EXPERIMENT REPORT

Student Name	Thirada Tiamklang
Project Name	Experiment B - Kaggle competition with Adaboost classifier model
Date	24 August 2023
Deliverables	<Experiment_B.ipynb> < Adaboost classifier model > <AUROC> <https://github.com/thirada2799/Thirada_AML_AT1.git>

1. EXPERIMENT BACKGROUND	
1.a. Business Objective	The goal is to predict whether a college basketball player will be drafted to join the NBA league based on their statistics for the current season. The results will be presented as player IDs and their corresponding probabilities or chances of being drafted by an NBA team. A higher AUROC score indicates that the model is better at correctly classifying positive examples as positive and negative examples as negative across various threshold values. This improves the accuracy of predictions, which is valuable for sport commentators and fans.
1.b. Hypothesis	The goal of the business is to predict the probability of the player being drafted in test dataset using AUROC score as a benchmark. The hypothesis underlying this approach is that the Adaboost classifier model with tuning hyperparameters will yield a higher AUROC value compared to the default model.
1.c. Experiment Objective	The business objective is to predict the probability of a player being drafted in the test dataset. To achieve this, we will train a model using the training dataset, which includes the target feature 'drafted.' Our focus will be on improving the AUROC score of the Adaboost classifier model by experimenting with various hyperparameters. Subsequently, we will use the trained model to predict the probability of drafting for the test dataset.

2. EXPERIMENT DETAILS

2.a. Data Preparation	Load the CSV files X_train.csv, X_val.csv, X_test.csv, y_train.csv, y_val.csv, y_test.csv, and raw_test.csv from the ../data/processed/ directory that we had saved from the previous experiment. Convert the arrays of y_train, y_val, and y_test into 1-dimensional arrays for training.
2.b. Feature Engineering	As we are utilizing the same cleaned data and feature selection choices for comparison, the dataset has undergone feature engineering, which includes the following steps: 1. Label encoding the categorical features including yr, type, team, player_id, and conf. 2. Scale data with StandardScaler 3. using Synthetic Minority Over-sampling Technique (SMOTE) since the ‘drafted’ is imbalanced. The new balance is 55,555 for each drafted and not drafted. 4. Split data into train, validation, and test set with ratio 80:20.
2.c. Modelling	With binary classification as the target variable, our approach involves training the first dataset, which includes the target variable, using an Adaboost classifier with hyperparameter tuning. This aims to enhance the AUROC scores across the training, validation, and test sets by experimenting with diverse hyperparameters. Subsequently, we will utilize the trained model to predict the probability of being drafted for the second dataset, which lacks the target variable. The process of selecting hyperparameters involves employing a random forest classifier and evaluating the Variance Inflation Factor (VIF) value. The final feature set, denoted as X, comprises four features: 'adjoe,' 'rimmade,' 'dunks_ratio,' and 'adrtg.' This approach enhances the model's predictive capabilities while focusing on interpretability and relevant features.

3. EXPERIMENT RESULTS

3.a. Technical Performance

Hyperparameters	AUROC
default	0.9530439222551397
(random_state=42)	0.9530439222551397
(n_estimators=100)	0.9578687494389083
(n_estimators=150)	0.9581509785438549
(n_estimators=200)	0.958140878893976
(learning_rate=0.05, n_estimators=150)	0.9736331807164019
(learning_rate=0.1 n_estimators=150)	0.974367088607595
(learning_rate=0.5, n_estimators=150)	0.9649631923871083

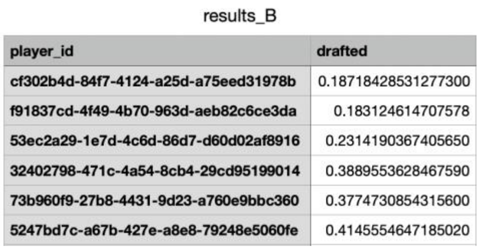
It can be seen that the Adaboost classifier model with *learning_rate=0.1, estimators=150 and random_state=None* provided highest score of AUROC on testing set. We will use this trained model to predict the probability of drafted on test dataset.

3.b. Business Impact

Based on the obtained results, the trained model's performance aligns with the hypothesis that tuning the hyperparameters will result in a higher AUROC value compared to the default model configuration, which consists of parameters 'learning_rate=0.1,' 'n_estimators=150,' and 'random_state=None.' This observation suggests that by fine-tuning the hyperparameters, we are able to achieve a more refined model that exhibits better predictive power. The improved AUROC score implies that the trained model can effectively discriminate between positive and negative outcomes, thereby enhancing its utility for real-world predictions.

Furthermore, the process of hyperparameter tuning helps mitigate the overfitting issue, as evidenced by the reduction in the gap between training and testing set performance. This indicates that the model's generalization capabilities have improved, making it more reliable in predicting outcomes for new, unseen data.

Sports commentators and fans can assess the probability of a player being drafted through the contribution of our generated result_B.csv file. An example of the result is presented below.



results_B

player_id	drafted
cf302b4d-84f7-4124-a25d-a75eed31978b	0.18718428531277300
f91837cd-4f49-4b70-963d-aeb82c6ce3da	0.183124614707578
53ec2a29-1e7d-4c6d-86d7-d60d02af8916	0.2314190367405650
32402798-471c-4a54-8cb4-29cd95199014	0.3889553628467590
73b960f9-27b8-4431-9d23-a760e9bbc360	0.3774730854315600
5247bd7c-a67b-427e-a8e8-79248e5060fe	0.4145554647185020

3.c. Encountered Issues

Tuning the hyperparameters on Adaboost has many ways to do that at first we do not know where to start to make the AUROC score improved. We solved this problem by changing one by one hyperparameters which included n_state, n_estimators, and learning_rate.

4. FUTURE EXPERIMENT

4.a. Key Learning	Tuning the hyperparameters of the Adaboost classifier model appears to enhance its efficiency, leading to higher AUROC scores on the testing set. Additionally, this process contributes to the reduction of the overfitting gap typically observed between the training set and the testing set. Despite these improvements, it's worth noting that the final AUROC score achieved after hyperparameter tuning is comparatively lower than that of the default model and the AUROC score obtained in the previous experiment. This outcome highlights the intricate interplay between hyperparameter settings, model complexity, and generalization performance. While hyperparameter tuning can enhance certain aspects of the model's performance, it can also impact other aspects, potentially leading to trade-offs between various metrics. It's important to strike a balance between model complexity and generalization ability to ensure that the model remains robust and reliable for new, unseen data.
4.b. Suggestions / Recommendations	The results clearly demonstrate that the AUROC score achieved using the Adaboost classifier model, specifically with hyperparameters 'learning_rate=0.1,' 'n_estimators=150,' and 'random_state=None,' outperformed both the default model and other models with different hyperparameters. Notably, the testing set yielded an impressive AUROC score of 0.974. Despite this success, it's important to acknowledge that there are still opportunities for improvement. Exploring alternative algorithms, models, or incorporating advanced feature engineering techniques, such as utilizing grid search for hyperparameter optimization, could potentially lead to further enhancements in predictive performance. Continuous experimentation and refinement are recommended to unlock the full potential of the model and achieve even better outcomes.