EXPERIMENT REPORT

Student Name	Simon Lim
Project Name	Models Predicting Potential NBA Basketball Player
Date	25/08/2023
Deliverables	Lim_Simon-24661225-week2_LGBMclassifier.ipynb LGBMclassifier https://github.com/SimonUTS24661225/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	The goal of the project is to build potential models that correctly predict if a college basketball player will be drafted to join the NBA league in the future based on player's various statistics for the current session. The project also aims at the ground truth of players who can be drafted into NBA league.
1.b. Hypothesis	1.It is hypothesized that LightBGM model will be likely be a potential model for this project, as the project involves binary classification. LightBGM is fast and accurate and particularly well suited for large datasets. This dataset has over 50000 data on the training dataset. The purpose of the project is mainly to distinguish whether players can be drafted to NBA league or not. Therefore, LightBGM would be an optimal model and will result in a high AUROC score. 2.Also, AUROC score will be likely to be a suitable performance metric in this project, in order to distinguish random players between positive and negative classes.

1.c. Experiment Objective	The expected outcome and scenarios of the experiment will be that LightBGM models will produce probabilities of predictions and for training dataset AUROC scores will be close to 1 (i.e., very high performance in distinguishing positive and negative classes). For testing dataset, probabilities of predictions will be submitted to Kaggle and will obtain ground truth of players who can be drafted into NBA league. Kaggle will provide AUROC scores for testing dataset. It is also expected AUROC scores for testing dataset will also be very high, close to 1, giving us the ground truth of the project. The score, close to 1, indicates that the model can correctly predict players who can be drafted to NBA league.

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

1.There were a lot of missing values in a couple of columns, almost 24 columns. There were three columns (e.g., ‘pick’, ‘dunks_ratio’, ‘Rec_Rank’), in which there were more than 50% of missing values in both training and testing datasets. I decided to delete these three columns rather than dropping missing values, in order to enhance the ground truth of the project (i.e., including all rows of data). The other columns were filled with means of the columns.

2. There were three object columns (‘ht’, ‘num’ and ‘yr’), in which string values were replaced with numeric values and data types were changed to integer or float.

3. In training dataset, numeric data were split into X_train, containing all numeric columns except ‘drafted’, target variable and y_train containing ‘drafted’ column. For testing dataset, X_test was only created, containing all numeric columns and there was no y_test, as ‘drafted’ column was not given in testing dataset. (In this project, instead of obtaining AUROC score for testing dataset through target variable, AUROC score for testing dataset will be obtained by submitting probabilities of predictions for data on Kaggle).

4. X_train, X_test and y_train were then saved as csv files and will be used in future experiments.

5.‘StandardScaler’ was used to remove outliers in data.

2.b. Feature Engineering

After removing columns and filling missing values with means of values in some columns, there were three object columns, **‘ht’** (Height of student), ‘**num’** (Player’s number) and **‘yr’** (Student’s year of study). ‘The values in the column **‘ht’** were ambiguous and difficult to replace with numeric values. Hence, I decided not to include it on building models. On the other hand, ‘**num’** and **‘yr’** were string values but could be replaced with numeric values. For example, there were only 5 unique values in **‘yr’** **(**freshmen, sophomores… etc.**.),** which can be replaced with 0 to 5 in order. Although there were some typo errors in ‘**num**’ column, such as ‘23B’ or ‘31/24’, values were successfully replaced with numeric values.

Lastly, data type of ‘yr’ was changed integer and data type of ‘num’ was changed float.

2.c. Modelling

There were two models of LightBGM used in this experiment. One was LBGMclassfier with the default model.

The second model was logistic LBGMclassifier with best hyperparameters. The best hyperparameters were tuned and obtained using RandomSearchCV.

In this experiment, LBGMclassifier with best hyperparameters showed a better performance than default LBGMclassifier on testing dataset. In future experiments, assessing different hyperparameter would be important steps to enhance the performance of models.

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	For performance metric, AUROC score was used as a main metric. The purpose of the project is to build models that can correctly distinguish random players based on statistics of players between positive and negative classes (whether drafted or not into NBA league). AUROC score measures performance of model's probability of predictions whether players belong to positive or negative classes. Therefore, AUROC score will be a suitable metric for this project.
3.b. Business Impact	LBGMclassifier models also showed a great performance in distinguishing random players between drafted and non-drafted classes. **AUROC score of default LBGM model, which was obtained on Kaggle, was 0.98366 and AUROC score of best hyperparameters LBGM model was 0.98392. LBGM classifier models showed slightly better performance than logistic regression models and therefore will be successfully able to distinguish players who can be drafted to NBA league.**
3.c. Encountered Issues	1. There were a lot of missing values in datasets. The missing values could not be dropped because of keeping the ground truth of testing dataset. Therefore, I tried to fill out missing values as long as it is possible, otherwise I had to remove the columns if they had too many missing values (solved). 2. In feature engineering part, values in **'ht'** column were ambiguous and difficult to interpret. Because the column indicates the height of players but values in the column were dates, such as '9-Jun', '3-Jul' (Unsolved).

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	1. Enhancing the ground truth of machine learning models is very important, especially when applying to real-world cases. It will increase accuracy against the real-world. 2. Filling missing values with mean of values in column is also crucial in improving the performance of models (In case the proportions of missing values are small). 3. Splitting into numerical and categorical columns is very efficient in building models and combining columns in future process. 4. Feature engineering is a crucial process in machine learning, which can potentially increase performance of models. 5. It is also crucial to find the optimal hyperparameters of model for better results.

4.b. Suggestions / Recommendations	1.Deciding main performance metric in advance is very helpful in building optimal models. 2. Rather than removing missing values, filling out missing values can potentially enhance the performance of models (in case the proportion of models is not large). 3. Feature engineering is an essential stage in machine learning, that can influence models and outcomes of the projects. 4. Optimal hyperparameters can be found using techniques, such as RandomSearchCV or GridSearchCV.