Name, program, year: My name is Michael Li and I’m a U1 Software Engineering student.

Why MAIS 202: I chose MAIS202 to explore and learn about AI. I thought it was a great opportunity to get an early introduction to machine learning which is a subject that strongly interests me. I’ve attended past workshops taught by Isaac and Frank who were very good at explaining every theorical concepts and demonstrated great knowledge of ML. The experimented workshop leads and the appealing syllabus motivated me to sign up for this bootcamp.

Hobbies: Soccer, Gym, Piano, Travelling

Goals for the future: Launching a successful AI Startup that makes a positive change in the world.

● Brief description on what your project does – e.g. it classifies pictures of dogs by breed

The Football Prediction Bot predicts the outcome of English Premier League fixtures.

● Inspiration for the project

Microsoft’s Bing search engine correctly predicted 15 out of 16 matches in the knockout stages and the winner of the 2014 World Cup.

● Data sets, Libraries used – how did you arrive at your library and dataset of choice?

<http://www.football-data.co.uk/> contains datasets of every EPL seasons since 1993. However, only the most recent one (season 2018-2019) was used for this project in order to avoid noise and changes between different seasons (new players, new teams, new managers, etc.). The dataset is updated weekly with its fixtures. It was chosen, because it provides detailed information about the result of each game and contains the following features:

* Full-time result
* Home team/Away team full time goals
* Home team/Away team half time goals
* Home team/Away team shot and shots on target
* Home team/Away team free kick and corner kick
* Home team/Away team number of fouls and # of yellow/red cards
* Data/location/referee

Keras, Pandas, Matplotlib and Numpy were the primary libraries used for this project.

● Challenges encountered

The dataset for the 2018-2019 is really small since the season is still on going at the moment this project was published. In addition, the features present in the dataset are only available at the end of every game since they are statistics on the corresponding fixture. Therefore, it is impossible to obtain information on each feature prior to the match. Every team only plays against each other twice during the entire season. Hence, we cannot predict using the results of the previous game, because we only have one sample of each matchup.

● Favourite part of the project

In order to solve the problems related to my dataset, I proceeded to engineer features by using my knowledge of this sport. First of all, Pandas allowed me to explore and extract insightful features from past games results. In addition, it is a known fact to any sports fan that the home team advantage exists notably due to the fact that they have less fatigue due to travelling, moral support from the fans and familiarity of the home pitch. In fact, it is possible to notice a higher win rate for the home team by analyzing the games from the last 2 decades. Therefore, the features were engineered taking in account that every team performs differently at home and away.

The attacking and defensive strength at home and away of each team were calculated using the following formula:

Home Team Attacking Strength

Note: X represents any EPL team

Away Team Attacking Strength

Home Team Defensive Strength

Away Team Defensive Strength

The next feature is the conversion rate which is calculated using the following formula:

The results of the last five games also serve as features to give an index of each team’s form.

A detailed documentation will be available in the GitHub repo linked below.

● Provide a link to your GitHub repo!

<https://github.com/mlej8>