EXPERIMENT REPORT

Student Name	Hitoshi Fukuda
Project Name	Kaggel Competition - NBA rookie career
Date	9 Nov. 2022
Deliverables	**<notebook name>** hitoshi_fukuda-week1_log-reg.ipynb **<model name>** Logistic Regression **<other>** The link of Git for the group 4: https://github.com/aprilgum/adv-dsi-2022-at1-grp4

---

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 goal is to build a model for predicting if a rookie player can stay as an NBA player for at least 5 years in the league based on his stats. The result could be used for rookie players to improve their performance and for related organisations and fans to speculate on the future performance of rookies.
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.* It is predictable by some performance ratings whether a rookie player can stay as an NBA player for at least 5 years in the league.
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 objective is to build a highly accurate model for speculating a rookie's future career and also to identify which variable is meaningful for the prediction.

---

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

1. Data Exploration
    - Create data frames based on the raw data
    - Grasp the overview of the data (such as the size of rows/columns)
    - Get information and description of the data:
      Crucial to check if there is any missing data or non-quantitative data
    - Visualise the data in correlation heatmap and plots:
      Important to understand the trend in the data with visualisation of it
2. Data Processing
    - Apply the Standard Scaler for standardisation of the data:
      To remove the difference in scales in the data and improve the accuracy of a model
    - Split the data into 3 different sets: training, validation, testing
      To measure the performance of a model with validation and test data and avoid over-fitting
    - Store the split data sets in the “processed” data folder
      To create reusable data set which are already cleaned and standardised

Then, in the future, I would have to address the issue of “imbalanced data”, in which some dominant classes would emphasise its presence more than the target variable. I will check if that situation is real in my data sets and, if true, make a solution to it.

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

I have not made so much to feature engineering, since the given dataset is originally clean to some extent. (no missing data, no duplicate data, all quantitative data)

However, I will have to take some steps from now on:

- Investigate the correlation of the dataset furthermore
- Investigate the plot furthermore and consider the effect of outliers in the dataset
- Check if there exists multicollinearity between variables and avoid it if needed
- Check if it is effective to create interaction features

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 chose the Logistic Regression model because the problem given is a binary classification, and this is why I did not choose Linear Regression. In the future, I would also consider other models such as Decision Trees, SVM, and Random Forest.

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.* <Result of the model based on the training data set> - Accuracy: 0.835 - RMSE: 0.406 - MAE: 0.165 Accuracy is not too bad, but it is probably possible to improve it. Regarding RMSE and MAE, RMSE is too higher compared to MAE potentially due to outliers in the dataset because RMSE is a squared indicator, which might affect the accuracy.
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 I have created for now is not so accurate, so it would result in not a significant prediction and make it impossible to improve the performance of a rookie player based on the dataset.
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.* - Difficulty in manipulating Git repository: Almost the first time using Git, but I checked the materials provided in/after the online session, and also my teammate gently caught up with me. - Difficulty in coding in Python: Also the first time using Python, but I reviewed the decks and did the exercises given in the class to get familiar with Python and its tools.

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.* - Feature engineering is needed furthermore with my experiment. - Building a good model is key, but feature engineering is also crucial to improve the quality of data and the accuracy of a model.
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.* 1. Further investigation into the features in the dataset 2. Check how can I improve the steps for feature engineering 3. More understanding of how to evaluate and interpret the results of the model 4. Consideration of the possibility to improve the model or apply another model