**: With the help of histogram analysing the performance level of players using the columns “GP”, “Min_per”, “rimmade”, and on “drafted”. • The testing data frame is left after cleaning the data in it, to access it later in order to assign its player id column with the obtained training dataframe probability scores. • **Baseline Model**: • **Custom Class: NullBinaryClassifier:** With the help of the custom class "NullBinaryClassifier", analysing the baseline of the model. • **Custom Function: print_classifier_scores:** This custom"print_classifier_scores" function is used to print the accuracy score and f1 score of both the train set and the validation set.

- The baseline model is the dumb model, and it is computed by applying the mode function to the target variable for classification models. The base line score can be analysed using the accuracy score matrix. The result of the baseline model should be low compared to the training set. If the training set has a score closer to 1, whereas the baseline model has a score less than the training set, then this means the model beats the baseline model; otherwise, it is stated that the model cannot beat the baseline model.

Model:

The Multi-Layer Perception (MLP) is the most common type of neural network. It is a function that maps the input to the output. It has single input, output layer and between them there can be any number of hidden layers.
The same set of neurons are present in both input layer and features, whereas hidden layer can has more than one neuron.
Every neuron is a linear function on which activation function is applied to solve complicated problems.

The output from each of the layer is given as an input to all the neurons in the subsequent layers

The hyperparameters used are:
Max_iter =300: It is used to give maximum iterations and assigning it to 300 iterations.
solver='lbfgs': The solver is used for weight optimization, and “ibfgs” is an optimizer in the family of quasi-Newton methods.
alpha=1e-5: It is a strength of L2 regularization term. This L2 regularization is dividing the sample size when adding to the loss.
hidden_layer_sizes=(5, 2): It represents the number of neurons in the hidden layer.
random_state=1: Assigning 1 to get same splitting sets when each time the code is run.

The custom function “fit_assess_classifier()” is used to fit the data, predict the data, testing and printing accuracy and f1 score.

	Models avoided: The model built using linear regression and algorithms that come under unsupervised machine learning should be avoided because here I am training the models using supervised machine learning binary classification, which has binary labels.

---

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	Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes. • **NullBinaryClassifier class :** The NullBinaryClassifier class is used to test the baseline model **Baseline Model Score**: 0.500 (accuracy score) **Multi Layer Perception Classifier:** **MODEL :** **Accuracy Score:** Training Set Score: 0.891 Validation set score: 0.895

	F1 score : Training Set Score: 0.890 Validation set score: 0.894 • **The Roc score obtained from Kaggle is 0.9498.** • • The performance matrix clearly indicates that the baseline score is low compared to the training set, and therefore, we can say that the model beats the baseline model score. • Comparing the training and validation scores, it can be concluded that the validation sets areslightly underfitting the data slightly. • The Roc Score is **0.9498**, which means the model has accurately predicted the probability of players getting selected in the NBA Draft based on their previous performances.
3.b. Business Impact	Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others) **Results:** The alternative hypothesis is true. The model is able to predict the probability of players being selected in the NBA draft based on their historical performances. These results clearly show that the model has given the best results and significantly good. This result, if used to select the players, will lead to earn more profit in the business by selecting eligible players in the team. The business will gain popularity and the number of audiences if the capable players show up in the match.

3.c. Encountered Issues	List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments. **List of issues:** • Selecting the suitable code to do the exploratory data analysis. • Decision the new functions to add in the update version of the custom package. • Uploading the custom package on tes.pypi.org and importing the custom package into the notebook to access the custom-defined functions. • Selecting the unused supervised machine learning algorithm and its relevant hyperparameters to obtain the probability results for uploading the score csv into Kaggle to get the Roc score, multiple attempts were made to manipulate the hyperparameters. • To reach the desired results, multiple models were made using various algorithms. **Issue in the future:** improper selection of algorithms and their hyperparameters may be one of the issues that has to be faced in the future.

---

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	Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. • **Outcome of the experiment:** Experimenting various models to get the au-roc score 1 has lead to get the result of 0.994. The insight gained from the model is that models have learned the patterns of unseen data, it could fit the model more accurately, and could easily distribute the data correctly among the classes.

	Therefore, it can be declared that this model will help the business to select the best players for the NBA draft event.
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. The best obtained from the model is one of the best results that had reached the accuracy and f1 score to 0.8% and au-roc score to 0.994 Therefore, the model has reached to the best score, it is recommended to deploy in the production.