EXPERIMENT REPORT

Student Name	Rui Hu Zhang (13627753)
Project Name	36114 Assignment 1 week 2 Team 3
Date	20/02/2022
Deliverables	**Jupyter notebook**: Zhang_RuiHu-13627753-week2_RandomForestCV20FoldsGridSearch.ipynb **Model name**: randforest_10cv_gridsearch_best.joblib **Github repo**: JKaur1992/adsi_at1 **Repo branch**: RZ_Week2_codes **Kaggle submission**: Richard_RandFor_CV_GridSearch_week2_20220220.csv

---

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 goal of this project is to create a machine learning (ML) model that predicts likelihood whether a new NBA player will be staying in NBA after 5 years, based on their in-game performance metrics. This allows the user of this ML model the possibility of making long term investment in a new player with the individual's in-game performance, and minimise making investment on short term players.
1.b. Hypothesis	The 19 player performance variables are sufficient in making accurate prediction in the likelihood of a new player continue after 5 years. The interaction effect between these performance variables are not influencing this prediction significantly, and the higher order effects are also not influencing this prediction. These variables are descriptive of the performance of a player, hence these variables are sufficient in describing the player's skills. Therefore no other effects are necessary in describing player skills.

1.c. Experiment Objective	A ML model is created using the 19 independent variables to predict the likelihood of a new player staying in NBA after 5 years.

---

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	The variable column Id is removed from the dataframe prior to model training. This is only a descriptive variable, and is not a variable that describes the player performance. Therefore this is not included in the dataframe of the model training. For the training dataset (training.csv), the target variable column TARGET_5Yrs is also removed prior to model training to avoid incorrect or accidental inclusion of these data values by the ML models. The remaining variable values are scaled to one standard deviation about the mean, using the Scikit learn method StandardScaler. This ensures the variations in the variable values are not influencing the parameters predicted by the model. For the polynomial features, the Scikit learn method PolynomialFeatures is used to generate second order variables. All interaction variables are generated. This ensures the ML model considers For train-test split, given the target variable TARGET_5Yrs shows some imbalance between 0 and 1 classes, a stratified selection is used. In addition to the above preparations, this week examines dataset balancing. The simple upsampling is used instead of the stratified dataset splitting used in the last week. In other words, this week considers two data preparation strategies: 1. Without dataset balancing, perform stratified dataset splitting on the dataset. 2. Perform upsampling dataset balancing, then perform dataset dataset splitting.

2.b. Feature Engineering	This week the second order and interaction variables are all dropped. Based on last week's results, ML models using the second order and interaction variables do not show much improvement over models using only first order variables. This week the cross-validation study is reduced to 10-fold, due to the number of grid search studies with relatively large grid. In addition to cross-validation, a simple grid search study is performed for all three ML models and algorithms outlined in Section 2.c below. This grid search is performed using standard variable value variations recommended in reference books and webpages. This study incorporates cross-validation in the model training. In a separate investigation, the data balance is considered. In this week, the data imbalance is addressed by simple upsampling using the Scikit learn method resample. This brings the 1331 counts of the minority class is brought to the same count of 6669 count of the majority class. The AUC score is the only metric used in the cross-validation and grid search experiments.
2.c. Modelling	Logistic regression – simplest linear model for classification ML model, given the small number of parameters and hyperparameters, this will also be fastest to train Stochastic gradient descent (SGD) – from reference books and Internet websites, this is one of the methods for shallow ML models that allows faster approach Random forest – another shallow ML model that is relatively robust for a general situation. This will also allow the consideration of variable interaction. Given there are a number of hyperparameters, this ML model has a large scope for optimisation. The following hyperparameters are varied in the grid search appraoch. Logistic regression: regularisation parameter, and algorithms used by Scikit learn SGD: loss function type, multiplier of regularisation term, regularisation parameter ratio, and learning rate algorithm Random forest: number of estimators, maximum depth of trees, and minimum samples per split

---

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

Tabulated below is the AUC score for all models trained this week.

Original Imbalanced Dataset	
Logistic regression	0.7082
Random Forest	0.7235
SGD	0.7023
Upsampled Dataset	
Logistic regression	0.7054
Random Forest	0.7267
SGD	0.6674

	From these results and the experiments done up to now, the model performances are expected. Specifically there is no ML model or hyperparameters that make a significant difference to the model performance. As the best ML model generated for this week has an AUC score of 0.7267, this is used for the assignment submission this week.
3.b. Business Impact	The highest AUC score of these models is 0.7267. This corresponds to a logistic random forest model with linear variables, cross-validation, and upsampled dataset. This suggests that without any further models being generated, this model has moderate accuracy, as the AUC is between 0.5 and 1.0.
3.c. Encountered Issues	For SGD models, there are a number of hyperparameter combinations where the model training process does not converge. These are filtered by the grid search algorithm. This does mean that the hyperparameter search is needed to ensure the model training can complete. Given the effort spent on the various algorithms this week, no improvement in the model performance is observed. This is despite the application of cross-validation and hyperparameter optimisation. This suggests that other ML models and algorithms should be attempted.

---

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	As shown in Section 3.a the experiments performed in this week generated ML models that have relatively weak prediction performance. Given the cross-validation and grid search experiments done in this week, the next step is to try different ML models. For example, XGBoost could be considered.
4.b. Suggestions / Recommendations	With both grid search and cross-validation codes are tested and written, these can be used for experiments on other models. The ML models tested up to this point have their hyperparameters optimised, therefore, new ML models can now be tested to see if they can form better predictions.