# EXPERIMENT REPORT

| | |
|---|---|
| **Student Name** | Daniel Alexander |
| **Project Name** | Assignment_1_Part_A |
| **Date** | 18/06/2023 |
| **Deliverables** | Experiment_A.ipynb<br>Random Forest Classifier<br>ROC Score<br><br>Github Repository:<br>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 model to be invested early and cheap before the young players are even drafted for NBA.<br><br>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** | The available statistics of the players alone are enough to determine the likelihood of them being drafted to NBA. If they are, then sporting apparel company should be offering those potential players beforehand to secure cheap investment. |

### 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 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 significant role to determine 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 70:30 ratio, hoping that by capturing the scarcity of the minority will produce better performing model.

Potential approaches to be used in future experiments are the resampling techniques, either change the ratio or the methodologies, 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
   - Ftr (Definition missing. Need definition, before using this feature)
   - Pfr (Definition missing. Need definition, before using this feature)
   - 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 succesful 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 scales sensitivity issues for some algorithm.

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

Potential approaches for future experiments, process feature ht as logically it's a significant factor for our purpose. Get clarification on the definition on ftr and pfr in case we could use it to make our model better performing.

## 2.c. Modelling

In this experiment we are using random forest classifier as this is a versatile algorithm which performs beautifully negating effects of outliers and is a good starting point to build our project. However, due to time constraint, we use the default hyperparameter as a starting point.

We did not use linear regression algorithm, as the algorithm is often inferior in terms of performance compared to the more sophisticated model such as random forest especially given the ample number of features available, assuming the relation is not linear.

For our future experiment we might try AdaBoost, polynomial log regression, and tweaking its hyperparameter and consider using grid search.

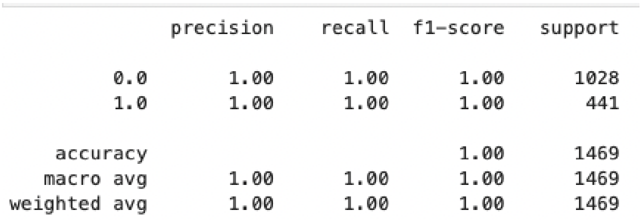
## 6. EXPERIMENT RESULTS

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

Naïve Model’s Recall score for positive class = 0

Experiment_A recall score= 1

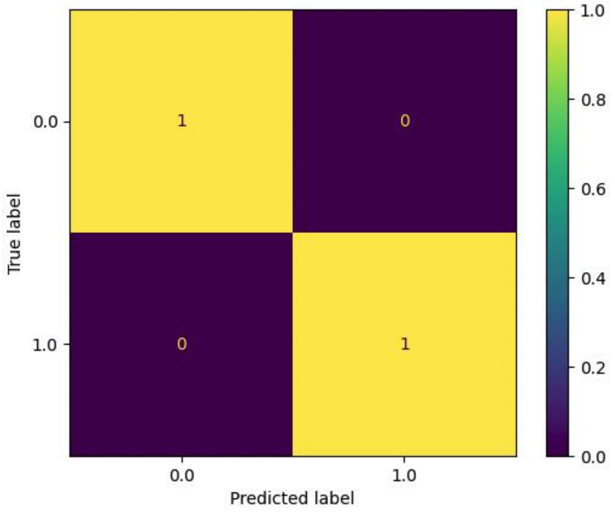


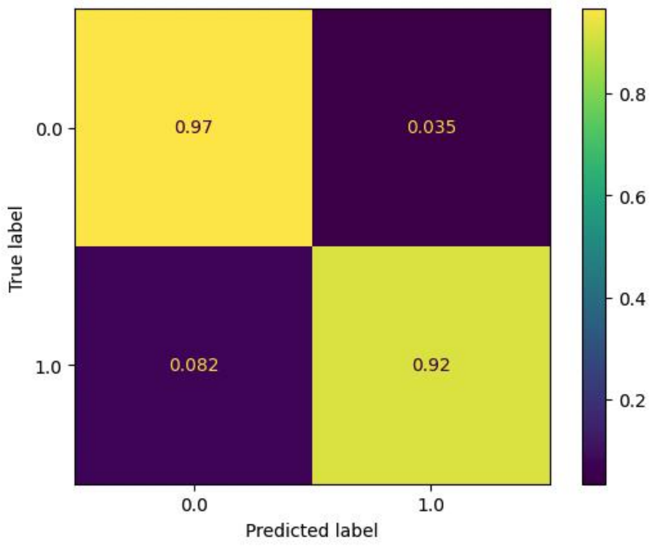
```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      1028
         1.0       1.00      1.00      1.00       441

    accuracy                           1.00      1469
   macro avg       1.00      1.00      1.00      1469
weighted avg       1.00      1.00      1.00      1469
```

Confusion Matrix for training (under sampled)



| | |
|---|---|
| | Confusion Matrix for validation (under sampled)<br><br> <br>Looks like the model that we use in this experiment is overfitting as the confusion matrix were not very consistent between training and validation. Potentially we can tweak the hyperparameter to reduce overfitting, or use more sample or different resampling techniques. |
| **3.b. Business Impact** | Since the model works like wonder in training but not general enough for unseen data, our model is not robust enough to be used for production. The model could lead to overconfidence in making investment decision to potential NBA draft player and will cause financial loss ultimately. |
| **3.c. Encountered Issues** | 1. Missing data definition for 2 features which need to be clarified. Therefore, we removed them from our experiment.<br>2. Bad data quality in ht and the Prescence of missing data disregarding ht from our experiment and working around with the missing observations either with data engineering or removing the observations.<br>3. Imbalanced dataset, which caused more complication and need extra effort to take care of the imbalanced dataset.<br>Issues number 1 and 3 are inherently present on the raw data, hence in the future we will have to deal with it. However, issues number 2 could be solved once explanation about the bad format and data dictionary have been provided. |

## 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** | The model produced from this experiment is absolutely better than the naïve model. However, it is overfitting and is not consistent enough to be used on unseen data. Nevertheless, we should do more experiments to compare different possible models. |
| **4.b. Suggestions / Recommendations** | Potential next step:<br>1. Try to mitigate the overfitting issue by tweaking the hyperparameter or use different resampling techniques or assumptions. Minimal effort and resources needed with potential improvement from overfitting issues.<br>2. Limit the features used to avoid overfitting. This approach would need more extensive domain knowledge and more trials to achieve better performing model.<br>3. Try different algorithm. This approach consumes much time and resources compared to the above 2.<br>4. If the company need any model in an instant, then we can arguably force this model to production as it will still give better prediction than the naïve model. |