EXPERIMENT REPORT

Student Name	Rushab Shah
Project Name	Kaggle Competition
Date	01/09/2023
Deliverables	<NBA Draft 3> <wk3_lr_pipeline> <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. Experimentation with features to determine the possibility of improvements to the AUC.

---

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. Filtering on 'pick' (undersampling technique) Ht column was corrected by converting the dates into corresponding height values in inches (Month assumed to represent height in feet)
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 Several new features were created for this experiment that aimed to aggregate other features. Metrics related to drafting percentage of teams, player efficiency, attacking, defensive and others were created and incorporated into the model. For categorical values, imputations to fill with mode and one hot encoding. For numerical values, imputations to fill with mean and scaling was performed
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 Sklearn was used for this experiment allowing for more customizability with the model. Hyperparameter tuning was not the focus of this experiment (more focused on feature engineering), however several different solvers were tested to determine any improvements to the classification model, however they all returned same confusion matrix.

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 = 0.8273 AUC = 0.8651 Recall = 0.7037 Precision = 0.8260
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 did not show much improvement, however when tested against the Kaggle data, it returned exactly the same score suggesting that the model the new features did not result in any improvement, or perhaps there maybe an error in the experiment.
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. Developing the pipeline for Sklearn. Including imputations with the feature engineering in the pipeline proved to be challenging and may not have been done correctly, since several columns required adjustment due to division by zero.

---

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. If the results are free from error, this implies that the engineered features were not useful to the model. It may be worth checking this against other algorithms before ruling these features out.
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 experiment conducted needs to be studied to determine why errors from division by zero were occurring despite imputations performed prior. Other algorithms should also be used to model to see if there are any improvements. Otherwise, more features will be generated and attempts at tuning hyperparameters.
