EXPERIMENT REPORT

Student Name	Thirada Tiamklang
Project Name	Experiment C - Kaggle competition with Polynomial Logistic, Adaboost, and Random Forest Classifier model using Grid search
Date	1 September 2023
Deliverables	<Experiment_C.ipynb> < Adaboost classifier model > < Polynomial Logistic regression model> < Random Forest 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 polynomial logistic regression, adaboost, and random forest classifier models with grid search will provide a different score of AUROC on the testing set.

1.c. Experiment Objective	The business objective is to predict the probability of a player being drafted using the test dataset. To accomplish this, we will train models on the training dataset, where the target feature is 'drafted.' Our aim is to enhance the AUROC score of the Polynomial Logistic Regression, Adaboost, and Random Forest classifier models through grid search. Subsequently, we will employ the best-trained model to predict the drafting probability 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 dataset, it has undergone feature engineering, involving the following steps: Label encoding the categorical features, including yr, type, team, player_id, and conf. Scaling the data using StandardScaler. Applying the Synthetic Minority Over-sampling Technique (SMOTE) due to the imbalance in the 'drafted' class. This results in a new balance of 55,555 samples for each drafted and not drafted class. Splitting the data into train, validation, and test sets in an 80:20 ratio. For this experiment, we will select all the features from the dataset after dropping and cleaning Nas. This choice of selection distinguishes it from experiments A and B.

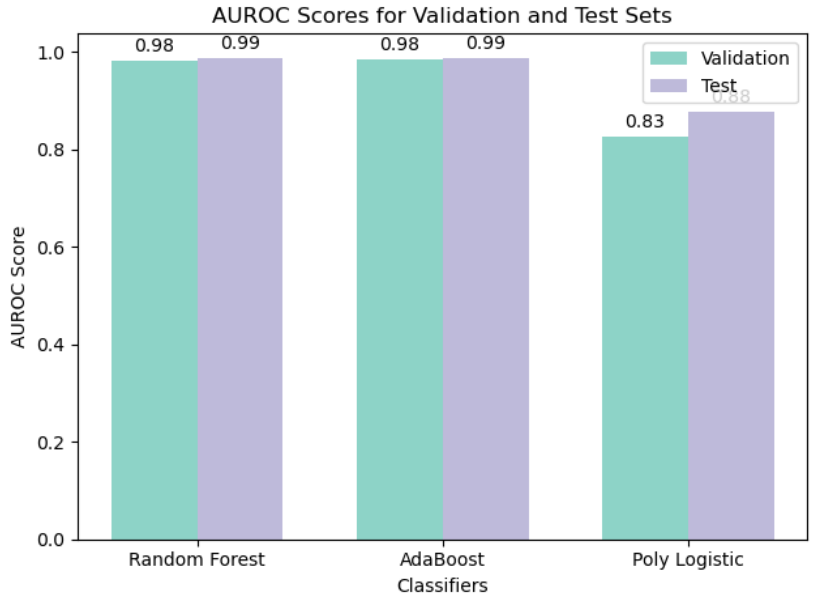
2.c. Modelling	With binary classification as the target variable, our approach involves training the first dataset, which includes the target variable, using Polynomial Logistic Regression, Adaboost, and Random Forest classifier models through grid search. This aims to enhance the AUROC scores across the validation, and test sets by experimenting with diverse models. Subsequently, we will utilize the best trained model to predict the probability of being drafted for the second dataset, which lacks the target variable. The hyperparameters used after grid search included: 1. LogisticRegression(random_state=42) 2. AdaBoostClassifier(learning_rate=0.1, n_estimators=200, random_state=42) 3. RandomForestClassifier(max_depth=10, min_samples_leaf=2, min_samples_split=5,random_state=42)

3. EXPERIMENT RESULTS

3.a. Technical Performance

Model using grid search	Set	AUROC
Polynomial Logistic Regression	Validation	0.8253925491717045
	Test	0.8768830236107371
Adaboost classifier	Validation	0.984433895439439
	Test	0.9885671963371937
Random forest classifier	Validation	0.9832030061567689
	Test	0.9866246970105036

It can be seen that the Adaboost classifier model with *learning_rate=0.1*, *n_estimators=200*, *random_state=42* provided the 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 results, the trained model's performance corresponds with the hypothesis that training multiple models using grid search yielded varied outcomes, with the best model, Adaboost, consisting of parameters 'learning_rate=0.1,' 'n_estimators=200,' and 'random_state=42.' On the testing set, the Random Forest and AdaBoost classifiers had virtually identical AUROC scores of 0.99. However, the AdaBoost classifier drove the Random Forest model significantly, scoring 0.9885671963371937 on the testing set, whereas the Random Forest scored 0.9866246970105036. This demonstrates that boosting is superior to bagging in this circumstance. The contribution of our created result_C.csv file allows sports broadcasters and fans to judge the likelihood of a player getting picked. A sample of the outcome is shown below. results_C player_id: drafted cf302b4d-84f7-4124-a25d-a75eed31978b: 0.23460271003027900 f91837cd-4f49-4b70-963d-aeb82c6ce3da: 0.24413019172308700 53ec2a29-1e7d-4c6d-86d7-d60d02af8916: 0.35606375920705000 32402798-471c-4a54-8cb4-29cd95199014: 0.2779346589146850 73b960f9-27b8-4431-9d23-a760e9bbc360: 0.41031817334618600 5247bd7c-a67b-427e-a8e8-79248e5060fe: 0.40876050153024900
3.c. Encountered Issues	Using grid search required a lot of space and time to run the model. Especially the Random Forest classifier model, which has many hyperparameters to decide on, and the concept of bagging We resolve this issue by reducing the hyperparameters for them to choose or search.

4. FUTURE EXPERIMENT	
4.a. Key Learning	This experiment attempts to determine whether model is superior in terms of bagging versus boosting. Furthermore, we use polynomial logistic regression to compare the results to the prior experiment while utilising all of the features as X. As a consequence, the Adaboost classifier outperformed the Random Forest (bagging) and the polynomial logistic regression in terms of AUROC. This demonstrates that boosting the weak learner improves the model.
4.b. Suggestions / Recommendations	The results reveal that forecasting the probability of a player getting picked using Adaboost with the following parameters: 'learning_rate=0.1,' 'n_estimators=200,' and 'random_state=42' provides the best performance among all the models we trained on the testing set, with 0.9885671963371937. This model may be used by the company to train and anticipate the target for sports broadcasters and spectators who want to know the results in real time for the current season.