EXPERIMENT REPORT

Student Name	Sean Blamires
Project Name	nba_career_prediction
Date	February 13 2022
Deliverables	**Blamires_Sean_14034019_1_data clean.ipnyb** Data cleaning and Linear Regression model

---

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 objective of this task is to devise a machine learning model that can be reliably implemented to predict what factors increase the likelihood of National Basketball League Rookies extending their playing careers beyond 5 years. The outcomes of which can inform Club Managers on future player recruitment and maintenance strategies.
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. At this stage in the model's development I have run a linear regression model and compared mean squared error as the metric of fit between the training and test data. From the ADSI Lecture Notes and other reading (e.g. Fenner M *Machine Learning with Python for Everyone*. Addison Wesley), it is apparent that regression models generally underfit the data, thus regularizations are required. Hence, I expected to find my model to underfit the data.

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. My EDA on the train set data could not see clear patterns that suggest any one variable is likely to be a stronger predictor of rookeries sustaining a career in the NBA over any other. However, research has suggested that declining skill in basketball (or any complex task) is associated with decreasing opportunity to perform of practice the task. Accordingly, I might expect rookeries who spend playing grater time early in their career will have longer careers. Any players with limited playing time, due to injury, illness, poor form, etc., will not likely have careers of 5+ years. A robust model should distinguish among parameters and predict the parameters enabling rookies to have longer careers.

---

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. I firstly ran an exploratory data analysis on the training dataset using the pandas profiling package. This enabled me to visualize the data graphically to ascertain the distribution and other attributes (e.g. missing or null values) of the variables to assess the suitability of the regression and other models and to devise a hypothesis about the expected outputs of future models. Should any data appear strongly skewed left or right on a standard curve meant that some kind of transformation may need to occur. I then uploaded the data and cleaned it to be ready for application of a linear regression model.
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. Upon running EDA on the training data, and cleaning up both the training and test set data, I checked the data types and shapes to ensure the various processing steps did not significantly alter the datasets in any way. I also ensured that all of the data/variables were stored as pandas dataframe series' so the data was able to be recognized for regression modelling. The player ID was set as the index so that it was not included as a variable in the training set and induced unintended overfitting of the linear regression model.
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. I built a linear regression model to compare parameters in the training and testing datasets using the LinearRegression subunit of the package sklearn (as sklearn.linear_model). A linear regression is usually the first model to be run in a machine learning project as it gives information about the nature of the data at hand, such as linearity among variables and bias or variance in the data around the regression line. In this instance I assessed the mean squared error values of the training and test datasets to ascertain the fit between them. The variance was low but bias likely to be high so some kind of regularization parameter (α parameter, or L1/L2 hyperparameter) will be included in future models.

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. The mean squared error (MSE) was low and similar across the training and test datasets (training = 0.12816 compared with testing = 0.15225). This means the training and test sets were a close fit and the model appropriate for running additional models and making predictions. Any underperforming might be due to bias in the model, however this was somewhat minimized in the data cleaning process.
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). The model is too early in development to make any clear predictions about factors influencing the likelihood of National Basketball League Rookies extending their playing careers. A regularized regression model combined with additional ML interpretations (e.g. Naïve Bayes etc.) are required to ascertain the influential factors and predict players that might have extended careers to inform Clubs' recruitment and retainment programs.
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. As the training and testing datasets were subsets of a large dataset that was split unevenly (around 60%: 40%), regression model parameters could not be calculated on full datasets. Rather the training set had to be further partitioned (I chose the first 3799 entries) to calculate the parameters. While the model was primarily interested in MSE scores as a measure of model performance, the R^2 values were small, thus the relationships between the parameters across test and training data were not strong and might affect the reliability of future modelling.

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. It is too early to predict how well, or what kind, of machine learning algorithm/model will best serve as a predictor of how and why rookies might have longer NBA careers than others. Nevertheless, the nature of the data at hand needs to be understood to inform future modelling, and to this end the modelling performed was useful.
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. Going forward, I will run a regularized polynomial regression on the datasets to gauge their performance when certain parameters have been waited. Other models (naïve Bayes, decision trees) that use cross validation and other techniques will then be run and compared against the regularized regressions. Receiver Operating Characteristic (ROC) curves will be calculated to assess their false positive/negative rate and the area under the curve calculated to assess performance. Upon running the models and tests I will use python tools to loop out the rookies predicted to have extended NBA careers.

Github Repo: https://github.com/jaseuts/nba_career_prediction