EXPERIMENT REPORT

Student Name	Ron Au
Project Name	NBA Career Prediction - Week 2
Date	2021-02-14
Deliverables	au_ron-week2_logistic_regression.ipynb https://github.com/ronvoluted/kaggle-nba

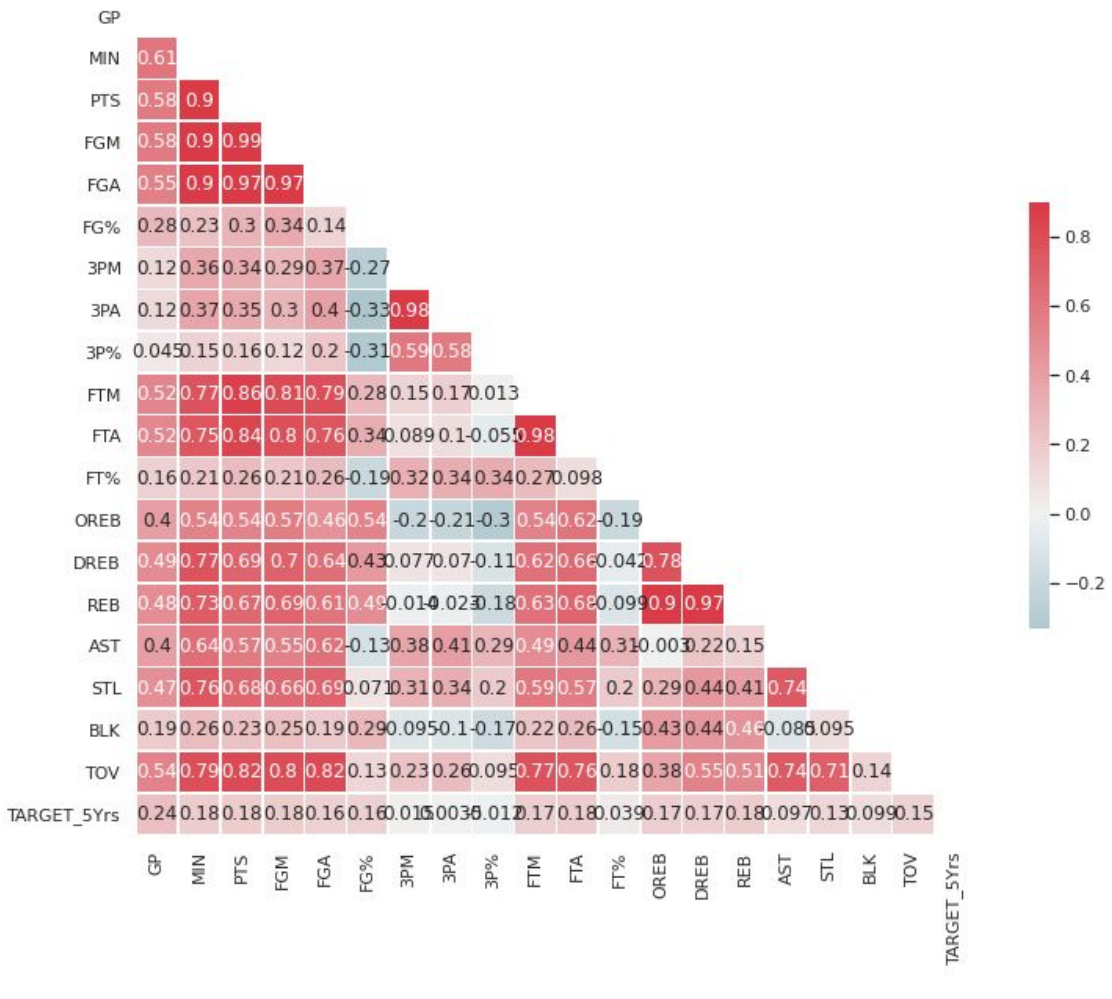
---

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	By employing visualisations, insights should be gained that can better inform choices in methodology, e.g. which algorithms may be more effective and which features have high multicollinearity. If we apply these insights, performance of the model should improve compared to other algorithms/features with default parameters. If we compare performance of models with default parameters, it would technically be possible to make a hypothesis of which models will perform better, given the appearance of data points in the visualisation. Of course, with zero hyperparameter tuning these results are expected to be inconclusive at first. The aim then is to tune these models with hyperopt and re-training with best values. This should more accurately present each model's viability for the data set and allow for meaningful comparison. Given that the objective is a classification problem, the expected end result is that the only linear model that would perform well is logistic regression. Tuning a model with Hyperopt should achieve better performance than manually testing via trial-and-error.

1.c. Experiment Objective	Prior experiment did not test hypotheses in-depth or establish learnings based on model results, as the objective was only to gain a working first experiment. It is intended for this experiment to take a more scientific approach, being driven by hypotheses and tested results. Its measurable objective is to determine the following: a) Can visualisations assist with choices in methodology? (data, features, algorithms) b) How do results compare between choices based on visualisation insights? c) How do results compare between default hyperparameter values and those tuned by Hyperopt?

