EXPERIMENT REPORT

Student Name	Olivia Dewi (Part of Team 2: Tim Wang, Federico Gonzales, Olivia Dewi)
Project Name	Kaggle competition - week 3
Date	23/11/2022
Deliverables	IamAHEAD/adv_dsi_assignment_3 (github.com) <dewi_olivia_week3_RandomForestRegressor> <RandomForestRegressor>

---

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?* We are given the data set of 8000 rookie NBA players with their playing statistics. We are asked to predict the probability of these rookies lasting at least 5 years in the league based on its stats. The results of this project can be used by professional NBA teams to recruit talented members at the earliest stage of their career before anyone else see their potential. The capability to detect talent at a very early stage allows poachers to approach the promising players who can then be mentored to reach their full potential. Basketball teams full of stars are very valuable to club owners.
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, - SMOTE can be used to create synthetic class samples of minority class ('TARGET_5Yrs'=0) to balance the distribution. *REASONS:* o There is a big skew in the training data in favour of players with 'TARGET_5Yrs'=1 o SMOTE will increase the size of the training data set, and because it is not generally duplicating (but creating new data points that are slightly different to the original), it also increases variety and reduces overfitting. - We can reduce overfitting by limiting the height of the Random Forest Model , and reducing the size of n_estimator

	REASONS: • From last week's experiment, I notice that the accuracy score tested on validation dataset constantly outperforms the Kaggle AUROC score. This suggests that the model performs very well on the training data, but not so good with unseen data. • To increase Kaggle AUROC score, I need to reduce the overfitting problem. - To address skewness in the dataset, remove outliers, and normalise the distribution shape. I cube root on the distributions of features with right tail *REASONS:* o For most of the data, the features are heavily skewed. o Cube rooting will normalise the distribution in dataset and reduce the distance between points.
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. - The creation of synthetic data in SMOTE is a superior oversampling option, in comparison to duplication (i.e. oversampling technique). AUROC score in Kaggle should increase with this technique. - There is a fine balance between overfitting and underfitting a model. Limiting the height and reducing the number of n_estimator can backfire and instead introduce bias. - Taking a cube root of data set will normalise the dataset distribution set and it should reduce overfitting.

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

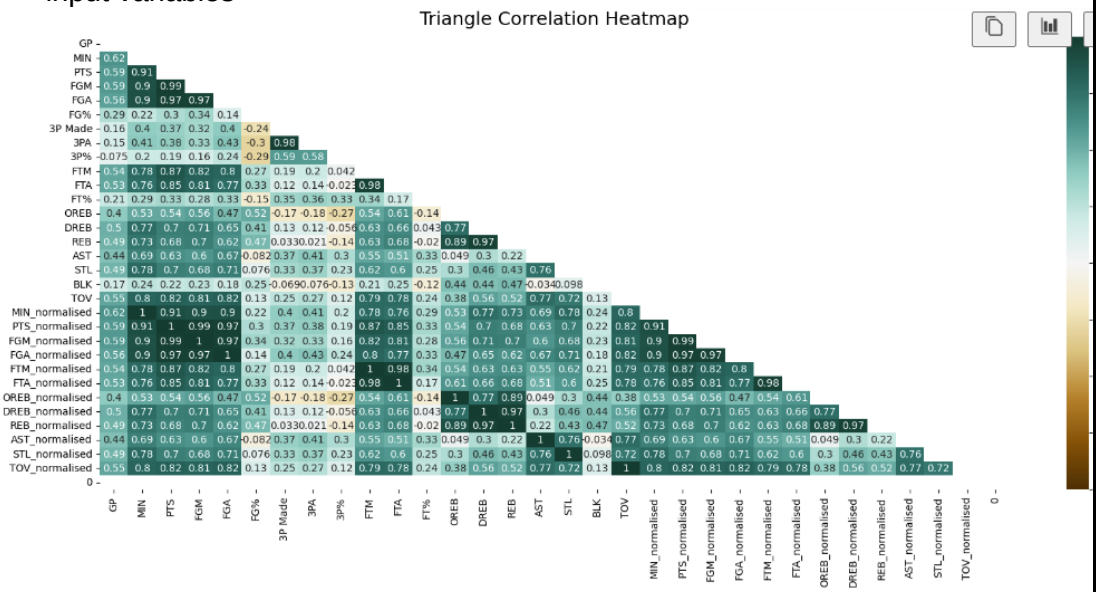
Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments

- Removing data with negative values in the training data.
- Identifying and removing columns with high correlation (>90%) and containing little information.
- Applying SMOTE to create synthetic dataset to balance the imbalance distribution.
- To address skewness in the dataset, remove outliers, and normalise the distribution shape. I cube root on the distributions of features with right tail

