1A EXPERIMENT REPORT

Student Name	AJ Duncanson 12823819
Project Name	[UTS AdvDSI] NBA Career Prediction
Date	7 February 2021
Deliverables	/notebooks/ duncanson_aj-12823819-week1_logreg03f.ipynb /models/ aj_logreg03f.joblib Aj_logreg03f_scaler.joblib /src/ All modelling is via Notebooks at this stage, although subfolders do contain function modules. Github repo https://github.com/adv-dsi-group4/nba-career-prediction

---

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?* The aim is to predict which players are likely to still be playing after 5 years. The team will be able to focus our investment on those players who can be expected to last the distance. If our model results in a false positive, then we risk investing in players who will not be a good long term return. If our model results in a false negative, then we risk under-investing in players with potential. False positives will lead to worse financial outcomes than false negatives.

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.* We hypothesised that a Logistic Regression model can be successfully applied to the problem, and after running a series of experiments to test the effect of different approaches the current hypothesis is that a combination of features selection, class balancing and cross-validation will improve upon the best result so far.
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.* The previous best results were an Area Under the ROC Curve of 0.71007 on the test set in Kaggle. The aim of the experiment is to get a higher AUC value.

---

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* • Data was examined to ensure all features in the training & test sets were the same, to ensure that there's been no error in data loading. • Verify that the target variable is a binary classifier, to inform the types of models we might employ. • Confirmed no null values, and hence no need to exclude samples or impute values. • Some variables contain negative entries. It is unclear what this can mean, since the values are supposed to represent game statistics. Either some of the statistics are 'net' values or there are data quality issues, and some basketball subject matter expertise is needed. Since this occurs in both the training and the test data I've not made adjustments at this stage. It does need to be investigated however.

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* Because this is a regression exercise, variables of different scales will distort the modelling. To prevent this, a scaler algorithm was applied to transform the data into the same scale. Most of my experiments so far have included all features available in the data, only excluding the ID and Old ID columns from the data. The most successful experiment, 'logreg03f', excluded 8 variables from the modelling to examine whether this might result in a more robust model. ● In a regression, variables that are strongly correlated with one another can sometimes create problems with individual model coefficients and so a pairwise correlation plot was constructed to identify variable correlations. ● It was found that several groups of variables were related: ○ Minutes played, points per game, and several aggregate stats. ○ 3pt shots made and attempted ○ Free throws made and attempted, and also total points per game ○ Rebounds, and the offensive and defensive rebounds. ● In each group, one one variable was retained and the others dropped.
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.* My experiments focussed on Logistic Regression models and Random Forest models. At this early stage in the project, these were arbitrary choices but agreed with my team so they could attempt something different. Both are relatively simple concepts to explain to stakeholders although Logistic Regression will deliver more easily explainable results. My initial Logistic Regression model fitted much better to positive cases than negative cases. This suggested the next iteration should include class weights, and this did indeed provide *much better* results overall. As described above, some potential features were dropped from the data and this gave a *marginal* improvement. I also trialled different regularisation penalties, both L2 and L1 and an ElasticNet mix, but no tangible difference was detected in any of the results. Perhaps this makes intuitive sense if I've already weeded out predictors with low additional predictive power. I also tried two different approaches to validation. Initially I split the training data 80/20 to produce a validation set, and compared results on that before deciding if a model was worth applying to the test set. Another *minor* improvement was made by instead

	doing a 5-fold cross-validation. The Notebook only deals with the experiment that gave the best result so far, which was a Logistic Regression, with some potential features excluded, class weights, L2 penalty and 5-fold cross-validation. My initial Random Forest model gave almost perfect fit to the training data and a poor fit to validation data. I applied a grid-search & cross-validation approach to the Random Forest model to attempt to find parameters that would not overfit. This focussed on limiting the depth, and increasing the minimum number of samples per leaf. However, *none* of the results were superior to the results from the Logistic Regression model. Random Forest code is not contained in this Notebook.

---

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.* Model logreg03f resulted in: ● On the training data: ○ Area under ROC curve = 0.70 ○ Precision 90% ○ Recall 64% ● On the test data: ○ Area under ROC curve = 0.71032
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)* A Precision score of 90% means that if the model predicts a player will last 5 years, then we should expect it to be right in 90% of cases. A Recall score of 64% means that of all the players who will go on to 5 years, the model will get 64% of them right. In business terms, this means that if we invest in players according to the model then ● The vast majority of the players we invest in will last 5 years or more. ● We will capture nearly two-thirds of the available pool of players with 5-year potential. They may exist a better model that captures more of the available pool, but it would not be preferable if it is at the cost of the strong 90% precision we have achieved here.

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.* Issues ● Negative values ○ As mentioned earlier, some player stats include negative values and we need to identify if that makes sense or if it is a data quality issue we need to act upon.

---

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 is a reasonable model, and could certainly be applied to the business problem. I think we've learned that this is about as good a model as we will get from a logistic regression approach, since a number of variations have not yielded anything better. We know that class weights make a bigger difference to the result than any of the adjustments to features or use of differing regularisation.
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.* I've concluded that this is as good as a Logistic Regression will get, without further data transformations. We also know, from other team members' work, that polynomial transformation can improve the results of the regression. Therefore it would seem that a good next step would be to apply the class weights in a range of polynomial transformations, using a cross-validation approach. In parallel, we also still have time to explore some other algorithms, such as KNN, SVM, XGBoost, Neural Nets.