**Assignment 2 Write-Up**

I started this assignment by firstly researching about football games in general, as I don’t have any knowledge in this area. As soon as I got familiar with the rules and terminologies, I started to look for datasets that may be useful. I mainly focused on <https://www.pro-football-reference.com/super-bowl/>”, because I found this site contains tons of information and data that are clean and straightforward. Next, I looked for the features that can potentially affect the outcome of the game, such as the sum of the team’s score in the season. I also found out that there are many features that can be correlated with other features, so I was careful picking the features I would be using. I eventually decided to the “Team Statistics” dataset for every season from 1970 to 2022, because the features seems to be representative and are likely the determinants of the strength of a team.

One issue I encountered when I created the data is that the website only has the “Team Statistics” dataset season by season, so I had to manually input each season’s data into Excel spreadsheets.

Another issue is that if I directly copied and pasted the ‘winner’ and ‘loser’ columns, this would automatically assign the winners and losers, which would potentially affect the performance of my model. Therefore, I checked these two columns to ‘AFL’ and ‘NFL’ so that the winners and losers are randomly assigned.

My final dataset contains 54 observations, and includes: Seasons, AFL team name, NFL team name, afl\_win(dummy = 1 if AFL wins, 0 otherwise), a set of features for AFL team, a set of features for NFL team. I collected this specific dataset because the features are very representative in terms of a team’s strength and its performance for all games in one season. I am very sure there are other features such as injury report, can potentially improve my model, but it would be very difficult to avoid omitted-variable bias.

Machine learning workflow:

* My dataset was collected very carefully at the first stage, so I did not need to much data cleaning and feature engineering. However, I did check summary statistics and correlations when I imported the data to Python. I also standardized the features to make sure these are in the same scale.
* My final model is logistic regression. Before that, I went through Naïve Bayes and KNN, but I doubted that the independence assumption for Naïve Bayes is very unlikely to hold in this context and the dimensions are too high for KNN. So, I implemented Lasso first, with repeated cross validation method, because Lasso can bring the coefficients of those unimportant features to 0, which can minimize the dimensions in my model. However, the output is not very pleasant, because it almost kicks out every feature, but the prediction from this model indicates that NFL team has a higher chance to win in 2023.
* Then, I decided to try another model, logistic regression, which is suitable for this context, as my outcome variable is binary. I also tested the robustness of the model by repeated cross validation. The final prediction indicates that the NFL team (Eagles) has a higher chance to win. The accuracy of the model is 66%, which is not bad.

I believe my model may have good out-of-sample predictive power, because the model uses team statistics that can be commonly find, and my final prediction is already based on data out of the sample. It is very easy to add another row of observation and make a prediction on the new team statistics data.