EXPERIMENT REPORT

Student Name	Anika Chauhan
Project Name	36120 Advanced Machine Learning Application
Date	01/09/2023
Deliverables	chauhan_anika-14188775-week3_XGBoost_features_imputed https://github.com/anikachauhan30/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 construct a prediction model to determine the probability of a college basketball player being selected to join the NBA league based on their performance statistics from the current season. The results can be used by NBA teams to identify and recruit talented players, by players to see their standing and make necessary improvements, and by college scouts to attract potential recruiters towards talented players. Accurate results will lead to optimized player selection, players making informed choices and heightened fan engagement. On the other hand, incorrect predictions might cause some talented players to miss valuable opportunities, and the teams to misallocate their resources. The model's accuracy is crucial and pivotal as it holds the potential to make or break careers.
1.b. Hypothesis	The hypothesis being tested is whether the college basketball players' current season performance statistics can effectively predict their chances of being drafted to the NBA. It is worthwhile to consider it as an accurate prediction can help players and agents make informed decisions and help the NBA teams to optimize their drafting process.

1.c. Experiment Objective	The expected outcome of this experiment is to develop a predictive model which can accurately predict the probability of a college basketball player being drafted to the NBA based on the current season's performance statistics. The goal is to achieve a high AUROC (Area Under Receiver Operating Characteristic) score which indicates a strong predictive performance.

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	The following steps were taken to prepare the data for this experiment: 1) The 'ht' feature was being interpreted as an object datatype instead of float. To correct this a function 'feet_inches_to_cm' was developed which splits the height string into feet and inches, handles various formats ('5-Jun' or 'Jun-00'), and converts them to centimeters. 2) The numerical features with the absolute correlation coefficient of 0.22 or higher with the target variable ('drafted') were stored in a list 'rc'. These features were: 'twoPM', 'porpag', 'dunksmade', 'dunksmiss_dunksmade', and 'dporpag'. Out of these only twoPM, porpag and dunksmade were used in training the model. Dporpag and dunksmiss_dunksmade were dropped as they were derived from porpag and dunksmade respectively. 3) Simple imputer was used to fill in any gaps in these features with the mean of that particular column. 4) StandardScaler was used to help prevent the high (or low) magnitude of some features from overpowering the results.
2.b. Feature Engineering	The 'ht' feature was being interpreted as an object datatype instead of float. To correct this a function 'feet_inches_to_cm' was developed which splits the height string into feet and inches, handles various formats ('5-Jun' or 'Jun-00'), and converts them to centimeters. In the future experiments, the focus will be on refining the imputed values and experimenting with some other features.
2.c. Modelling	For this experiment XGBoost classifier was used as it is known to be able to handle complex relationships in data and also due to its robustness against overfitting. It was chosen for its suitability for binary classification tasks like predicting whether a player will be drafted (yes or no). Random Forest was not used as it did not do well with missing data and logistic regression was treated as the baseline. In the future experiments, I intend on employing feature selection techniques along with gradient boosting algorithms to uncover better performance under varied circumstances

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	This experiment yielded an AUROC score of 0.9741. This score is lower than the best score obtained till date. But as this only the score for 50% of the data, it might just be a step in the correct direction. While it looks like an almost perfect score, such high value can indicate overfitting. To overcome that, feature selection and hyperparameter tuning will be applied in future experiments.
3.b. Business Impact	Such a high AUROC score indicates that the model's ability to discriminate between drafted and non-drafted players is exceptionally high. However, incorrect results can lead to missed opportunities for talented players, misguided draft choices and a lot of wasted resources.
3.c. Encountered Issues	1) The 'ht' feature had a formatting issue which was later fixed by developing a function which first converts strings to float and then changes inches to cm. 2) Null values posed a problem when it came to evaluation. To fix this simple imputer was used. 3) This experiment did not explore any feature engineering techniques. This will be explored in future experiments. 4) XGBoost's black-box nature might have skewed the results and hindered the interpretability. This will be resolved by tuning the hyperparameters and further using tools like LIME to gain insights into the model's decision-making process.

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	Although the high AUROC score gives the experiment a lift, the lack of feature engineering suggests room for further exploration. This is in no way a dead end but there is a still a long way to go for further refinement.
4.b. Suggestions / Recommendations	The next steps involve refining the model using feature engineering, hyperparameter tuning and introducing tools like LIME. Upon achieving desired performance, the model can be deployed to be integrated in the NBA teams' selection procedures as an aid to their existing recruiting methods.