---

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	CSV was read into dataframe and copied into cleaned version to preserve original input. 'TARGET_5Yrs' column separated as target label. 'Id_old' and 'Id' columns removed as they are not data features.. Many negative values noted: 5,216 values across 4,954 rows. These are certainly incorrect given our domain knowledge that Games Played and 3-Pointers-Attempted, etc cannot be less than zero. They could be dropped/imputed at this stage but were left as-is until results were generated and thus able to be compared before/after. Data also showed an imbalance of 6,669 '1' classifications and 1,331 '0' classifications, an 83% to 17% ratio. Again, this was left as-iis until results could be compared to determine if balancing the dataset would actually improve the model. Dataset standardised using StandardScaler as value ranges were quite disparate. This will also aid the accuracy of visualisations. Some models perform worse/better with standardised data however, so this should be kept in mind for future experiments. Data later split into training, validation and testing sets.
2.b. Feature Engineering	A Pearson Plot visualisation was generated to observe potential insights: https://drive.google.com/file/d/1lSIc77MEhI_NdPpJQ7MPOdlLyFKFLgyO



From the plot, it appeared that:

- There was high multicollinearity in the ‘FGA’ column
- There was high multicollinearity in the ‘MIN’ column
- There was moderate multicollinearity in the ‘FGM’ column
- There was moderate multicollinearity in the ‘PTS’ column

