EXPERIMENT REPORT

Student Name	Vu Ngoc Anh Le - 14195743
Project Name	Kaggle Competition – NBA Draft
Date	03 September 2023
Deliverables	**Jupyter Notebook:** le_vungocanh_14195743_week3_baseline le_vungocanh_14195743_week3_randomforest **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 **Support Vector Machine Classifier** model is able to predict the probability of a player being drafted more accurately than the Logistic Regression and Adaboost in Experiment 1 and Experiment 2.
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. This Experiment is expected to perform better than the Logistic Regression Classifier model in Experiment 1 with higher AUROC scores on training and validation datasets.

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 that are irrelevant to the prediction task, including 'type', and 'player_id'
2) Column ‘yr’: Replace invalid values with NaN values
3) Column ‘ht’: Convert data shown in date format to float data type
4) Column ‘num’: Replace any non-numeric values with NaN values
5) Remove columns with missing values containing > 50% of the total values, including 'Rec_Rank', 'dunks_ratio', 'pick'
6) Apply LabelEncoder technique to categorical columns, including ‘team’, ‘conf’, ‘yr’
7) Fill missing values in the dataframe with their mean values
8) Apply SMOTE oversampling technique to address the imbalanced dataset
9) Feature elimination with correlation matrix to select the most relevant features of the dataset
10) Scale data:
    - Apply StandardScaler method on X_train, X_val, X_test
11) Split the training dataset:
    - Split X_data and y_data → X_train, X_val, y_train, y_val with test size of 20%

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

1) Below columns were removed from the dataset:

- Column 'type': There is only one value in this column, which is 'all'. Hence there is no need to keep this column.
- Column 'player_id': This column does not play any role in the prediction task.
- Columns 'Rec_Rank', 'dunks_ratio', 'pick': These columns contain > 50% of the total values

2) Column ‘ht’: Realising this feature is vital for the prediction task, I decided to convert the data (displayed in date format) to float data type.
3) Apply LabelEncoder technique to categorical columns, including ‘team’, ‘conf’, ‘yr’
4) Feature elimination with correlation matrix to select the most relevant features of the dataset → From 62 features in the original training dataset, after performing the feature selection, there are only 43 features left.

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 RandomForest model to train the dataset. The recall scores on the training and validation datasets when trained with default hyperparameters were 1.0 and 0.9998734307453669 respectively.

To reduce the overfitting issue, I conducted manual hyperparameters tuning on n_estimators, max_depth, min_samples_leaf and max_features to find the best hyperparameters.

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.

1) **Baseline model:**

   AUROC score on training dataset is 0.5

2) **RandomForest Model without Default Hyperparameters:**

   AUROC score on training dataset is 1.0

   AUROC score on validation dataset is 0.9998734307453669

3) **RandomForest Model with Hyperparameters Tuning n_estimators=90:**

   AUROC score on training dataset is 1.0

   AUROC score on validation dataset is 0.999869927420505

4) **RandomForest Model with Hyperparameters Tuning n_estimators=90, max_depth=30:**

   AUROC score on training dataset is 0.9999965275615045

   AUROC score on validation dataset is 0.9996575246817134

5) **RandomForest Model with Hyperparameters Tuning n_estimators=90, max_depth=18:**

   AUROC score on training dataset is 0.9994360241472875

   AUROC score on validation dataset is 0.9987519901315267

6) **RandomForest Model with Hyperparameters Tuning n_estimators=90, max_depth=18, min_samples_leaf=2:**

   AUROC score on training dataset is 0.9995628033374554

   AUROC score on validation dataset is 0.9988653520539349

7) **RandomForest Model with Hyperparameters Tuning n_estimators=90, max_depth=18, min_samples_leaf=6, max_features=12:**

   AUROC score on training dataset is 0.9991857853148767

	AUROC score on validation dataset is 0.9983015638064566
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) RandomForest model by far was the most effective model in generating accurate predictions with high AUROC scores (more than 0.999) with minimal overfitting compared to the previous Logistic Regression and Adaboost models.
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. 1) **Column 'ht'** was a challenge when I tried to convert it to numeric values. At first, I thought the data was datetime data type and spent a lot of time to work out of this assumption. However, after checking the data type, it was actually string. Then I approached it by splitting the string by the hyphen. 2) With this Experiment, I tried to **apply customised functions** to improve the code execution robustness. However, in this Experiment, some functions were not able to be employed, including those to save and load datasets, split dataset, assess baseline and model performance. The reason is because of the file structure.

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 this experiment with RandomForest model, the performance was the best amongst the three experiments so far, although there is still small overfitting.
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 steps: ▪ Learn how to apply pre-populated functions to the project, instead of copy and paste repetitive codes to different work books.