EXPERIMENT REPORT

Student Name	Yatindra Vegunta
Project Name	NBA_Career_Prediction_Week3
Date	15/11/2022
Deliverables	Model Testing.ipynb Experiment report

---

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? **Goal:** To predict if a rookie player will last at least 5 years in the NBA league based on their current stats. **Use:** The results will help the business direct their resources towards players with the greatest talent and potential. **Impact of accurate/inaccurate results**: With accurate results, the NBA league can be more selective and they will be able to cultivate more successful players and teams. This can increase profits for the business by encouraging viewership, sponsorship deals, merchandise sales etc. and heighten the status of their 'brand'. Inaccurate results could have an adverse impact on players' career paths. Those with the potential to do well may be weeded out from the league unnecessarily.
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 The insights we are seeking revolve around our goal of predicting if a rookie player will last at least 5 years in the NBA league based on their current stats. As I tested the logistic regression with some columns removed, I want to test the same with other models without removing the columns. Models with automatic hyperparameter tuning will be utilized to improve time efficiency over trial and error.

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. Improved XGBoost model score compared to previous week.

---

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 The steps taken for data preparation are as follows: ● Split train data into multiple files ● Split data into 80:10:10 ratio (Training, validation and test) ● Removing Id column as it is not required ● Using all columns except for Id and Target_%yrs for normalisation ● Using Min Max Scalar for training dataset Cleaning the dataset is the most important and vital aspect of any data exploration which is directly corelated to the accuracy of the model. It is vital to understand the outliers and noise in the dataset.
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 Features were selected using sequential feature selection, where 13 features listed below were selected: `['GP', 'MIN', 'PTS', 'FGM', 'FGA', 'FG%', '3P Made', '3PA', 'OREB', 'DREB', 'REB', 'BLK', 'TOV']`

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
	
	• XGBoost: Model from week 2 was executed using teams cleansed file, to see if the ROC AUC would increase compared to previous week. This however resulted in a lower Kaggle score of 66.15 when compared 68.76 previous week. This model was then improvised with a new model
	
	• A new Xgboost model was created to increase the score, with the best fit hyperparameters used are as below:
	`hyp_params['max_depth'] = best['max_depth']`
	`hyp_params['learning_rate'] = best['learning_rate']`
	`hyp_params['min_child_weight'] = best['min_child_weight']`
	`hyp_params['subsample'] = best['subsample']`
	`hyp_params['colsample_bytree'] = best['colsample_bytree']`
	`hyp_params['min_split_loss'] = best['min_split_loss']`
	`hyp_params['max_delta_step'] = best['max_delta_step']`
	`hyp_params['grow_policy'] = grow_policy[best['grow_policy']]`
	`hyp_params['booster'] = booster[0]`

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.

Baseline Accuracy:

	model	t_accuracy	t_precision	t_recall	t_F1	v_accuracy	v_precision	v_recall	v_F1
0	Base	0.833594	0.833594	1.0	0.757942	0.83375	0.83375	1.0	0.758161

Training ROC:

```
Average absolute error: 16.625%
ROC: 0.70114
```

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)

Our baseline model performs very poorly and struggles to predict between classes within the imbalanced dataset. Our initial expectations were correct as our Random

	Previous XGBoost model performed better than our Baseline AUROC. Without hyperparameter tuning however, it is clearly overfitting with a score of 1.00 for the training set. Whilst it can perfectly predict the probability of a rookie player lasting longer than 5 years in the NBA league with the given data, it performs significantly worse with unseen data as evidenced through the validation AUROC of 0.701. The implementation of this model would result in a misallocation of resources towards new players that may not have as much potential to succeed, affecting the business in terms of future revenue and reputational value. Automatic hyperparameter tuning was performed via sequential feature selection. The AUROC validation score has also been improved to 0.71, resulting in better predictive ability.
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. 1. Complicating the model understanding and not able to generate the desired results. I had take a step back and re think on how to approach the model.

---

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. We will be focusing on the final team report for the coming week by identifying the best fit model among the team members
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. Further work will be carried out as part of self learning