EXPERIMENT REPORT

Student Name	Yatindra Vegunta
Project Name	NBA_Career_Prediction_Week2
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. By performing the data exploration again this week, I have not removed any columns except for Id. The nulls and invalid datapoint were removed. By performing the below models, I expect to choose the best fit model and have hyperparameter applied to the result. Models experimented are as below: • Logistic regression • KNN Euclidian • KNN Manhattan • XGBoost The Objective is to find the best fit model for further exploration in the coming 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:

- Load the dataset and display the first 5 rows
- Present the summary of each column and dimension of the data frame
- Display the descriptive statistics
- Plot the 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. Each column of both test and train data were assessed for duplicates, nulls. The descriptive statistics highlight the potential outliers and unreasonable values of each feature and provide guidance on the data cleansing step. The visualisation provides further insight on the distribution of the target variable and the correlation between each feature. These graphs imply the existence of common data issues - imbalance data and collinearity.

2.b. Feature Engineering

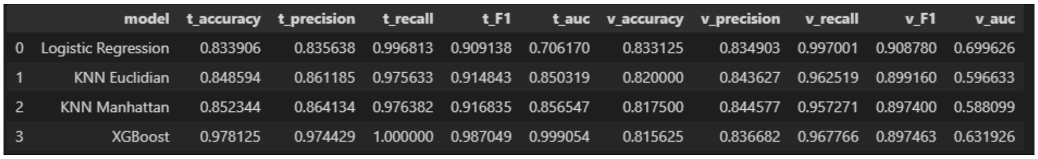
Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments

No features were created in this weeks explorations, as I wanted to explore the data without creating any features.

2.c. Modelling

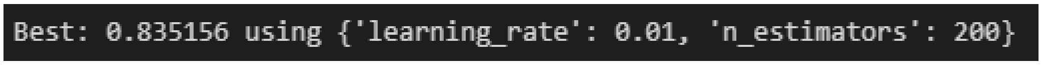
Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments

- A Baseline model was created to understand the results, which was later compared with the above listed models.
- The multi model provided the below result, which proved the xgboost was the models to be selected for investigation and exploration.



	model	t_accuracy	t_precision	t_recall	t_F1	t_auc	v_accuracy	v_precision	v_recall	v_F1	v_auc
0	Logistic Regression	0.833906	0.835638	0.996813	0.909138	0.706170	0.833125	0.834903	0.997001	0.908780	0.699626
1	KNN Euclidian	0.848594	0.861185	0.975633	0.914843	0.850319	0.820000	0.843627	0.962519	0.899160	0.596633
2	KNN Manhattan	0.852344	0.864134	0.976382	0.916835	0.856547	0.817500	0.844577	0.957271	0.897400	0.588099
3	XGBoost	0.978125	0.974429	1.000000	0.987049	0.999054	0.815625	0.836682	0.967766	0.897463	0.631926

- A grid search on xgboost was performed to understand which hyperparameters resulted in the best performance of the model. Below is the result of this exercise:



```
Best: 0.835156 using {'learning_rate': 0.01, 'n_estimators': 200}
```

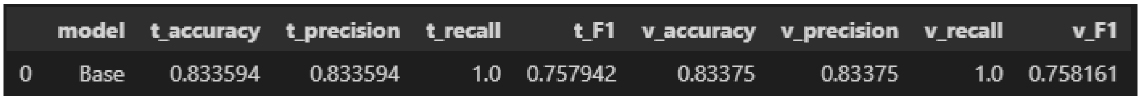
3. EXPERIMENT RESULTS

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

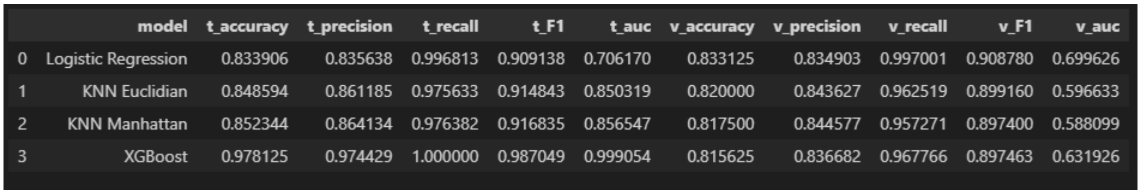
Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.

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 Accuracy:



	model	t_accuracy	t_precision	t_recall	t_F1	t_auc	v_accuracy	v_precision	v_recall	v_F1	v_auc
0	Logistic Regression	0.833906	0.835638	0.996813	0.909138	0.706170	0.833125	0.834903	0.997001	0.908780	0.699626
1	KNN Euclidian	0.848594	0.861185	0.975633	0.914843	0.850319	0.820000	0.843627	0.962519	0.899160	0.596633
2	KNN Manhattan	0.852344	0.864134	0.976382	0.916835	0.856547	0.817500	0.844577	0.957271	0.897400	0.588099
3	XGBoost	0.978125	0.974429	1.000000	0.987049	0.999054	0.815625	0.836682	0.967766	0.897463	0.631926

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 Forest models 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 0.67. 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 Random Search to reduce this overfitting. Overfitting has been reduced on the training set with an AUROC score of 0.78. 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. Issues encountered during data preparation and feature engineering were listed and discussed in sections 2.a. and 2.b.
2. Overfitting with Random Search. Countered with Hyperparater tuning - Random Search.
3. Collinearity - Points Per Game (PTS), Field Goals Made (FGM), Field Goals Attempts (FGA) are all highly correlated at around 97-99%. Similar case for (Free

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. This week, I executed the code with more cleansing of data through removing null and duplicates, which we missed in the last experiment. Though the model performed well on the training dataset, the score on kaggle proved that there is scope for improvement in the model. I would like to continue working on the xgboost model in the coming weeks and try and enhance the overall score. I would like to try further experimenting with the methods last learned in this weeks session.
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. Potential next steps include: 1. I would like include the methods learnt in this weeks class to observe how drastic will be the change in scores.