EXPERIMENT REPORT

Student Name	Daniel Alexander
Project Name	Assignment_1_Part_C
Date	01/09/2023
Deliverables	Experiment_C.ipynb Adaptive Booster ROC Score Confusion Matrix Github Repository: https://github.com/daniel-alexandr/adv_ml_application_assignment_1

---

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	This project is aimed to help the company as one of the leaders in the sporting apparel industry to maximize potential investment return on young players. The result of this project will be used for determining potential sports models to be invested early and cheap before the young players are even drafted for the NBA. Inaccurate results will impact the scale of financial loss, for example offering contracts to players that is not drafted will not generate profit for the business.
1.b. Hypothesis	We can get a better performing model than the random forest model from our experiment B. Model B will be our benchmark, if we can't get a better performing model, then we will use the model from experiment B for production, otherwise we will consider the model from this experiment.

1.c. Experiment Objective	The objective of this experiment is to build a reliable predictive model that utilizes the statistics of college basketball players' current season performance to assess their likelihood of being selected in the NBA draft. If the experiment proves our hypothesis is correct, then the model should be pushed into production which then will be used as part of business as usual to help the player acquisition department to invest in the right people. If the model used in this experiment does not satisfy, then we would either pursue more statistics (features) of the players, or try other potential models. If the results are way off, then we might also consider terminating the project.

---

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. Removing the missing observations from these following columns. From the EDA, we suspect those features play a significant role in determining the likelihood of being drafted, while the number of missing observations are insignificantly small. Therefore, we keep these features and remove missing observations.

   - Rimmade
   - Drtg
   - Mp

2. Performing resampling, specifically under sampling to mitigate the imbalance dataset. The minority class represented are less than 2% of the raw dataset. We resample the dataset with a 70:30 ratio, hoping that by capturing the scarcity of the minority will produce a better performing model.

3. Since the test data are separated from the training data, it requires different treatments for the missing values. We do not want to delete any of the missing observations on the test data.

4. In this experiment, instead of splitting training and validation, we are using cross validation method with 5 fold cross validation.

Potential approaches to be used in future experiments are the resampling techniques, either changing the ratio or the resampling methodologies such as SMOTE, which might increase the performance of our model.

2.b. Feature Engineering

1. Removing these features as our EDA indicates that it is insignificant towards our objective, also to avoid overfitting caused by unique features such as ID

   - Team
   - Yr
   - Type
   - Num
   - year
   - player_id
   - ht (Values are absurd, need clarification)
   - pick (Will make the model only consider those who were given the draft order)

2. Fill in missing observations with 0 for these columns because it's appropriate or to capture the missingness or inapplicableness.

   - dunks_ratio || *caused by no successful dunks were made*
   - rim_ratio || *caused by no succesful rimmshot were made*
   - mid_ratio || *caused by no midmade_midmiss were made*
   - Rec_Rank || *to capture the state of unranked*

3. Fill in the missing value with its average, because these features are not supposed to be missing, significant for our purpose, and is appropriate to be filled with its average.

   - ast_tov

4. Scaling the numerical features using standard scaler arguably to avoid scale sensitivity issues for some algorithms.

5. Applying one hot encoding for conf feature as the only categorical feature that we are using in this experiment.

Below are the data engineering treatments which we applied to the test data so it is compatible with the trained model:

1. Remove these following columns to match what were removed from the training data:

   - Team
   - Yr
   - Type
   - Num
   - year
   - player_id
   - ht (Values are absurd, need clarification)
   - pick (Will make the model only consider those who were given the draft order)

