**Introduction**

Can the game of football, complex and unpredictable by nature, be understood better with the help of artificial intelligence and machine learning?

Historically, we can agree predicting the outcomes of certain matches, scores and league table positions is difficult. This randomness has proven lucrative for the gambling industry, highlighted by the fact that over half of the teams in the top two divisions of English football have a gambling sponsor on their shirt.( <https://www.theguardian.com/sport/2020/jan/08/how-the-betting-industry-has-become-inextricably-linked-to-football>)

Using historical training data from the first half of a season, we aim to build a system that will accurately predict match outcomes and hopefully, over the course of the second half of the season, provide a positive return on bets we would place using these outcomes.

There are many metrics to be measured during a football game, and part of the art will be choosing which inputs will have a positive impact on the systems predictions. What weighting does the number of corners taken by a team have on the final score? How about yellow cards? Possession statistics?

Our data sets will be taken from the top tier of English football, the Premier League, with the features we are most likely to use including:

* Final score
* Number of shots off target
* Number of shots on target
* Possession
* Number of passes
* Passing accuracy
* Yellow cards
* Red cards
* Home and Away
* Corners
* Fouls

The challenge for our system will be to analyse the noise each metric produces in relation to the final outcome, and then through learning determine a weighting for each input that will result in more accurate predictions.

Therefore, we will evaluate our system based on how accurate our predictions were based on historical data. As we can’t see into the future, it makes sense to train our system on the first 190 games of previous seasons, and evaluate its performance against the following 190 games. We will then do a final test of the system on matches from the second half of the 2020/2021 season up until the project deadline, and produce a final evaluation of the system’s accuracy.

Another form of evaluation would be comparing our predictions to the spread bets from betting companies, i.e. is our system better at predicting outcomes than following the wisdom of the crowd?

Data collection (web scraper) and implementing/extending an algorithm

**Work for TB2**

The first half of the term will centre around writing or adapting an existing algorithm to

Validate solution against actual outcome

Folding – train algorithm on 90% of data, validate on 10%, avoids overfitting

Sk learn and python will show how this works

**Method**

As this is a competitive market, we are fortunate enough to have many datasets at our disposal. As mentioned previously, part of our analysis will centre around which inputs are most important, as too many may impact our running time and memory usage.

The method we will be using is k-fold cross validation, over the last decade of Premier League seasons, and then finally evaluating our model on the current season. If we are to split each season by a ratio of 9:1, we will use the first 340 matches as our training dataset, and the subsequent 40 matches as our testing dataset.

Multinomial classifier

Label for dataset: win lose draw

Use a neural network, starts with arbitrary random weights

Propagation: if label says win but you predicted loss it will change weights to minimise risk.

Binary classification

Weights are what you learn on

World rank as a feature? Average pay of members

Win ratio

Normalisation: scale each number you feed in between -1,1