EXPERIMENT REPORT

Student Name	Rohan Rocky Britto
Project Name	Assignment 1 Week 2
Date	25/08/2023
Deliverables	Britto_Rohan-24610990-week2_multiple_algorithms.ipynb Random Forest, Logistic Regression, SVC and Adaboost 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 multiple algorithms like Random Forest, Logistic Regression, Support Vector Classifier, AdaBoost to predict the possibility of a player being drafted into the NBA league. I will compare the performance of all these algorithms and use the best 2 for my predictions this week. I will also use these 2 algorithms in my experiment next week and try to improve their performance using hyperparameter tuning.
1.c. Experiment Objective	This experiment will give us a good insight on which algorithms work well and which do not with the given dataset.

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	In the previous experiment, I had dropped ht feature on the basis that it had bad data. But as per the discussions in Canvas, I got to know that this was an issue with Kaggle and/or Pandas. Hence, I had to redo the data preparation steps. Looking at the unique values in the ht feature, I could gather that the month represented the feet value and day represented the inch value. I used frequency encoding on ht feature as well to transform 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	AdaBoost, Random Forest and SVC were the top three best performing models. However, the AUROC score for Random Forest did not change much compared to the previous experiment. AdaBoost returned an AUROC value of 0.9838 on the test set on Kaggle which indicates some overfitting. SVC returned an AUROC value of 0.9709

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 score of 0.9979 for the training set and 0.9960 for the validation set after adjusting the hyperparameter. Logistic Regression achieved a score of 0.9905 for the training set and 0.9908 for the validation set. SVC achieved a score of 0.9958 for the training set and 0.9953 for the validation set. AdaBoost achieved a score of 0.9965 for the training set and 0.9960 for the validation set. As per the public leaderboard, I achieved a score of 0.98387 and 0.97093 for 50% of the test set for AdaBoost and SVC respectively.
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 to make sure the models perform consistently on unseen data.
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	AdaBoost and SVC have been able to achieve a very high score, especially AdaBoost. Further experimentation will be required to increase the performance and reduce overfitting.
4.b. Suggestions / Recommendations	We will be experimenting further to try and increase the performance of the models and to reduce overfitting.