EXPERIMENT REPORT

Student Name	Olivia Dewi (Part of Team 2: Tim Wang, Federico Gonzales, Olivia Dewi)
Project Name	Kaggle competition - week 2
Date	16/11/2022
Deliverables	<dewi_olivia_week2_RandomForestRegressor> <RandomForestRegressor>,<XGB>

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, - Removing outliers is essential to prevent our model from learning an incorrect pattern. For example data with -ve values need to be removed (such as -ve games played). - To obtain better results, we need a curated training data so the population is a good representative of the test population**.** - Players with good attributes such as higher points per game, 3P Made, STL are likely to have career length >= 5 years. - There is a high correlation between the attributes of these players, we need to take these out to prevent multi collinearity. - There are measures within the data that contain overlapping information. These need to be trimmed. - The training data seems skewed towards players with career length >=5 years, with column ‘TARGET_5Yrs’ = 1 for approximately five times of the

	‘TARGET_5Yrs’ =0. Undersampling techniques potentially can improve our predictions on unseen data. - Oversampling the minority may not improve the prediction results because it may cause overfitting (as it is basically copying exactly the same data). - There are many columns in the training data that are represented in different unit of measurement (e.g. some in %, some in points, some in time). A tree based model may cause an overfit, but for now let’s experiment with this. - Adding the number of forests to the RandomForestClassifier and XGBoost models will increase the accuracy of the predictions. - XGBoost may produce more accurate predictions than RandomForestClassifier, as it assigns weighting to the trees (and essentially eliminate the poor performing tree in the process).
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. - A tree based model with wide columns may cause an overfitting problem. - Removing too many columns may reduce the accuracy of our predictions. To incorporate ‘feature importance’ feature into our model to measure the predictive power of each input column. - Identification and elimination of outliers in the training data. - Finetuning hyperparameters to achieve better accuracy, without sacrificing too much of agility (i.e. optimising training time). - The experimental way to oversample minority data may not achieve the ideal outcome. To look at ways to optimise the sampling ratio on unseen data.

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

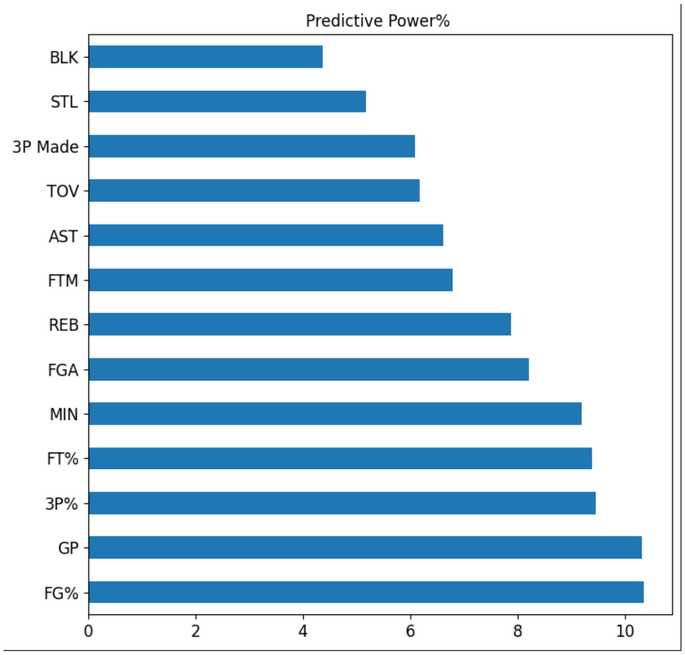
- Undersampling data for majority class, to fix the imbalance in the training data set. In the future, to explore different methods (e.g. SMOTE) to fix the imbalanced data.
- Remove column Id from the input column, the id number should not have an impact to the features of these players
- Oversampling the minority did not improve the prediction results because it had caused overfitting (as it is basically copying exactly the same data).
- Removing columns with high correlation (>90%) and containing little information.
- To detect and remove outliers from the dataset (e.g. negative values in Games Played).
  - Removing these may exacerbate the imbalanced dataset (as I am shaving off the bottom performing players and thus reducing the population in the TARGET_5Yrs = 0 bucket).
  - I need to reconstruct the training data and ensure its statistics are representative of the test data population
-

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

- The columns FGM/MIN/PTS/FGA have extremely high correlation. To prevent multi collinearity, I have removed these columns from the input columns and kept FGA only.
- The columns MIN/PTS/FGM have extremely high correlation. To prevent multi collinearity, I have removed these columns from the input columns. I will take just the MIN
- There is a high correlation between 3P Made and 3PA. To prevent multi collinearity, I have removed these columns from the input columns. I will take just the 3P Made
- There is a high correlation between FTM and FTA. To prevent multi collinearity, I have removed these columns from the input columns. I will take the FTM
- There is a high correlation between OREB/DREB/REB To prevent multi collinearity, I have removed these columns from the input columns. I will take the REB

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

<table>
<tr>
<td></td>
<td>


</td>
</tr>
<tr>
<td><b>2.c. Modelling</b></td>
<td>
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 have tried Random Forest Classifier and trained it on the TARGET_5Yrs column and generate probability for a player lasting >=5Yrs in the league.
- For this experiment, I have chosen the default setting of a random forest regressor with unconstrained depth and 300 trees. The hyperparameters can continue to be fine tuned later.
- I have tried XGBoost and trained it on the TARGET_5Yrs column and generate probability for a player lasting >=5Yrs in the league.
- I have decided not to use Logistics Regression model because the performance was poor and it is too simple. The more complex model would perform better.
</td>
</tr>
</table>

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 = 0.68522 (improvement of ~0.023). It can be improved further with the following logic: - aligning the training data with the test data - vetting out the outliers - experimenting to get the most optimal mix of 'TARGET_5Yrs' to mimic the real population (by using SMOTE) - experimenting with other models e.g. SVM/GBM -fine tuning the tree model hyperparameter by using datagrid - feature engineering experiments
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 0.68522, 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: feature engineering & finding the correlation between input columns - Not solved: ○ vetting out the outliers, ○ finding the solution to overfitting tree-based model, ○ finding the optimal mix to represent the population, ○ Finetuning the hyperparameters to achieve better AUROC score.

---

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. - Input columns with high correlation can cause overfitting - Data exploration is crucial, to identify abnormalities in the dataset. - Increasing the number of forests in the tree-based model also increases accuracy, at the expense of longer training time. - XGBoost is versatile and easy to use. I intend to spend more effort and experiment time is to finetune the hyper-parameter for better AUROC score.
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 15% in accuracy - (2) Hyperparameter finetuning – uplift 10% in accuracy - (3) Fixing imbalanced data - uplift 10% in accuracy - (4) Feature engineering - uplift 5% in accuracy - (5) Identification and elimination of outliers – uplift 5% in accuracy