Project description and goal

This project is inspired by the Kaggle competition called Football Data Challenge. In this project, all the "Serie A" football matches of the last 8 years have been provided. The goal of this challenge is to predict the final result of future games with the odds of various bookmakers at my disposal. As I looked around the internet, it was hard to get external variables surrounding the matches such as the weather condition, the head to head against a current team, the teams current form, all of which can really affect the outcome of a game. Therefore I will have to choose another football league to analyze such as the English Premiere league which is more popular, and thus will be easier to find such information around the games. The essence of this project still remains the same; I have just changed the dataset in order to get more relevant information in order to learn.

Benefits to the client

Sports betting is something that is very hard to predict and it is important to remember that even if all of the odds are in a team's favor, there are no number of variables that can guarantee a victory as a sports game is not a game which is played on paper but it is played on the field!
Since we now understand that football and sports is also a game of luck and unpredictability, our aim is to remove as much as possible the unpredictable factors and to have an increased chance of actually predicting the right outcome of a game when placing a bet. With the help of our in house data scientists, the model that we will create will be able to predict the result of football games more often than not, which will make it almost like a sure investment in the short and long term. People invest in stocks, land, etc, but once you use our prediction model, you will get an immediate return of investment, no matter how little or big the investment you make in betting for particular games.

What do we want to achieve

We would like to predict the result of a football match as accurately as possible. In the kaggle competition, the winner was able to get a score of 0.66885 and the runner up was able to get a score of 0.53442. Our aim is to predict the result of the football games with an accuracy above 0.5. If we do that, we will consider this a success as we are able to predict the outcome of a game more often than not.

How we expect to achieve it

In order to achieve those results, we will have to clean the data, to enrich the data and to make sure we can add additional variables to the data such as maybe the Tier of a team (such as Tier A, B, C which will determine the strength of the team in general. For example the current top 6 teams of the English premiere league can be considered as Tier A teams etc…), the weather conditions of the match, the teams' current form according to the previous games they have played, the past meetings between the two teams, the Referee assigned to the match.

Since this is a multiclass classification problem (as there are more than 2 outcomes) there could be a number of ways to do this and I will choose between the following two:

- If we change the problem and make it so that we train our model to find whether "the home team wins or not?" This changes it to a binary classification problem which will be easier to deal with as if they don't win it means it will either be a draw or that the away team will win. For this we can use Logistic regression. We could then repeat the same process to find out if the away team will win or not and then that would mean that the result will be either a draw or a win for the home team. If we join those two results and find that the probability for a home team win and an away team win is the same, then we can predict that the result will be a draw.
- If treat it like a multiclass classification problem, we can apply neural network algorithms or decision trees in order to get the predicted result between a home win, away win or a draw. This method might be more difficult but there will be fewer steps involved compared to the first method.

What are the ethical implications?

There are no ethical implications as the perfect model for sports prediction does not exist. If it did, the effect it would have is to take the fun out of sports and the value of sports would decrease, including all the investment made in sports in general (salary of players, people's interest in sports will decrease). If there are any ethical implications, they would be linked to gambling.