EXPERIMENT REPORT

Student Name	Sampath Pitchandi
Project Name	NBA Career Prediction
Date	21-Feb-2021
Deliverables	P_Sampath_Week03-01-XG-model-train.ipynb P_Sampath_Week03-02-XG-model-train.ipynb P_Sampath_Week03-03-XG-model-train.ipynb P_Sampath_Week03-04-XG-model-train.ipynb P_Sampath_Week03-05-RF-model-train.ipynb P_Sampath_Week03-06-RF-model-train.ipynb P_Sampath_Week03-07-LoReg-train.ipynb P_Sampath_Week03-08-GMM-train.ipynb P_Sampath_Week03-09-Stage1-train.ipynb P_Sampath_Week03-10-Stage2-final-model-train.ipynb P_Sampath_Week03-11-predict.ipynb

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	----- the section below this line is from the previous experiment ----- The main objective of this project will be to determine/predict if an NBA rookie will last at least 5 years from the given set of data. With this experiment, we are trying to see if a rookie player will be able to survive the first 5 years of the NBA league. Joining the NBA league is a big deal for any basketball player. And this prediction model can be used by various sets of people to track the performance of each player through their journey. For sports commentators & fans, they can keep track of the most important players. And for the teams, they can focus on picking the right players from the very beginning based on the predictions.
1.b. Hypothesis	Was intrigued by the conversation we had during our last class on model stacking. This week's training was to try model stacking + derive some insights from Variable Importance by Permutation and Partial Dependence Plot. The idea is to perform the following STAGE 1 Model: - Train an RF with Up-sampled Data - Train an RF with Up-sampled Data - Train an XG with Down-sampled Data

	- Train an XG with Down -sampled Data - Train a Logistic Regression model - Get the Probability from GMM STAGE 2 Model: - XGBoost Stack all the outputs from these models into the raw data, then apply XGBoost without any tunning. Doing this for the first time so not sure if this is the right way, I already had the models trained the first 4 models and wanted to improve upon it, so continued with it. Using a new function to plot test and validation curve in one place instead of using the scores. The experiment is to see if I can get a better fit than last week's training.
1.c. Experiment Objective	The objective for this week is two folds: - Model Stacking - Understand the impact of columns on the prediction using Variable Importance by Permutation and Partial Dependence Plot

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

No New data preparation step was done for this week's assignment. It is more to explore the behavior of the data on the model prepared.

----- the section below this line is from the previous experiment -----

The CSV was loaded into a data frame without any issues. There are 8000 rows and 22 columns. Checked the following to see if there were any anomalies in the data

- duplicate check using `duplicated()`
- descriptive statistics to check data issues using `describe()`
- datatype check using `info()`
- null value checks using `isna().sum()`

Looking at descriptive status found negative values in the following columns.

- 'GP' , '3P Made', '3PA', '3P%', 'FT%' , 'BLK'

Investigating further I found that the percentages of these columns are incorrect. FG%, 3P%, FT% were not representative of the actual data. Also, there were 0 and negative values in '3P Made', '3PA', which would result in incorrect (or inf) values when calculating percentages.

To fix these issues, I applied the following formula

- Update all negative column values as zero
- If values of FGM, FGA, 3P Made, 3PA, FTM, FTA is zero then update all the column values as zero, the **new percentage column is also set as zero**
- Set the new percentage column for each category of data for all non-zero records using the formula.

CALCFG% = FGM/ FGA*100
CALC3P% = 3P Made/3PA*100
CALCFT% = FTM/ FTA*100

- ~~Create a new target column called **'Target_5Yrs_Inv'** to store the inverse of **'Target_5Yrs'**. I made this change as I was not able to tell if the model is doing well based on the RECALL and PRECISON scores.~~

Stats of Rows & Features used for the experiment
Number of Rows: 8000
Number of Features: 22

2.b. Feature Engineering	Used Standard Scaling on the feature columns before training the logistic regression model.
2.c. Modelling	Used multiple models - Random Forest - XGBoost - Logistic Regression - GMM

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

A score comparison between Assignments 1, 2 & 3 is given in the table below.

For Assignment 3 I have only written down the score of the Stage 2 model.

Model	Accuracy	F1 Score	Precision	Recall	ROC_AUC	Kaggle Score
Baseline	0.83	0.9	0.83	1		
Assignment 1 - Random Forest + CLASS_WEIGHT						
Train	0.74	0.83	0.92	0.76	0.79	**0.69437**
Validate	0.66	0.77	0.87	0.71	0.66	
Test	0.68	0.79	0.87	0.72	0.65	
Assignment 2 - Random Forest + Down Sample						
Train	0.99	0.99	0.99	0.99	0.99	**0.70416**
Validate	0.7	0.63	0.63	0.62	0.63	
Test	0.69	0.62	0.59	0.56	0.63	
Assignment 2 - XGBoost + Down Sample						
Train	0.71	0.65	0.67	0.7	0.64	**0.69223**
Validate	0.66	0.64	0.64	0.65	0.64	
Test	0.66	0.62	0.61	0.6	0.63	
Assignment 3 – Model Stacking						
Train	0.99	0.98	0.99	1.0	0.98	**0.59892**
Validate	0.65	0.82	0.90	0.96	0.84	
Test	0.64	0.81	0.89	0.95	0.84	

The models as part of Stage 1 I trained many models, but only picked the ones that were good fits. However, I stopped Stage 2 without tunning because while performing Variable Importance by Permutation and Partial Dependence Plot found a lot of gaps in terms of potential new feature columns that can be created/tuned further. I think I will investigate the subject matter first before training the next set of models.

