EXPERIMENT REPORT

Student Name	Rohan Rocky Britto
Project Name	Assignment 1 Week 1
Date	18/08/2023
Deliverables	Britto_Rohan-24610990-week1_random_forest.ipynb Random Forest Algorithm Github Link: https://github.com/rohanbrit/Adv_ml_asgn1

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 NBA draft is an annual event in which teams select players from their American colleges as well as international professional leagues to join their rosters. The task is to build a model that will predict if a college basketball player will be drafted to join the NBA league based on his statistics for the current season. An incorrect prediction can impact the reputation of the organization that uses them.
1.b. Hypothesis	I will be building a Random Forest model to predict the possibility of a player being drafted into the NBA league. Random Forest is an ensemble technique that combines the power of multiple decision trees to make a prediction. It could be a good starting point to know if there is a good relation between the features and the target variable.
1.c. Experiment Objective	As Random Forest is a very powerful model, there is a high chance that it will be able to make good predictions. However, this model also tends to overfit, and I might have to be watchful of that.

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. There were a couple of features like 'Rec_Rank', 'dunks_ratio', 'pick' that had a lot of null values in them. I had to drop them as filling them up could lead to deviation from the real-world data. 2. type feature had only 1 unique value and would not help the model in making predictions 3. num and player_id are identifiers and can lead to overfitting 4. ht feature did not have good data and hence, I have dropped it 5. Abnormalities in yr feature were adjusted with the mode value 6. Null values in numeric features were replaced with mean value 7. Categorical data was converted into numerical values using frequency encoding as they had a lot of distinct values 8. All the numerical data was scaled 9. There was a class imbalance noticed in the dataset and hence, I had to use SMOTE oversampling technique to balance it
2.b. Feature Engineering	The features like 'Rec_Rank', 'dunks_ratio', 'pick' could be important from a future perspective, and we should check if we can get the actual values for these features from the business.
2.c. Modelling	Random Forest is an ensemble technique that combines the power of multiple decision trees to make a prediction. As expected, it did a very good job in predicting the target variable. There was slight overfitting noticed and I tuned the max_depth hyperparameter to overcome it.

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	I used AUROC as the evaluation metric for this experiment as specified in the task description. Random Forest achieved a perfect score of 1 for the training set and 0.9998 for the validation set. After adjusting the hyperparameter, it was able to achieve 0.9916 for the training set and 0.9909 for the validation set. As per the public leaderboard, I achieved a score of 0.97051 for 50% of the test set which indicates that the model is still slightly overfitting.
3.b. Business Impact	Though the results achieved are very high, there is still some overfitting noticed. I will have to further fine tune the model and try other models to make sure the model performs consistently on unseen data. However, a high score surely indicates that there is a good correlation between the features and the target variable.
3.c. Encountered Issues	The presence of null values in potentially important features, imbalanced dataset and overfitting of the model were the major issues encountered in the experiment. If we are unable to achieve good results after hyperparameter tuning and trying other algorithms, we might have to go back to business and check if we can populate the null values with actual data and find some more data to balance out the dataset.

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	We have been able to achieve a very high score which indicates a good correlation between the features and the target variable. Further experimentation will be required to reduce overfitting of the model and to try other algorithms.
4.b. Suggestions / Recommendations	We will be experimenting further to try and reduce the overfitting in Random Forest and will also try other algorithms.