1C EXPERIMENT REPORT

Student Name	AJ Duncanson 12823819
Project Name	[UTS AdvDSI] NBA Career Prediction
Date	21 February 2021
Deliverables	/notebooks/ duncanson_aj-12823819-week3_rforest05a.ipynb /models/ aj_rforest05a.joblib /src/ All modelling is via Notebooks, 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.

The team has produced good results forms several algorithms, including a logistic regression, a polynomial logistic regression, an SVM classifier.

We have not been able to improve on these results using an XGBoost Classifier.

Therefore, this experiment is to test the Hypothesis that a Random Forest algorithm might deliver better results. It is hoped that it might outperform the logistic regressions by avoiding the linearity assumption, and that it might be easier to tune than the XGBoost has proven to be.

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 Area Under the ROC Curve on the test set in Kaggle was 0.72159, obtained by a polynomial logistic regression with feature selection.

The aim of the experiment is to get a higher AUC value.

It is unclear before commencing whether this will be achieved, or what level of outperformance we might expect. Given that many prior experiments have all landed in a band of AUC values between 0.7 and 0.712, it seems likely that any outperformance will be small.

---

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

As in the previous week's experiments:

- Data was examined to ensure training & test sets contained the same features to ensure that there's been no error in data loading.
- It was verified that the target variable is a binary classifier, to inform the types of models we might employ.
- We confirmed no null values, and hence no need to exclude samples or impute values.
- Some variables contain negative entries. These have been set to the absolute value this week; despite the slightly better result by keeping them negative last week, it really doesn't make any practical sense.

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

This experiment consisted of three tuned models:

(05) - trained on the full initial feature set available in the data.

(05a) - trained on the full initial feature set with the addition of an extra feature, generated by clustering algorithm as described in the next section. It was hoped that applying a new approach might give more information to the main algorithm and balance out some of its errors. **This is the Notebook I have submitted.**

(05b) - trained on the full initial feature set plus the cluster feature, but with feature 'GP' removed as described in the next section. After viewing a sample of LIME plots, a new hypothesis was formed that this feature might be contributing to many false negatives.

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 initial Random Forest model gave almost perfect fit to the training data and a poor fit to validation data, suggesting significant overfitting.

First tuned model (05) - initial feature set

I used both manual search and a HyperOpt search space to tune hyperparameters, focussing on those that might reduce the overfitting: the maximum depth, the number of estimators and the minimum number of samples per node.

I used the class-balancing weights built into the sklearn RandomForestClassifier class, to manage the imbalance between positive and negative observations. Stratified k-fold cross validation was also used to avoid overfitting and to manage the class imbalance issue.

Second tuned model (05a) - with the addition of clusters

To generate a Cluster feature, I used a Gaussian Mixture algorithm. I performed a search for the optimum number of components using the Bayesian Information Criterion, trained the model on the training data and then predicted the cluster for each observation in each of the training, validation and test data. Then the Random Forest process above was repeated from the beginning with the new extended datasets.

It was observed that the proportions of positive and negative classes were quite different for different clusters - ranging on the training data from 7% negative in one cluster up to 46% in another - giving rise to hope that they might help the model along.

After running 05a, I ran a number of inspections of the results using Importance by Permutation and Partial Dependence Plots to examine which features make the biggest impact on results, and how it responds to different values of those features.

The biggest impacts:
FG%: 0.00742
FT%: 0.00427
FGA: 0.00265
3P%: 0.00195
GP: 0.00177
3PA: 0.00117
TOV: 0.00100
3P Made: 0.00093

All other features reported negligible impact (or negative impact), indicating that they may not be contributing anything useful to the model.

I also ran a LIME analysis on a small sample of observations from the validation data that were being falsely reported as negatives. In each of the 3 sampled observations it was reported that feature 'GP' was the largest contributor to the incorrect classification. There's nothing to say that the samples are in any way representative of a large group, but it seems worth an experiment to see what happens if I leave out just that one feature.

Third tuned model (05b) - removal of features with low permutation importance

Model (05b) tested the above hypothesis by excluding the 'low permutation importance' data from the feature set before following exactly the same process as 05a.

Fourth model (05c) - removal of feature 'GP' to test hypothesis about LIME analysis

Model (05c) tested the 'GP' hypothesis by excluding it from the feature set. I decided to include 'GP' in the clustering model, as that will allow a much weaker 'GP' effect in the model. Otherwise, the modelling followed exactly the same process as 05a.

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.

Models resulted in the following Area under ROC curve:

(05) - initial feature set

- Training data 0.75
- Validation data 0.70
- Test data 0.70564

(05a) - with the addition of clusters

- Training data 0.75
- Validation data 0.70
- Test data 0.70797

Finding: the addition of the cluster feature appears to have made only a very small improvement.

(05b) - only including features with permutation importance > 0 in (05a)

- Training data 0.75
- Validation data 0.70
- Test data 0.70475

Finding: the removal of the features with low permutation importance has resulted in worse performance.

(05c) - like 05a but remove feature 'GP' after the clustering is complete.

- Training data 0.73
- Validation data 0.69
- Test data 0.67767

Finding: definitely not! Clearly, even though 'GP' might be contributing to some false negatives, taking it out also has other unwanted effects on other observations.

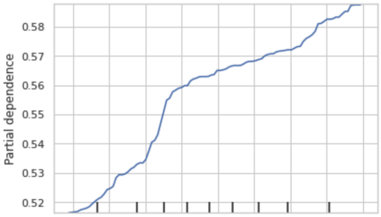
Best model

- (05a)
- Precision 90% (validation set)
- Recall 67% (validation set)

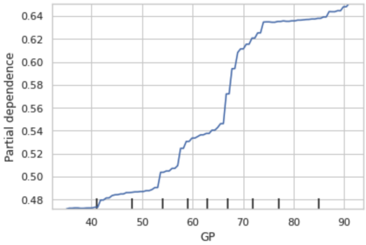
Partial dependence plots (05a) - there are several interesting observations from the plots of the key features.