	2. Since we do not want to remove any observations from our test set, therefore, we fill the missing observations for these following columns with their average value, because there are only 1 missing observation for each of these columns and filling them with the average value is insignificant. - Drtg - Adrtg - Dporpag - Stops - Bpm - Obpm - Dbpm - Gbpm - Ogbpm - Dgbpm - ast_tov 3. Filling missing observations for these columns with 0 because it is logical to use that value considering that the dependent columns (ratio's relating to these columns) support our logic as no shots were made. - Rimmade - Rimmade_rimmiss - Midmade - Midmade_midmiss - Dunksmade - Dunksmiss_dunksmade - Rim_ratio - Dunks_ratio - Mid_ratio - Rec_Rank Potential approaches for future experiments, process feature ht as logically it's a significant factor for our purpose.
2.c. Modelling	In this experiment we are using an Adaptive Booster Classifier with hyperparameter: random_state=42, n_estimators= 468, learning_rate=1, base_estimator=DecisionTreeClassifier (max_depth=7), algorithm='SAMME' The model selection used in this experiment is focused on improving the performance from the previous experiment, we are aiming at >98 roc_auc score for this experiment. We chose this algorithm because usually adaptive boost algorithm fits better compared to random forest algorithm For our future experiment we might try polynomial log regression and compare how they perform against the random forest classifier. It is also worth noting that other resampling methodology such as SMOTE could give us better performing model.

6. 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

Experiment_B roc_auc score = ~98

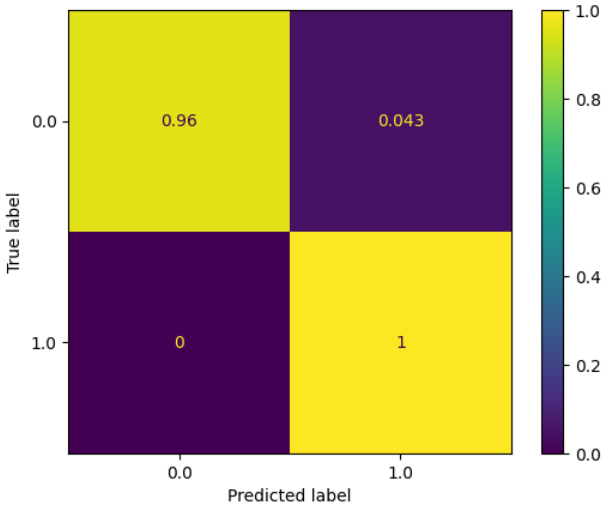
Experiment_C train dataset roc_auc score= Safe to assume that roc_auc score for this experiment is >98 although it's not very consistent indicating that overfitting might be present (not that significant, judging from this score)

```
array([0.99135549, 0.9832457 , 0.98814722, 0.97601146, 0.98285893])
```

Experiment_C train dataset recall score= It looks like the recall scores are not consistent, the more reason to believe that the model is overfitting

```
array([0.8877551 , 0.92857143, 0.95918367, 0.91836735, 0.96938776])
```

Experiment_C full dataset confusion matrix= Below is the confusion matrix for the full dataset (not undersampled). Performance is not an issue here, the only concern is whether the model is overfitting.



	From the results, it looks like that we were able to improve the performance of the model, as the scores are better than in experiment B. However, this model seems to have slight overfitting issues, decreasing our confidence that it would perform equally well on unseen data.
3.b. Business Impact	This model is definitely better than model A and perform marginally better than model B, however, the results suggest that it could be overfitting. We think that model B is better suited to be pushed for production despite lower roc_auc score but with better generalization for the unseen data. If model C is used instead, we could be overconfident on making the right decision on the unseen data and end up gaining less profit compared to model B.
3.c. Encountered Issues	1. Bad data quality in ht and the presence of missing data. We solve the issue by disregarding ht from our experiment and working around with the missing observations either with data engineering or removing the observations. 2. Imbalanced dataset, which causes more complications and needs extra effort to take care of the imbalanced dataset. 3. The test data came separated from the training data which require us to do different data preparation and engineering, also extra care should be applied to ensure that the test data set has the exact same features, format, and orders as the processed training dataset, i.e. we had to add dummy columns just to match the training, and reorder the features.

---

7. 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	Adaptive Boost algorithm seem to be able to achieve better scores than random forest classifier, however for this particular dataset and experiment, it also shows some degree of overfitting. We think that the marginal scores improvement may not be worth the risk of inconsistency on the unseen data.

4.b. Suggestions / Recommendations	Potential next step: 1. Push model B to production 2. Try different resampling techniques to model B and C 3. Compare different algorithms such as or polynomial LogRegression.