EXPERIMENT REPORT

Student Name	April Nommesen
Project Name	NBA Career Prediction
Date	9/11/2022
Deliverables	notebook name: at-group4-week1-eda.ipynb at-group4-week1-eda.ipynb nommesen_april-week1-logreg1.ipynb model name: logreg_sklearn_subset1 GitHub repo: https://github.com/aprilgum/adv-dsi-2022-at1-grp4/tree/master/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 rookie plate==yer likely to last 5 years can be safely granted a longer contract. Any business sponsor can decide whether its worth investing on a rookie for endorsement. Based on short research, rookie players joining in the first round of NBA draft are given four-year contracts where the first two years of the contract are guaranteed with the NBA team and the third and fourth years can change. Rookie players joining in the second round of NBA draft and undrafted players can sign contracts that can be anything from one year to four years and that are either fully guaranteed or not guaranteed at all. The value of rookie players contract is tied to the salary cap and the team they will play for. A rookie player who lasts for at least five years are considered successful and a player who does not last at least five years is considered risky. Being able to have a data-based decision to decide whether a rookie is a potential success or a potential risk greatly impacts the team's budget to pay the players' salary and also impacts the team's performance.

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, For this week's experiment, I would like to begin with the question, who are the rookie players who are likely to last at least 5 years? It is worthwhile to check both sides – whether to predict who are risky (ie not last 5 years) or who are potentially successful.
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 goal for this experiment is to be familiar with the data. Possible scenarios would be like: 1) Realise whether I have enough data. If I eliminate the bad data, would I still have enough? How can I address it? 2) Check what are the features that have a big impact to the target.

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* For week 1, I investigated the structure of the data (ie dimensions), data quality (missing values, duplicates, distribution of each feature, proportion of success/fail. I used all the records available because from what I learned this week, the data is quite clean. I decided to begin my experiment with using standardized features. Because I started with using all the features, I would like the values of the features to be on the same scale.
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. My team member used all the features and put them all in the model. I would like to experiment on producing a model of comparable or better performance using less features. I began by loading all the features to a logistic model. Any feature that has a p-value that is not smaller than 0.5 will be excluded on the next round. I repeated this until only features with the p-values less than 0.5 remained.
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 was most familiar with logistic regression to use for predicting binary outcomes so I started with that. I used both Statmodels and SKLearn because I would like to compare how these two works. I ended using them both where I used Statmodels to exclude features and SKLearn as the final modelling method.

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. Accuracy of logistic regression classifier on train set: 0.83 Accuracy of logistic regression classifier on test set: 0.84 Training set RMSE 0.407 MAE 0.166 Validation set RMSE 0.405 MAE 0.164
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)
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. My main issue is not being very good a python so the process is slower than I would like to be.

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 think it's not yet a dead end! My main insight is that I learned a lot in setting up a proper experiment

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 would like to address the imbalanced target variable next time because I think it is important to resample.