The clearest relationships:

- FG% (Field Goal %) where there is a strong contribution to the probability of a 5 year career as the percentage increases, with a particular step change at around 41%.

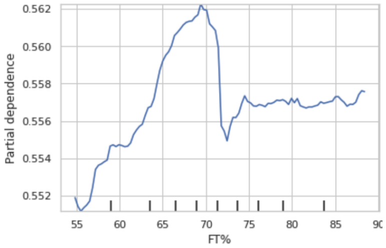


- GP (Games Played) where there is also a strong contribution to the probability of a 5 year career as the value increases, increasing even more after around 66 games and possibly plateauing after 70 games.

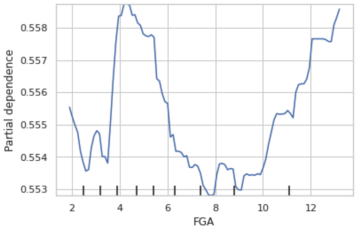


Curious relationships:

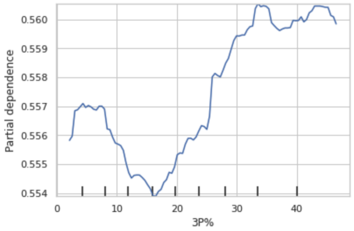
- As FT% (Free Throw %) increases towards 70%, there's a greater contribution to probability of a positive. Over 70% though, it drops back to a lower plateau level. It's not clear what to make of this. It suggests that you are more likely to last as a player if your free throw % is average, than if it is excellent.



- FGA (field goals attempted) is all over the place. It suggests that an average of 4 or 5 is a good sign, relative to anything more or less than that. It is possible that the increase in the curve after an average of 10 is not to be trusted as it is based on limited data.



- 3P% (3 point %) is almost the inverse of the FGA. If you have an average over 25 it's a great indicator. But it is also possible that having a very low % is better than being in the 10-20% range. Possibly this relates to good players who specialise in other parts of the game, eg defense?

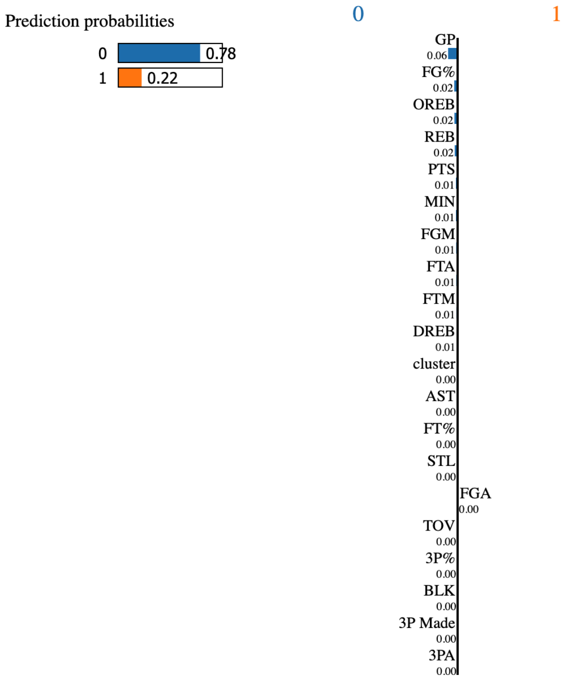


LIME - Local interpretable model-agnostic explanations

- This is a sample of LIME output for a false negative observation with a very low predicted probability.

  It suggests that pretty much every feature is suggesting it's a negative, even though it is in truth a positive. And that no one feature has a very large impact - making it very difficult to identify any potential remedy.

  The experiment in 05c showed that removing the 'GP' feature made the model worse overall. And little difference to these false negative cases.



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 67% means that of all the players who will go on to 5 years, the model will get 67% 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 two-thirds of the available pool of players with 5-year potential.***

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

	• In terms of teamwork, we've had less communication between us this week. One outcome of this is that we've not been working in identical environments, resulting in some shared code modules not working for everyone. • A flow-on effect was that this week I ran my experiment entirely in the notebook with no new modules created. This has made the notebook rather long and prone to error when edited. Moving some of the functionality out into a code module, and making better use of pipelines, should avoid repeating code and de-risk the process.

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.* I've learned: • There's not a huge amount of difference in performance between the different algorithms that the team has applied. • I now have a better understanding of how to manage hyperparameter tuning and validation,but all the extra effort hasn't delivered a better model! • Although it was interesting building the Gaussian Mixture clusters, it turned out that they contributed little to the performance. An additional analysis that might be worthwhile is to examine the • Even though the Permutation Importance values in 05a suggested that a number of features were not contributing to the model performance, model 05b showed that when you remove them all, the model is not as good. My conclusion here is that, although removing 1 feature may make no difference, interactions between two or more features can contribute a (in this case, small) performance boost. • It's a similar story with the 'GP' feature in 05c. It may well be indicated by the LIME analysis but it's impact on those cases is small and elsewhere it is making a positive contribution to model performance. It's better left in.
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.* A next step might be to take the 1 thing that's made the most difference in all the work the team has done - the polynomial transformation - and see if that, or some other data transformation, makes any difference to any of the models we've tried. It seems unlikely though, as any tree-based algorithm is not necessarily going to care about the scale of a variable in the way that a linear model does.

	Or perhaps, since the team does have a model that can be used in practice, we might conclude that more business benefit is to be gained by implementing what we have rather than trying to finesse it any further!