EXPERIMENT REPORT

Student Name	Rushab Shah
Project Name	Kaggle Competition
Date	17/08/2023
Deliverables	<NBA Draft> <model name> <other>

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 objective of this project is to build a model that will accurately predict if a college basketball player will be drafted to join the NBA league based on his statistics for the current 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, ● Hypothesis: College basketball players' statistical performance during the current season can effectively predict their likelihood of being drafted into the NBA. ● Question: Can player statistics in college accurately forecast their NBA draft selection? ● Insight: Analyzing player data to determine if metrics like points, rebounds, assists, field goal percentage, and more can reliably indicate their potential for success in the NBA. This hypothesis is worth exploring as it could provide a data-driven approach to enhance NBA draft decision-making, identifying players who demonstrate the skills and consistency needed to excel at the professional level.
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. Initial experiment to determine the best algorithm and features to predict if a player will be drafted.

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.

Ht column was corrected by converting the dates into corresponding height values in inches (Month assumed to represent height in feet)

Imputations to fill numerical missing values with median value. High number of numerical missing values so the median was taken in their place to determine if the features were still relevant. Mode and mean will also be tested.

SMOTE sampling technique to balance the target ‘drafted’ class as there is a major imbalance, with only approximately 1% of ‘drafted’ players.

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 additional features were created for this experiment, however several features were removed. Features with greater than 50% missing data and features that did not add any value were removed.

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

Pycaret was used to automatically train several different classifiers to determine the best one. XGboost returned the highest overall metrics but was slow to train, while light GBM was faster to train and returned a high AUC and F1 score, therefore both models were trained, but only light GBM could be tuned due to slow training time for XGboost.

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.
	Accuracy = 0.9915 AUC = 0.9826
	Recall = 0.4481 Precision = 0.5749 F1 Score = 0.5009

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) The experiment had a low number of correctly predicted drafted players. There was a larger number of drafted players that were not predicted, highlighting room for improvement. In this case, a type II error (false negatives) is better than type I error.
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. Data preparation issues: Solutions listed in data preparation issues.

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. Fixing the height values and imputing missing values proved to be an effective method to improve the AUC. Feature selection also showed that not all the features contribute to the AUC, indicating that there may be further potential to optimize engineered and selected features.
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. The next step would be to build a better/improved model and manually hypertune it. The generation of more features will also be tested, to achieve the same objective. Similarly, an optimal selection of features needs to be identified for the best AUC.