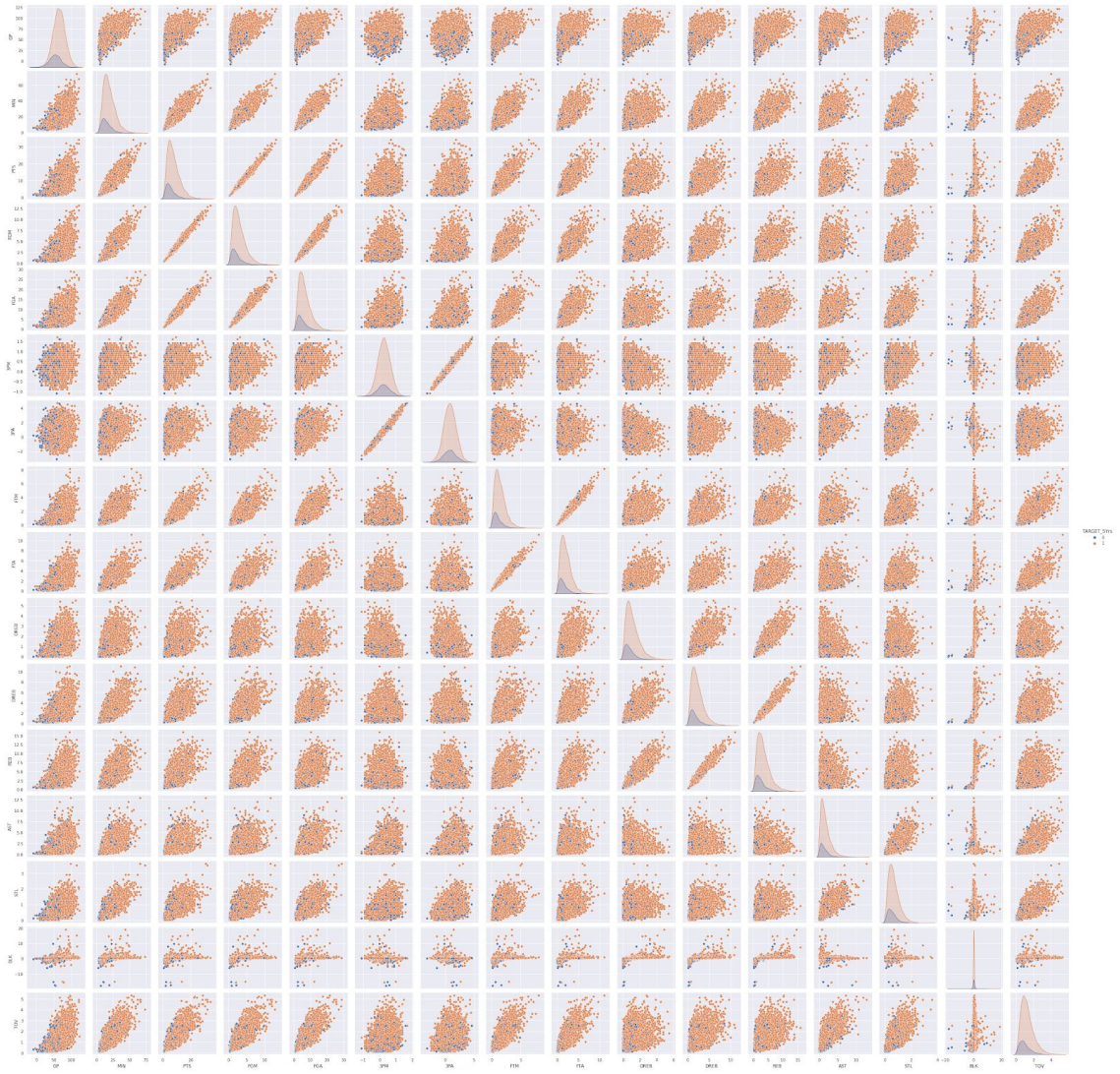
Various combinations of dropping these columns were evaluated. Results are not listed here, as no combinations resulted in higher accuracy scores.

It is possible that imputing the values or engineering new columns, instead of dropping them, may result in improved performance.

2.c. Modelling

A Seaborn pair plot visualisation was generated to observe potential insights:

https://drive.google.com/file/d/1-9EftoSFilL5ojpkoO8xhfXcQZBRMt9n



From the plot, it could be assumed that:

- A linear classification algorithm (i.e. logistic regression) may not be ideal as there is no clear separation of classifications that a line could be drawn through
- A clustering algorithm may not be ideal as there is much overlap of classifications and no clear concentrations of groups

It was now hypothesised that a trained logistic regression model or a trained kNN model would not perform as well as a trained random forest model. a random forest classifier was chosen as the first model (not including baseline predictions). This was tested against

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

First run results

A control set of results using default parameters (other than Random Forest max_depth=5 to prevent 100% overfitting) was performed:

Model	AUROC (Validation)	AUROC (Training)	Accuracy (Validation)	Accuracy (Training)
Baseline		0.5		0.723046875
Logistic Regression	0.69780407	0.70973723	0.825	0.83710938
K Nearest Neighbors	0.60183193	0.85103258	0.80078125	0.85253906
Random Forest Classifier	0.69200372	0.77177527	0.81953125	0.83984375

As expected, results are not conclusive as there is no tuning and kNN showed heavy overfitting. They do appear to suggest that simply choosing to use random forest may result in a slight increase in accuracy over logistic regression. The difference fits within even a small margin of error however.

