**Data Available**

The most popular sport in the world soccer has the most data available as in big leagues like the premier league nearly everything is recorded. The Premier league in England is known for being one of the most popular leagues in the world. It is also known as one of the toughest leagues in the world to predict as often the bigger teams can be beat by the smaller teams. This makes it difficult to predict the outcome of the matches and even the league table at the end of the season.

Because soccer is extremely popular there is a wide range of data available on the internet. This data includes everything from goals and assists to even x and y coordinates of where a player was on the pitch when he scored. This amount of data is very beneficial when it comes to analysing and predicting the outcomes of upcoming matches.

**Source of data**

I’m sourcing my data by scraping it from websites and storing it in csv files. The website I’m scraping from is Understat.com. This website has very details amount of data from six leagues in European soccer. The leagues are the English Premier League, French Ligue 1, Spanish La Liga, German Bundesliga, Italian Serie A and Russian Premier League. It contains data for each of these leagues going back to the 2014/15 season. Admittedly it doesn’t go that far back as 8 seasons isn’t a whole lot, but it makes up for that with the amount of detail in the data that is stored.

It stores that on the league itself and the teams, fixtures, and players in the league. For each team it stores data on the individual players showing their goals, minutes, appearances and much more while for the team itself it stores formations used, shot zones, most frequent time for the team to score in a match and much more. For fixtures it stores data on the players that either play or get subbed on as well as the events that happen in the match such as a yellow or red card and a goal or shot and there is plenty more data that they have also. Then for the players it stores data on the goals a player scored whether by left or right foot and also the position of the player when they score as well as much more data.

**Data Chosen**

The reason I chose Understat.com to scrape data from is because it has lots of data available and has the data I need for my project. (Add more as I add to project) What I need for my project is data on players goals and shots as well as matches played in, and teams scored against. I also need data for each fixture so I can get not only the score but who scored and what other events happened. Understat.com has all the data that I need for my project.

**Research Domain**

The research domain I’m working in is the prediction algorithm domain. This is where the goal is to make an algorithm that can predict future outcomes by analysing past data that you have collected. If done right the algorithm can have high accuracy rates and be very beneficial to the person who created, it. These algorithms are not fixed and are usually revised regularly to incorporate any changes that have occurred in the data. If new data is analysed by the algorithm, it will most likely change the future outcome of the prediction and means that the prediction will have to be recalculated.

**Prediction Algorithms**

Prediction algorithms can be used to predict just about anything but for my project I’m trying to predict soccer match results and the chance of a player scoring against a certain team. The goal of my project is to make it where my prediction algorithm can predict the outcome of a soccer match or the player scoring against an opponent with good enough accuracy. Obviously, it is never gonna be correct all the time as its impossible to predict the outcome of soccer matches as with the past data you have collected you can’t account for things you can’t measure in data, for example if a team is playing bad and loosing and they substitute on a player, and he turns the game around. How would you account for that player coming on?

Even though you can’t account for things like team morale and maybe even an inspiring team talk from the captain to lift the team’s mood, there is a lot of data available that a prediction can be made that is accurate enough and should be close to what the actual outcome of the match would be. Factors like goals, expected goals, chances created, team form and defending influence the outcome of a match greatly. So given the data available when making the prediction you should come out with a fairly accurate prediction for the outcome of the match.

**Choosing the Model**

When researching different models to build my prediction algorithm I came across the Arima Model, Bayesian Hierarchal Model and Poisson Distribution Model. Arima is a time series model that predicts events based on past values. This model is used with data that exhibits patterns and is not random events. I decided not to use this model and soccer is a random sport. Bayesian Hierarchal model is a time series model but is more suited to predicting outcomes in soccer as it uses the observed data to account for all the uncertainty that is in the data.

Poisson distribution is used to calculate how often an event will occur over a given period. Poisson distribution calculates events independent of each other. For example, when trying to calculate the outcome of a football match Poisson distribution when predicting which team will score doesn’t take into account the score current score in the match. If a team were losing by a goal they might push forward and try to score and equaliser which effects the likelihood of both teams scoring.

**Poisson Distribution Model**

This model suited my needs as I’m predicting single match outcomes and this model is suited to this. Poisson distribution as said above calculates events independent of other events in the match. This while not being completely accurate still gives an accurate enough prediction of what the outcome of the match should be. The Poisson distribution model suits my match prediction as I’m using goals scored and conceded for both teams as well as my player scoring prediction where I’m using goals scored by the player and conceded by the opposing team.

The most important aspect of the game of soccer is obviously the scoring of goals. Despite the significance of other factors in the game like possession, defending and chances created but to win a match a team has to outscore the other team. My match prediction algorithm aims to predict goal scoring in a match and from that predict the most likely outcome of that match. My Player prediction algorithm aims to predict the chance of a player scoring against an opposing team.