EXPERIMENT REPORT

Student Name	Tim Wang
Project Name	ADSI Week 3 Project Report
Date	23/11/2022
Deliverables	<wang_tim_week3_xgb.joblib> XGBoost with Smote Upsampling and Hyperparameter Tuning> Github: https://github.com/timtamothy/adv_dsi_lab_2/tree/Tim_W

---

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	Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results? The goal of this project is to see if one can predict if a basketball rookie player can be retained for longer than 5 years. This may give insights to how a team is making decisions and what KPIs are important to monitor and improve. It can then be used to help players improve their game and impact to the team.
1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, Hypothesis: With last week's SMOTE upsampled data, different algorithms other than last week's Random Forest may improve the predictive power and a better model may be presented. Hypothesis 2: Hyperparameter tuning of these models can further improve the model. Hypothesis 3: The use of the entire training set (recombining the split train + validation set) may improve the score further after an optimised model is selected.

1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. Expected Outcome: An improvement in both the AUROC compared to Random Forest, last week's model, and last week's submission to Kaggle. Last week's AUROC score with Random forest was 0.7172 and a corresponding Kaggle score of 0.699. I anticipate that a better model will improve both the AUROC score on the validation set and the score on Kaggle.

---

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	Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments
	
	Steps:
	1. Utilise the SMOTE upscaled data to test new binary classifiers: Naive Bayes, K-Nearest Neighbours, Support Vector Machines, XGBoost
	2. Compared each model's AUROC score with each other and the Random Forest algorithm
	    a. Chose XGBoost for a high AUROC score
	3. Tuned XGBoost Hyperparameters to increase AUROC score higher
	4. Predicted probabilities using highest score trained model
	    a. Used only the training data without validation data
	5. Predicted probabilities using retrained model
	    a. Used training and validation data to retrain model
	6. Output for Kaggle submissions
	7. Retuned XGBoost Hyperparameters to increase generalisability

2.b. Feature Engineering	Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments *Note: These steps were the same as the prior week* 12 Features were created from existing features with poor distributions. Theses included: MIN2, AST2, PTS2, FGM2, FGA2, FTM2, FTA2, OREB2, DREB2, REB2, STL2, TOV2 Most of these had a left skewed distribution starting at or near zero with a right tail. After cube rooting these features, the distributions appeared normal. Future work: FT% and BLK don't exhibit great distributions after the cube root. A different function or manipulation may be used to make this distribution better.

2.c. Modelling	Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments I chose classifiers that had more complexity than decision trees and random forest. These classifiers also contained hyperparameters that could be tuned to increase its predictive power. Out of all the models trained, the best was XGBoost. The XGBoost model with the highest AUROC score for the validation data was n_estimators = 100, learning_rate = 0.1, and max_depth = 5. However, this model was determined to have a lower score on Kaggle than last week's Random Forest. This may indicate overfitting. Therefore, the learning_rate and max_depth are hyperparameters that may potentially improve this model in the future to better generalise the model to new data.

---

3. 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	Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.
	
	1. [Prior week] Random Forest and SMOTE Upsampled Data:
	    a. AUROC = 0.7172
	2. Gaussian Naive Bayes
	    a. AUROC = 0.6578
	3. K-Nearest Neighbors
	    a. AUROC = 0.7274
	4. Support Vector Machines
	    a. AUROC = 0.7175
	5. XGBoost
	    a. AUROC = 0.8960
	
	Last week's Kaggle score: 0.69904
	This week's Kaggle scores:
	1. XGBoost with only test data
	    a. Kaggle Score = 0.6119
	2. XGBoost with test data and validation data for training
	    a. Kaggle Score = 0.62666
	3. XGBoost with reduced AUROC (detuned hyperparameters)
	    a. Kaggle Score TBD, Team ran out of submissions

3.b. Business Impact	Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others) The improvement in the AUROC score tells us that we may have found a good model for this particular dataset. It is able to correctly classify the validation dataset better than other models by a good margin. However, the lower kaggle score indicates that we may be dealing with some overfitting. In addition, utilising both the training and the validation data to retrain the model showed an improvement in the Kaggle score. This may also be an indicator that the model is overfitting. The impact to the business is that we may not realistically achieve a score as high as the XGBoost's AUROC. Also, this indicates to the business that the training data may be a bit too specific and/or more data could be helpful in improving the accuracy of the data.

3.c. Encountered Issues	List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments. ISSUES: 1. UNSOLVED: a. Jupyter Lab/notebook is not currently plotting data b. Saving the dataframes for future use is currently not working in the anaconda python environment c. A de-tuned XGB with AUROC score still higher than other models needs to be submitted to Kaggle to determine if detuning has the intended effect of better generalising.

---

4. 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	Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. High AUROC isn't necessarily great. While exciting to see on the validation data, that doesn't always translate to a better model when working with the test data. However, I do think that it's a good thing we found a model with a high AUROC. That means we can titrate the AUROC score down to generalise better rather than trying to increase an AUROC score with an unknown ceiling. This way, we might find the point at which AUROC and the Kaggle score meet in the middle.
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. Next steps: ● Determine Kaggle score for de-tuned model. ● Continue to tune hyperparameters, maybe even do a grid search to find best hyperparameters that generalise best.