The last page has all the scores for all the models trained for this week’s assignment. This time around, I think I manage to tune models better than last time.

3.b. Business Impact

For now, with all the models I have trained so far, none of them perform well enough to provide consistent prediction with enough justifications around the features selected and hyperparameter chosen.

3.c. Encountered Issues

Many potential subject matter-related fine-tuning that can be done which would require further investigation. A few examples that I understood are:

- Understand the impact of rebounds on the data
- Calculate the Rebound Rate, Offensive Rebound Rate, Defensive Rebound Rate
- Understand the impact of Minutes played in the game with other stats. The plot_partial_dependence showed that there is a high probability of players who played less time to survive 5 years in the NBA. Which does not seem to be correct.
- The same goes for FGA (Field Goals Attempts), less and high value shows the chance of survival higher, which does not make sense.

For now, all my questions are related to the subject matter are unresolved. Tunning the model will not help unless I get the right input data for 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

This assignment has raised more questions than it solved. And due to lack of time, I was not able to investigate further into the data and its impact on the models. This leads to my conclusion for this assignment.

Performing Variable Importance by Permutation and Partial Dependence Plot is a very important step!!

As soon as one model is trained, the key is to get insights into the model, the goodness of fit, and the impact of the columns on the performance of the model.

More model does not solve any problem (hard lesson learned this week).

4.b. Suggestions / Recommendations

- Train a model which is a good fit, investigate the impact of the feature columns by implementing Variable Importance by Permutation and Partial Dependence.
- Investigate and see if any improvements can be done to the data by feature engineering new columns to assist the model performance.

Model Scores for Stage 1

Jupyter Notebook	Description	Models	Run Type	AUROC	Accuracy	F1	Recall	Precision
P_Sampath_Week03-01-XG-model-train.ipynb	XGBoost	Model 1	Train	0.500	0.834	0.909	1.000	0.834
			Validation	0.500	0.834	0.909	1.000	0.834
			Test	0.500	0.834	0.909	1.000	0.834
P_Sampath_Week03-02-XG-model-train.ipynb	Random Forest, invesrse the Target	Model 1	Validation	0.685	0.834	0.000	0.000	0.000
			Train	0.711	0.834	0.000	0.000	0.000
			Test	0.700	0.834	0.000	0.000	0.000
P_Sampath_Week03-03-XG-model-train.ipynb	XGBoost, with Data Upsampled	Model 1	Train	0.688	0.638	0.636	0.633	0.640
			Validation	0.686	0.640	0.642	0.648	0.637
			Test	0.677	0.629	0.623	0.612	0.633
	XGBoost, with Data Upsampled + Hyperopt	Model 2	Train	0.500	0.500	0.000	0.000	0.000
			Validation	0.500	0.500	0.000	0.000	0.000
			Test	Not Tested				
P_Sampath_Week03-04-XG-model-train.ipynb	XGBoost, with Data Downsampled	Model 1	Train	0.686	0.641	0.642	0.645	0.639
			Validation	0.683	0.638	0.656	0.690	0.626
			Test	0.668	0.612	0.610	0.607	0.614
	XGBoost, with Data Downsampled + Hyperopt	Model 2	Train	0.500	0.500	0.666	1.000	0.500
			Validation	0.500	0.500	0.667	1.000	0.500
			Test	0.500	0.501	0.668	1.000	0.501
P_Sampath_Week03-05-RF-model-train.ipynb	Random Forest, with Data Downsampled	Model 1	Train	0.679	0.634	0.636	0.639	0.633
			Validation	0.671	0.624	0.636	0.657	0.617
			Test	0.669	0.640	0.650	0.667	0.633
	Random Forest, with Data Downsampled + Hyperopt	Model 2	Train	0.677	0.622	0.618	0.612	0.624
			Validation	0.679	0.622	0.638	0.667	0.612
			Test	0.655	0.598	0.596	0.592	0.601
P_Sampath_Week03-06-RF-model-train.ipynb	Random Forest, with Data Upsampled	Model 1	Train	0.676	0.620	0.618	0.616	0.620
			Validation	0.678	0.628	0.627	0.625	0.629
			Test	0.659	0.614	0.606	0.594	0.618
	Random Forest, with Data Downsampled + Hyperopt	Model 2	Train	0.674	0.614	0.613	0.610	0.615
			Validation	0.677	0.621	0.621	0.620	0.621
			Test	0.658	0.609	0.602	0.592	0.613
P_Sampath_Week03-07-LoReg-train.ipynb	Logistic Regression	Model 1	Train	0.707	0.644	0.751	0.642	0.903
			Validation	0.715	0.641	0.750	0.644	0.897
			Test	0.695	0.645	0.752	0.644	0.902
		Model 2	Train	0.675	0.696	0.802	0.737	0.879
			Validation	0.698	0.699	0.803	0.737	0.883
			Test	0.684	0.697	0.801	0.733	0.883

Note: Model Score selected for Stage 1 is marked in **bold.**