EXPERIMENT REPORT

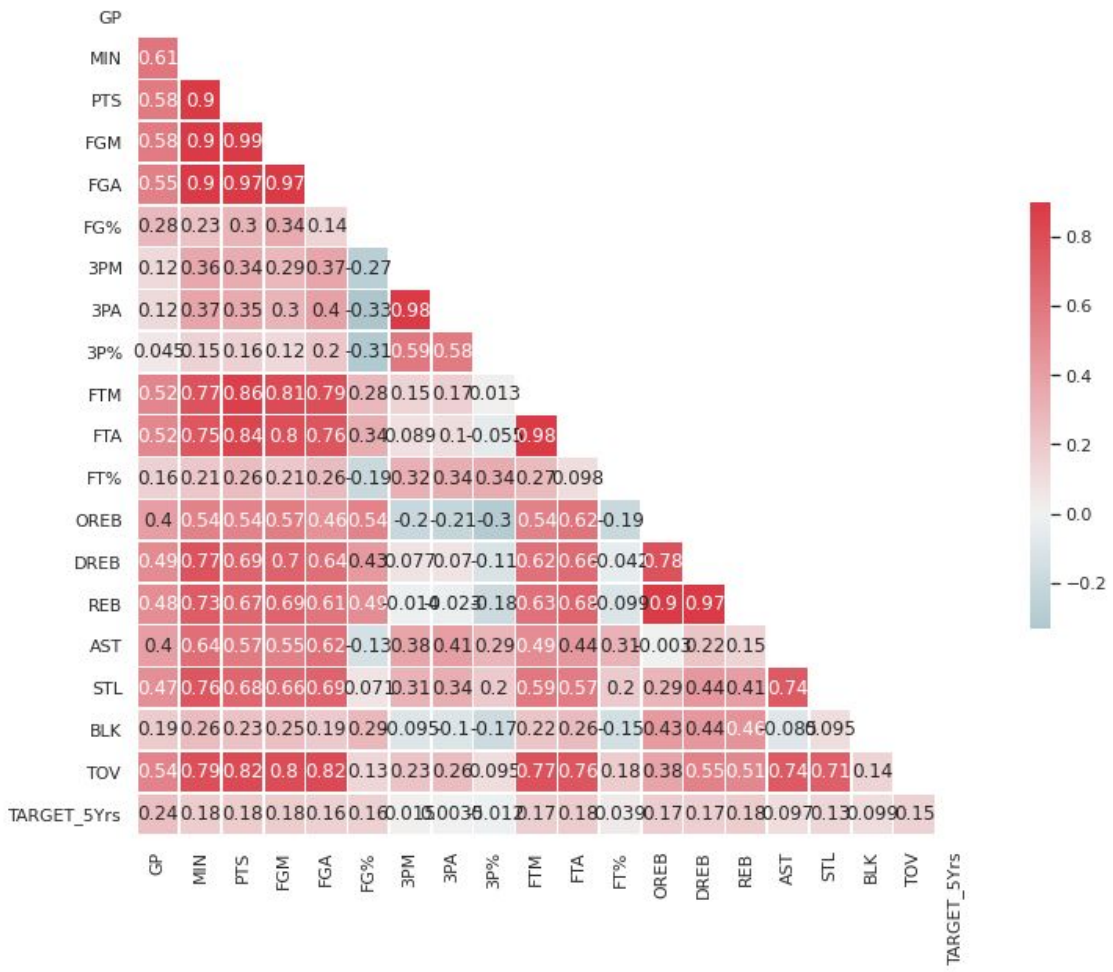
Student Name	Ron Au
Project Name	NBA Career Prediction - Week 1
Date	2021-02-07
Deliverables	au_ron-week1_logistic_regression.ipynb https://github.com/ronvoluted/kaggle-nba

---

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	Using a dataset of 8,000 NBA rookies and 19 features, train a model to predict their probability of having an NBA career longer than 5 years. The application of this problem is a Kaggle competition where a list of 3,799 unlabelled samples need to have their >5Yr probabilities predicted and recorded on a range from 0 to 1. The objective is to use the competitive environment to practice improving our data science methodology and assessing how different techniques are more, or less relevant to specific training problems. By operating in teams, it will also promote the use of good code sharing practices.
1.b. Hypothesis	Given that the target label is 0 or 1, this is a classification problem. Other than logistic regression, linear algorithms may not be appropriate. It is important to first get results and then spend time iterating. As such there is no hypothesis of actual scoring, but improvements are expected with further data preparation, feature engineering, model selection and hyperparameter tuning.
1.c. Experiment Objective	Since I am relatively new to data science and especially the SciPy stack, I am hoping to practice good code sharing methodology and machine learning process, but expect average success with actual model accuracy. For this Kaggle competition I will be satisfied with an AUROC > 0.6. If my submissions score less than, or close to the randomly generated base predictions then I will definitely have failed somewhere. Regardless of prediction scoring, I intend to improve my methodology over the following weeks.

---

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 Firstly I separated the 'TARGET_5Yrs' column as that was the target label. Secondly I removed the 'Id_old' and 'Id' columns as they were not features of the dataset. I noticed many rows had negative values which, given the domain knowledge of it being impossible to have negative games played, meant that the values were not useful. At first I removed all columns but this ended up being nearly half the dataset! With such a large cull, it would be a waste not to use the other usable values in these rows, and by reducing such a large amount it could also increase the chance of overfitting. In the end I decided not to remove these samples but I will revisit again. Another option, used by my teammate, was to absolute the values, but I thought it was more important to get some results first. Some columns representing percentages also seemed to not match the columns they should have represented. I used my teammate's 'recalculate_percentage' module to test, but did not seem to have a greater score so decided not to perform this operation in the end. I did run the dataset through a StandardScaler though so that relationships between features could be mapped more efficiently. I also split the data set into training, validation and testing sets.
2.b. Feature Engineering	I visualised a Pearson Plot to see if there were any high multicolinear columns and saw that the 'FGA' and 'MIN' columns showed particularly high colinearity with other columns:



I tested dropping these columns but doing so produced lower scores, so I chose not to continue with it.

2.c. Modelling

Since it was a classification problem, I didn't use true linear models. I chose logistic regression as it is a linear model that performs like a classifier.

I needed to increase the max iterations for the given dataset.

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	**Local** Accuracy: 0.837109375 F1: 0.9107257546563905 Recall: 0.9955534753100866 R2: 0.837109375 AUROC: 0.7097380631410375 **Kaggle** AUROC: 0.70709
3.b. Business Impact	This ranked ~4.5 out of 6 teams' highest scores. It was a great first attempt higher than I was expecting but there is much more tuning/experimenting I can do.
3.c. Encountered Issues	Setting up the Dockerfile and image took a lot of research, but I was able to replace the 3.5GB jupyter/scipy Docker image with a 1.5GB custom image.

---

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	Machine learning requires much experimentation and reasoning of data. The effects of data preparing, feature engineering, models selection and hyperparameter turning can be subtle but logical.
4.b. Suggestions / Recommendations	I need to continue experimenting with a wider range of methods.