EXPERIMENT REPORT

Student Name	Trung Kien Hoang
Project Name	Kaggle Competition: The NBA draft
Date	25/08/2023
Deliverables	<Experiment 2> <Logistic Regression Classifier, Polynomial Logistic Classifier, KNN model using Euclidean distance, SVC with default hyperparameters, Decision tree with default hyperparameters and AdaBoost Classifier> <Data preparation: Label encoding(yr) + Converter(ht)> Github link: https://github.com/KenUTS/adv_mla_assignment_1.git

---

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 main objective of this project for the business is to build a prognostic model capable of evaluating the probability of a collegiate basketball player getting selected for the National Basketball Association (NBA) draft, using their performance information from the current season. The outcomes of the model will be used by diverse stakeholders, including NBA teams, players, sport commentators and fans. Suppose the model accurately forecasts which player will be chosen. In that case, NBA teams can use the model's predictions to help them decide how to pick players in the draft. By finding possible high-performing players, they could improve their team performances. However, if the model's expectations are wrong, teams might not pick the best players with skills in the draft. The accuracy model has the potential to assist players in making informed decisions between NBA draft and their college careers. However, selecting an incorrect model has the potential to lead players to make misguided decisions in their future careers. For sport commentators and fans, the model engages in more comprehensive discussions pertaining to players. However, this method has the potential to generate inaccurate feedback and therefore disappoint supporters in the event that their favorite player is not selected.

1.b. Hypothesis	Hypothesis: The statistical data of collegiate basketball players during their current season have the potential to enhance the likelihood of their selection in the National Basketball Association (NBA) draft. Question: Are there any elements such as height of players or year of study that contribute to an increased likelihood of being selected in the NBA draft? It is worth examining due to its potential to assist players in their preparation for the selection process.
1.c. Experiment Objective	The projected result of the undertaking is a model that effectively forecasts the likelihood of players being selected in the NBA draft. Based on the requirement of the project, the metric used to assess model performance is AUROC (Area Under ROC). It is supposed to be smaller than 1 and greater than 0.8. Scenario 1: Train the selection of algorithms; for all data sets, if the best model's performance is less than 0.8 of the AUROC score, take one or more actions: Change data preparation, feature engineering, hyperparameter optimization, or algorithm selection. Scenario 2: Train the selection of algorithms; for all data sets, if the performance of a model is greater than 0.8, finish the project and deploy the model to the business.

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. Adding a new column for shots missed at or near the rim, missed two point shots that were not made at or near the rim and missed dunks for training and testing data. 2. Drop player number as it is not meaningful information, columns with large numbers of missing values such as Rec_Rank and pick columns. 3. Drop ratio columns where it is insufficient due to the players may have drunk made and missing equal to 0 for instance. 4. Fill all missing values of numerical columns with median values.
2.b. Feature Engineering	Because, the shots made and missed could be zero, so the ratio between the shots made and shot missed could be undefined if the number of shots missed is zero. Therefore, for this experiment, a new column for shots missed at or near the rim, missed two point shots that were not made at or near the rim and missed dunks are added for training and testing data. Furthermore, Polynomial Transformation is used to capture the complex among selected numerical features. Label encoding(yr) + Converter(ht): 1. According to many sources, it has been suggested that the classification of students based on their year of study follows a certain pattern. Freshmen are often referred to as first-year students, sophomores as second-year students, juniors as third-year students, and seniors as fourth-year students, indicating a hierarchical ranking based on their progression through their academic programme. Hence, the use of a label encoding tool is considered the most effective approach for converting categorical data into numeric variables. 2. It seems that there is an issue in the height column, where either Kaggle or Excel has mistakenly converted players' heights into dates. To illustrate, let us use the measurement of 6 feet 11 inches, denoted as 6 - 11 while this measurement was recorded into a date format, namely 11/06/2023. Consequently, a function is established with the purpose of converting such records.

2.c. Modelling	In this experiment, there are six models: Logistic Regression Classifier, Polynomial Logistic Classifier, KNN model using Euclidean distance, SVC with default hyperparameters, Decision tree with default hyperparameters and AdaBoost Classifier are trained in training, validating, and testing dataset. Specifically, Logistic Regression and KNN are standard models in binary classification; SVM is used to find the best boundary to separate both classes; and a decision tree is a simple algorithm for modeling non-linear relationships. The AdaBoost model combines multiple weak learners to create a strong classifier. 'n_estimators' hyper parameter is trained in the AdaBoost model. Other hyperparameters will be trained for future experiments.

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	Logistic Regression Classifier: the AUROC score on training is 0.982 and validating sets is 0.984. Polynomial Logistic Classifier: the AUROC score on training is 0.982 and validating sets is 0.983. KNN model using Euclidean distance:the AUROC score on training is 0.993 and validating sets is 0.747. SVC with default hyperparameters: the AUROC score on training is 0.982 and validating sets is 0.980. Decision tree with default hyperparameters : the AUROC score on training is 1 and validating sets is 0.660. AdaBoost Classifier: the AUROC score on training is 0.995 and validating sets is 0.988. It can be seen that AdaBoost Classifier is the best model with highest accuracy predictions but there is slightly overfitting between training and validating the dataset.
3.b. Business Impact	The elevated AUROC score signifies the model's proficiency in properly discerning collegiate players who possess a high likelihood of being selected in the draft. NBA teams may use this knowledge to enhance their draft plans, players can develop a strategic approach towards their careers, and fans experience heightened enthusiasm for their favorite players. The consequences of inaccurate outcomes, such as an NBA team, may include a decline in team performance and a loss of competitive edge. A player with a lower anticipated probability may risk losing a chance to be picked. A subpar model might also diminish the level of anticipation among fans.
3.c. Encountered Issues	Because there are missing values on ht(height) column, it raised an error: “TypeError: 'float' object is not subscriptable” when i tried extract month and date to feet and inch measurements. It can be solved by “isinstance” and “isdigit” functions that can identify float and int objects.

---

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	The chosen algorithm for this experiment is deemed sufficiently effective in completing the project. However, it is crucial to acknowledge the presence of overfitting across the training and testing datasets.

4.b. Suggestions / Recommendations	Experiment 3: Use the suitable tools for fixing imbalance dataset, Perform some features selection techniques, and also tune hyperparameters for other models. Rank 1: Smote + Stratify Rank 2: Feature selection Rank 3: Train models with hyperparameters if there is overfitting