EXPERIMENT REPORT

Student Name	Brandon Ji
Project Name	Experiment 1
Date	18-06-23
Deliverables	Ji_brandon_24579183-week1.py Decision tree Github link: https://github.com/bswji/Advanced-machine-learning

---

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 where NBA teams have an opportunity to add a valuable player to their rosters. By being able to predict which college players are drafted, teams can figure out the which metrics are most important when valuing players. If the prediction model were incorrect and teams select a lesser player, then the results would be quite detrimental as it could possibly delay a year of progress for the team.
1.b. Hypothesis	I think that the most valuable insight that can be gained from this experiment would be to be able to identify the key metrics that affect a player's draft status. I hypothesise that the most important metrics affecting the draft status of a player will be offensive rating, true shooting percentage and defensive rating. This is because these three metrics combine multiple other metrics together which I believe provide a more holistic representation of a player's performance.
1.c. Experiment Objective	I am expecting to create a somewhat accurate model that is able to predict the draft status of a player. With the model, the most important metrics determining draft status can also be identified.

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	There were many missing values within the data. Variables with more than 100 missing values were removed as they could cause incorrect results if left in the final model. I decided not to use the mean or median values to replace the missing values as I believe that this could create unrealistic values. I also wanted to see how the model would perform with these variables removed. Player id, num and type were removed because they are not relevant to the current problem. All categorical variables were also removed for various reasons. Ht was removed as it did not seem to have any relationship with the listed description in the metadata file. Team and conf were removed as there were too many different unique values which I believe would have little correlation with being drafted. In the future these variables could be changed into numerical values to be used in the data.
2.b. Feature Engineering	No new features were generated.
2.c. Modelling	I chose to use a decision tree and random forest. I chose to use both these models as they are robust to outliers and run well on large datasets such as this one. The model also does not require the variables to be scaled, making it ideal for an introductory experiment. The hyperparameters that were tuned for the random forest were the number of estimators, maximum tree depth and maximum number of features. The number of estimators was chosen as it represents the number of trees in the model. A higher number of trees in a model generally yields with better results, however due to the long training time that can occur. Thus, finding the number of trees that can balance both of these aspects is beneficial to the development of the model. This was also true for both max depth and max features. As for the decision tree, the minimum number of samples to split an internal node, maximum number of features and maximum depth were chosen to be tuned. These were all chosen to be tuned as they are important in preventing over fitting. In the future, other hyper parameters such as criterion may be considered as currently the models have underperformed. Other models such as KNN and Logistic regression should also be considered.

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	The ROC score for the decision tree was 0.60 whilst the score for the random forest was 0.58. Both of these are poor results likely caused by the processing of the data (removal of columns) and imbalanced dataset. The selection of the models used may have also had an impact on the performance of the model.
3.b. Business Impact	If these models were to be used as the final model for teams to use to determine the value of a player, then the results would be disastrous. This is because the current models are extremely inaccurate and would advise the drafting of less valuable players.
3.c. Encountered Issues	There were many missing values which I thought the best way to deal with them would be to remove them. However, with the poor results, I may need to come up with a new approach. Categorical variables were also removed however, some variables may be useful.

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	From this experiment I found that the results were poor and would likely need an entirely new approach to achieve better results. I learnt that the imbalanced data, missing values and categorical features should all be dealt with differently in order to achieve better results.

4.b. Suggestions / Recommendations	1- Deal with imbalanced data differently – (Would probably require upsampling/downsampling which would result in the creation/deletion of a lot of data) 2- Deal with missing values differently – (There was an enormous amount of missing values in the dataset. By deleting features containing many missing values, a lot of data was lost. By changing my approach, this data could be saved and used for the model.) 3- Use different models – (Current models did not work very well. Different models would solve the problem differently which would lead to new results.) 4- Deal with categorical features differently – (I think that the categorical features have low value in this problem, however they should still be considered as the current results are poor)