Second run results

After testing with some manually tuned hyperparameters, the three models were optimised with hyperopt:

Model	AUROC (Validation)	AUROC (Training)	Accuracy (Validation)	Accuracy (Training)
Baseline		0.5		0.723046875
Logistic Regression	0.68944993	0.711476	0.825	0.83574219
K Nearest Neighbors	0.63933741	1.0	0.821875	0.1.0
Random Forest Classifier	0.66566852	1.0	0.81875	1.0

Unfortunately these results were evidently not meaningful for the purpose of the hypothesis. Rather than showing differences in the models' viability, they showed there

was an issue in the training or evaluation methodology.

Third run results

The number of folds in `cross_val_score` was increased from the default of 5 to `cv=10` for all models which were then retrained:

Model	AUROC (Validation)	AUROC (Training)	Accuracy (Validation)	Accuracy (Training)
Baseline		0.5		0.723046875
Logistic Regression	0.69118389	0.71135028	0.825	0.83613281
K Nearest Neighbors	0.63705826	0.72864153	0.81796875	0.83496094
Random Forest Classifier	0.67215957	0.98071662	0.8171875	0.87285156

Results were much more tangible compared to the second run but there was still heavy overfitting.

Predictions from this run were used on Kaggle for final testing of the hypothesis:

Model	AUROC
Logistic Regression (optimised)	0.70739
Random Forest Classifier (optimised)	0.69995

Hyperopt `max_evals` observations

Separate to the hypothesis, effects of different `max_evals` on accuracy were also observed:

`max_evals`	5	10	50
Logistic Regression	-0.8349609375		-0.8349609375
K Nearest Neighbors	-0.8345703125	-0.83515625	-0.835546875
Random Forest	-0.8353515625	-0.8361328125	

3.b. Business Impact	Disappointingly, experiment results did not indicate that an optimised random forest classifier is more ideal for this data set than an optimised logistic regressor, which was the hypothesised outcome. While the conclusion was gained via thorough experimentation, it seems more plausible that the results suggest issues with insufficient feature engineering or poor methodology. Gaining visualisation insights is not sufficient if methodology is naive or incorrect. However, the experiment objective of following a good scientific approach was achieved however. All choices and conclusions were hypothesis/results driven. More consideration to feature engineering It means going forward I can rely/not rely on X and Y As Methodology is bad
3.c. Encountered Issues	Models always exhibited overfitting, but experimentation with different data preparation and feature engineering choices has not yet yielded improvements. Optimising hyperparameters with a library like hyperopt is expected to produce good results, but this experiment was not able to take advantage of it properly. Despite clear insights from visualisations, the experiment was not able to take advantage of them sufficiently.

---

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	a) Can visualisations assist with choices in methodology? **Yes** - Multicollinearity and feature relationships very observable b) How do results compare between choices based on visualisation insights? **Inconclusive** - Experiment did not bear out hypothesis c) How do results compare between default hyperparameter values and those tuned by hyperopt? **Inconclusive** - Experiment did not bear out hypothesis Methodology is lacking. Simply generating optimised hyperparameters with hyperopt will not significantly improve results. A logistic regressor is quite viable for this data set. Dropping rows or columns always results in worse results. Evidently, their values should instead be imputed or new columns added, or remain unchanged.

	Increasing `max_vals` with hyperopt has exponentially diminishing returns in loss reduction.
4.b. Suggestions / Recommendations	Methodology needs to be thoroughly re-examined. Classification imbalance was assessed in this experiment but not handled. Next experiment should examine the effects of a balanced dataset on the results gained from this experiment. Explore how to utilise the multicollinearity insights gained from the Pearson plot. If data preparation/feature engineering has been explored, and methodology has improved, it may be worth running hyperopt with more evaluations and granular space values.