EXPERIMENT REPORT

Student Name	Rui Hu Zhang (13627753)
Project Name	36114 Assignment 1 week 1 Team 3
Date	13/02/2022
Deliverables	**Jupyter notebook**: Zhang_RuiHu-13627753-week1_LogisticRegressionCV20Folds.ipynb **Model name**: logreg_cv20folds_default.joblib **Kaggle submission**: Richard_LogRegression_CV_corrected_week1_20220213.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.
2.b. Feature Engineering	This week the feature engineering is the generation of second order variables as well as interaction variables. No feature selection is made this week. For the linear variables only, a 20-fold cross-validation is made to the three ML models with default hyperparameter values. This is used to test the code for cross-validation and check that the model predictions made without the cross-validation returns similar predictions.

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. All three models are trained using the default parameter values of Scikit learn. This week's focus is on getting the basic experiment procedure and associated code to work, hyperparameter searching is not performed this week. For the SGD models, given there are a number of models that can be applied, with a number of hyperparameters, these can be checked in the coming weeks. Similarly the random forest model has a number of hyperparameters to be checked using, say, a grid search method in the coming week to obtain the optimal model.

---

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.

	Linear variables only, no cross-validation	Linear variables only, 20 fold cross-validation, best model	Second order variables, no cross-validation
Logistic regression	0.707122003	0.747613285	0.70553539
SGD	0.702266348	0.714071856	0.577236758
Random forest	0.666801468	0.736753765	0.662810982

The logistic regression models can be seen to be the best model type out of the three model types tried. This is expected as this is the simplest model type out of the three. For both SGD and random forest models, the choice of hyperparameter values are expected to influence the accuracy of the models.

3.b. Business Impact

For most models, the AUC score of these models are higher than 0.5. Hence these models provide moderate prediction for the new player staying in NBA.

The highest AUC score of these models is 0.7476. This corresponds to a logistic regression model with linear variables and cross-validation. 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	As the code representing basics of the experimental procedure (e.g. data splitting, ML model training, etc.) are confirmed to work, the next step is to start the hyperparameter optimisation process to find the best set of hyperparameters for each model. From the Section 3.a, the two advanced models (SGD and random forest) no optimisation of hyperparameter values are made this week. This will be examine in the coming weeks.

---

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. Therefore the next step is to perform model tuning experiments to improve the predictive performance of ML models. With the relatively few models generated at this stage there is no telling in whether a ML model should be discarded.
4.b. Suggestions / Recommendations	Even without trying other ML models, the hyperparameter tuning should improve the accuracy of these ML models. This can be made with cross-validation. With the more complex models (e.g. random forest), the number of folds used in the cross-validation during the hyperparameter optimisation stage can be decreased to, say, 5 or 10, to decrease the time per iteration. This then can speed up the cycle of experiments, and lead to further hypothesis testings and experiments being conducted.