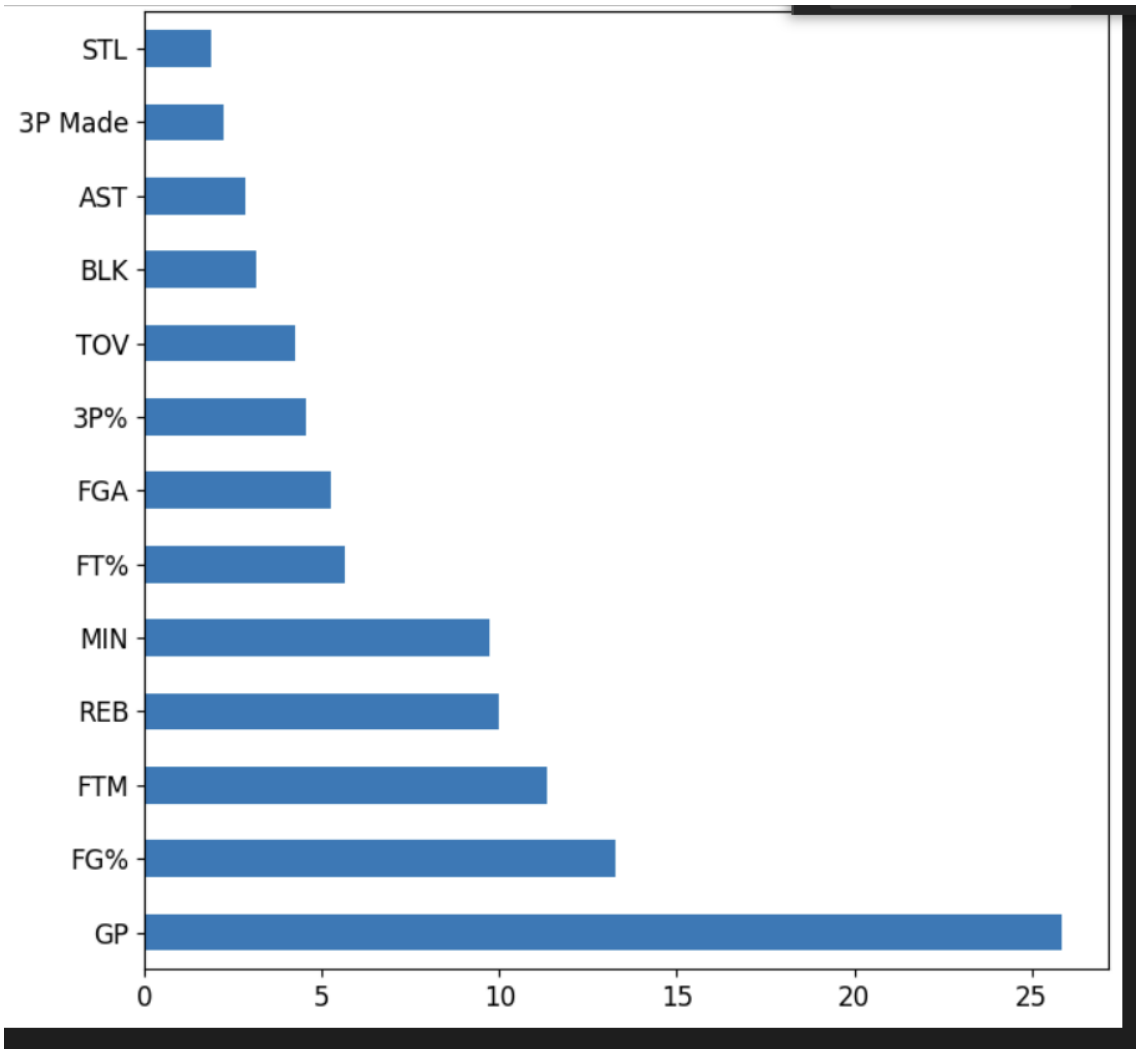
2.b. Feature Engineering

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

- I have added a triangle correlation heatmap to see the relationship of all the input variables



- Incorporating regressor feature importance into the report, for continuous improvement and transparency on contributing features to the model (below)



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 tried incorporating XGBoost and finetuning the parameters, but the ROC score is still lower than Random Forest.

- I find Random Forest Classifier works best with my dataset, but last week I have made too many assumptions without much experimenting.
- I find that the accuracy score on validation dataset constantly outperforms the Kaggle AUROC score. This suggests that the model performs very well on the training data, but not so good with unseen data.
- This week, I experiment to fine tune the hyperparameters to avoid overfit.

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.

AUROC Score = 0.71 (improvement of ~0.02). Potential root causes:

- Overfitting: I find that the accuracy score on validation dataset constantly outperforms the Kaggle AUROC score. This suggests that the model performs very well on the training data, but not so good with unseen data. The hyper parameter in Random Forest Classifier model can still be fine tuned for better accuracy.

- SMOTE sampling: The training dataset was tripled after SMOTE sampling, from ~6000 lines to 18000. A large portion of the training dataset was synthetic. I need to run experiments to determine the right balance.

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)

- With a score of 0.71, the model is not reliable and cannot be operationally deployed to poach players.
- We could lose good players, and pick up weaker players into the team.

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.

- Solved: SMOTE is a superior oversampling technique to imbalance data set.

- These issues are still not solved:
    - vetting out the outliers,
    - finding the solution to overfitting tree-based model,
    - Finetuning the hyperparameters to achieve better AUROC score.
    - Experimenting with the number of synthetic samples in SMOTE to balance the 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. - Continue to fine-tuning Random Forest Classifier and XGBoost hyper parameters to reduce overfitting (by reducing the maximum depth, the number of trees, etc) ○ Experimenting with Hyperopt to search for the best combination - Continue to experiment with the size of SMOTE sampling. - Increasing the size of n_estimators in RandomForestClassifier caused the model to overfit. I will not add more to the n_estimator
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. - (1) Model selection - uplift 5% in accuracy - (2) Hyperparameter finetuning – uplift 10% in accuracy - (3) Fixing imbalanced data via SMOTE- uplift 5% in accuracy - (4) Feature engineering - uplift 5% in accuracy - (5) Identification and elimination of outliers – uplift 5% in accuracy