EXPERIMENT REPORT

Student Name	Simon Lim
Project Name	Models Predicting Potential NBA Basketball Player
Date	25/08/2023
Deliverables	Lim_Simon-24661225-week1_Logistic.ipynb Logistic Regression 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 logistic regression is likely to be a suitable model for this project, as the target variable is binary. That is, the purpose of the project is mainly distinguishing whether players can be drafted to NBA league or not. Therefore, logistic regression would be a powerful model. 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 logistic regression 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.

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.Data were already split into training.csv and testing.csv. Hence, after importing and briefly describing two datasets, the next step was to investigate any missing values.

2.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.

3. Next step was to split into numeric columns and categorical columns in two datasets. Numeric columns would mainly be used in training and testing models and only 'player_id' would be used in categorical columns for future Kaggle submission.

4. EDA was performed with histogram and boxplot to check distributions and outliers of data. There were many outliers in each column, and therefore decided to normalize data.

5. 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).

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

7. I tested two normalization techniques to reduce outliers of data, 'StandardScaler' and 'MinimaxScaler'. 'StandardScaler' showed greater scores than 'MinimaxScaler'. Therefore, 'StandardScaler' would be used in future experiments.

8. Default logistic regression and logistic regression with regularization were trained, probability of predictions was obtained and AUROC score was provided for training dataset (Results mentioned in next sections).

9. Two models of logistic regressions were visualized using plots with different colors.

10. For testing dataset, probabilities predictions were obtained through training model, and probabilities of being drafted predictions were than combined as a dataframe with 'player_id' in categorical column and then sent to Kaggle as a csv file.

2.b. Feature Engineering	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). Fortunately, the other columns only had a small proportion of missing values, which were filled with means of the columns.
2.c. Modelling	There were two models of logistic regression used in this experiment. One was logistic regression with the default model. The second model was logistic regression with regularizations (penalty= ‘elasticnet’, l1_ratio=1, solver=’saga’). When assessing two models in training dataset, default logistic regression model showed a better AUROC score than logistic regression with regularization. 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 between positive and negative classes (whether drafted or not into NBA league). AUROC score will be a suitable metric for this project.
3.b. Business Impact	Logistic regression models showed a great performance in distinguish random players between drafted and non-drafted classes. **AUROC score of probability prediction of being drafted, which was obtained on Kaggle, was 0.98243, which is a very high score. High AUROC score indicates that logistic regression model is one of suitable models that can correctly distinguish random players between drafted or non-drafted classes.**
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).

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. Outliers' detection and removal through the two different normalization techniques (MinMaxScaler and StandardScaler) 4. Splitting into numerical and categorical columns is very efficient in building models and combining columns in future process.

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).