EXPERIMENT REPORT

Student Name	Tim Wang
Project Name	ADSI Week 2 Project Report
Date	17/11/2022
Deliverables	<wang_tim_week2_model2> <Random Forest with Smote Upsampling>

---

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: Using an upsampling technique will improve the prediction power of the model. It is worthwhile because the target variable contains fewer class 0 observations than class 1 observations. The hypothesis is that balancing the model will help prevent the model from putting too little training on one class. Hypothesis 2: SMOTE upsampling technique may be better than pure upsampling as it introduces a bit more variability in the data so that it may generalise better. Both will be compared.
1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate t he goal you are expecting. List the possible scenarios resulting from this experiment. Expected Outcome: An improvement in both the AURUC compared to baseline, last week's model, and last week's submission to 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 Due to the poor performance of last week's logistic regression model, this experiment will not reduce any features that are displaying collinearity. Steps: 1. Improve the distribution of all features using a **cube root function.** a. This could be improved using other functions. Specific features could do better with other methods b. FT% and BLK don't exhibit great distributions after the cube root. This could be improved in the future. 2. Removal of Id column 3. Rebalancing of the data using: a. Upscaling b. SMOTE upscaling 4. Splitting the data into train/validation sets for both upscaled and SMOTE upscaled data a. Stratified split used to make sure the train and validation data contain equal proportions of target values
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 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 For this experiment, I decided not to use a linear model such as logistic regression as it appeared to perform poorly in prior weeks. It also is restricted due to not being able to use features with multi collinearity. For this week, all features were left here. A Random Forest model was used to test whether or not better results could be attained. Random Forest and Decision Trees also have the benefit of determining feature importance. Should this model perform well, removing some features following the analysis of feature importance could improve the model.

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. AUROC Score using Random Forest and Upsampled Data: a. 0.7088 2. AUROC Score using Random Forest and SMOTE Upsampled Data: a. 0.7172 Both of these models performed substantially better than the prior week's model. In addition, it appears that upsampling using the SMOTE algorithm provides better predictive power than the simple upsampling. A reason for this may be because SMOTE introduces some more variability into the upsampled data rather than just simple replication of the data. This could better generalise and capture the minority class. Last week's Kaggle score: 0.5 This week's Kaggle score: 0.69904
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 predictive power is a good sign that this model will have favourable results for the business and the players. With this model, understanding feature importance would be helpful in understanding the importance of the features which can be used as KPIs for players. This impacts the business as it can then help the team and coaches make decisions on how to improve these KPIs.
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: Reinstallation of Windows necessitated reinstallation of python and all related packages. Used Anaconda for quick and dirty installation. a. Some packages were not found in Anaconda's library b. Jupyter Lab/notebook is not currently plotting data 2. SOLVED: EDA package crashes when trying to show correlation plot if a feature and its cube rooted feature are used for analysis. a. Removed the original feature and correlation plot worked. 3. UNSOLVED: Need a better method to split X and y from dataset. Current use of .pop changes the original dataset.

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. The unbalanced nature of the outcome did influence the predictive power of this single algorithm. With the data balanced, now other algorithms could also perform better as well. Balanced data should now be used on other algorithms so we can compare how different models perform.
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 feature importance and reduce features to see if model performs better ● Use balanced dataset on other algorithms to see if they have better predictive performance ● Use balanced dataset - unimportant features to see if that has better predictive performance as well