EXPERIMENT REPORT

Student Name	Vu Ngoc Anh Le - 14195743
Project Name	Kaggle Competition – NBA Draft
Date	20 Aug 2023
Deliverables	**Jupyter Notebook:** le_vungocanh_14195743_week1_baseline le_vungocanh_14195743_week1_logisticregression **Github Repo:** https://github.com/anhlevn149/adv_mla_at1

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 is an annual event in which teams select players from their American colleges as well as international professional leagues to join their rosters. This project aims to build a machine learning model which can generate the probability of being drafted for each player. To do this, a number of machine learning models will be used. Ultimately, the best model with the most accurate probability will be selected. The results will be useful for sport commentators and NBA fans as they can have predictions of which players that will be drafted for the next season.
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. I want to test a hypothesis if **Logistic Regression Classifier** model is able to accurately predict the probability of a player being drafted, taking into account of all available features, except for the ones with object datatype and those with missing values consisting of > 50% total values. At this stage, feature elimination has not been taken into account.
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. As target value 0 is highly dominant in the training dataset (99%), the baseline model is expected to predict value 0 all the time. The Logistic Regression Classifier model is therefore expected to have overfitting issue.

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

1) Remove columns with object datatype from the dataframe, including 'team', 'conf', 'yr', 'ht', 'num', 'type', 'player_id'
2) Remove columns with missing values containing > 50% of the total values, including 'Rec_Rank', 'dunks_ratio', 'pick'
3) Fill missing values in the dataframe with their mean values
4) Split the training dataset:
   - Split X_data and y_data → X_train, X_val, y_train, y_val with test size of 20%
5) Scale data:
   - Apply StandardScaler method on X_train, X_val, X_test

2.b. Feature Engineering

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

No feature engineering was performed in this experiment.

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

I used the Logistic Regression Classifier model to train the dataset. Although the model performed better than the baseline model, the improvement was minimal.

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.

The AUROC scores of the training and validation dataset are 0.9891360804890494 and 0.9874001668651553 respectively. Although the AUROC scores are high, they are biased towards the ‘not drafted’ variable. The reason is due to the imbalanced data of t he target variable in the training dataset.

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) As the model will always generate predictions that a player will not be eligible, the results are meaningless and sport commentators and fans are not able to make any predictions. Hence the project objective is not achieved.
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. I am a newbie to Github and in this project, it took me some time to get familiar with its workflow. I hope I will get better at it as practising more and more.

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. With high AUROC scores on both training and validation datasets due to biased dataset, Logistic Regression Classifier model did not generate accurate predictions to achieve business objectives. Therefore, more experimentation will be implemented to find the optimal model.
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. Next step is to deal with the imbalanced dataset, perform feature engineering to find relevant features, and train them with other models.