EXPERIMENT REPORT

Student Name	Thirada Tiamklang
Project Name	Experiment A - Kaggle competition with polynomial logistic regression model
Date	18 August 2023
Deliverables	<Experiment_A.ipynb> <polynomial logistic regression model> <mse, mae, 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 polynomial regression model will yield a higher AUROC value compared to the baseline 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 comparing the AUROC score of the polynomial regression model with that of the baseline model. Subsequently, we will use the trained model to predict the probability of drafting for the test dataset.

---

2. EXPERIMENT DETAILS	
2.a. Data Preparation	1. Load two dataset by reading CSV. file. The train.csv dataset used for training the model and the test.csv dataset which is not contain the drafted feature will be used for predicting probability in the final result. 2. Clean NA: - Drop features that have NA more than 10k. - Replace the NA by zero in the columns that have NA in year 2009. - Drop the feature which is not relevant with the target. (num and ht) - Replace numeric by mean and categorical by mode
2.b. Feature Engineering	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, we will train the first dataset containing the target variable using a logistic regression classifier with polynomial features. This will allow us to assess the AUROC scores of the training, validation, and test sets in comparison to the baseline models. Subsequently, we will employ the trained model to predict the probability of being drafted for the second dataset, which does not contain the target variable. The hyperparameters were chosen using a random forest classifier and the VIF (Variance Inflation Factor) value. The final feature set, denoted as X, includes four features: 'adjoe,' 'rimmade,' 'dunks_ratio,' and 'adrtg.

---

3. EXPERIMENT RESULTS

3.a. Technical Performance

Model	set	AUROC
Baseline		0.5
Polynomial logistic regression	Training set	0.9718561683981886
	Validation set	0.970257766795252
	Testing set	0.9791487117335488

It can be seen that the polynomial logistic regression provided higher score of AUROC. 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 aligns with the hypothesis that the polynomial regression model will yield a higher AUROC value compared to the baseline model. This suggests that utilizing this trained model for predicting the target on the test dataset will offer enhanced discrimination between positive and negative outcomes.

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

results_A

player_id	drafted
cf302b4d-84f7-4124-a25d-a75eed31978b	8.9401244423429e-06
f91837cd-4f49-4b70-963d-aeb82c6ce3da	3.10369936908342e-09
53ec2a29-1e7d-4c6d-86d7-d60d02af8916	0.0001559227584378240
32402798-471c-4a54-8cb4-29cd95199014	0.002349644766462260
73b960f9-27b8-4431-9d23-a760e9bbc360	0.0008587856870263560
5247bd7c-a67b-427e-a8e8-79248e5060fe	0.00992100557671046O

3.c. Encountered Issues	A significant number of missing values are present in various features. Deciding whether to drop these values or replace them with zeroes poses a challenge, as the outcomes may vary for each individual player. To address this, several methods for handling missing values were applied, as detailed in the data preparation step. Additionally, the situation is compounded by the existence of two datasets that require attention. The first dataset contains multiple features, including the target variable, while the second dataset does not. This challenge was resolved by training the model on the first dataset and subsequently applying it to the second dataset."

---

4. FUTURE EXPERIMENT	
4.a. Key Learning	Efforts toward refining the accuracy of predicting a player's probability of being drafted hold profound implications. Achieving higher precision in this prediction has the potential to ignite enthusiasm among fans and stimulate various sports-related businesses. The accurate forecasting of players' draft probabilities not only generates excitement but also carries potential economic and strategic benefits for stakeholders in the sports industry
4.b. Suggestions / Recommendations	It's evident that the AUROC score achieved through polynomial logistic regression training surpassed that of the baseline model, registering an impressive score of 0.97. However, there remains room for enhancement by exploring alternative algorithms, models, or implementing advanced feature engineering techniques. It's worth noting that while the current AUROC score stands out, the optimal score is 1, indicating flawless classification performance. Further experimentation is also recommended.