**Team Members**

* Rebecca Adaimi (rma2625)
* Stephan Boner (sb54652)
* Christopher Gang (cg37877)
* Nick White (nww295)

**Motivation**

People have always tried to predict the outcomes of soccer matches. In the 2010 World Cup, Paul the Octopus correctly predicted the outcome of every one of Germany’s seven matches. Although many attempts have been made, no one has come close to the accuracy of the predictions made by Paul the Octopus. There are many commercial services that make predicting soccer matches their primary business. They utilize data about previous matches, team performance, and players to make their predictions. They use various software programs and algorithms to aid their data tracking. Furthermore, they utilize the advice of experts in the area as part of their predictions. What if there was a way to build a model that could accurately predict the outcome of soccer matches?

**Goal**

Our goal is to investigate whether machine learning and deep learning algorithms can be used to predict soccer matches. Moreover, we aim to analyze and test whether scraping recent news about players is important for predicting the outcome of a match.

Thus, the main questions that we aim to answer at the end are:

* Can we build a model to be used for the purpose of match predictions?
* Is recent news about players important when predicting the outcome of a match?
* How much player history is relevant?

**Data Collection and Scraping**

We decided to base our model on data collected from the English Premier League (EPL), which is the soccer league that produces the most revenue in the world [1]. We will form the basis of our dataset by collecting and concatenating public datasets. There are several that we have found on Kaggle and soccer betting websites. We will add additional information to our model by scraping the internet for news and interviews relevant to the soccer teams leading up to their match. We will use natural language processing (NLP) on these news articles and interview transcripts to get the general sentiment of the teams before the match.

**Exploratory Data Analysis**

The datasets provided at [2] contains many features. On one hand, we have the data from the game for several features, e.g. Result, Goals, Halftime Results, Yellow Cards, Corners etc. (see table below). Furthermore, we have also odds from many odd makers. In general, we have to distinguish between features, which are known before the game (like the playing teams where home and away are also important, date, time, weather, referee, etc.) and features, which are not known before the game, like e.g. number of yellow cards, corners etc. There are two approaches to deal with the unknown data: We could just ignore these or we could try to predict e.g. the amount of the yellow cards or corners for every team. Furthermore, we have to work out a score for every team which represents the quality of the team and also the current fit of the team, like how they did in the last games and how they feel within the team.

*Rating each Team*

To rate each team, we need a score of the quality of every team. This could either happen by hand or out of features like the outcome of the last X games, the amount of scored and received goals in the last X matches, the market value of a team (e.g. the sum, maybe also include std. deviation, median or sth like this) to distinguish between a team with a high market value at average or just a few expensive players, g (e.g. last three, five, ten or a combination of these features), the score from the scraped news about the team and the number (and quality) of injured or suspended players. Also an influence could be how many players have the right amount of yellow cards (3,7,11,...) that they would miss the next match when they get another one in this game. The data about the market value and injured players we can fetch from [3] or a similar homepage.

*Rating the Referee*

Since a referee might have a big influence on the game, we could try to give the referees a score. This is probably really hard since it’s not really measurable out of the data. We could measure the quality of a referee by the number of yellow cards, fouls etc., but since this is also really depending on the playing teams, this might be very inaccurate. We might include e.g. the average of yellow cards or fouls of a referee by predicting the number of yellow cards.

*Including the Odds*

Since we can fetch at least some of the odds directly via the homepage of the odd makers, we can include them like normal features. For some homepages, we have to register as affiliates, but then we should have access to these data.

*Predicting Game Stats*

To get a more accurate result, we should try to predict the value of yellow cards, half time result, corners etc. For this, we will use a combination of the following features. If we e.g. want to predict the amount of yellow cards for the home team, we take the average number of yellow cards of the home team in the last X matches, the average number of yellow cards the opponent of the away team got in the last X matches and the team score of both teams.

*Unpredictable things*

Since there are every year teams coming and leaving the Premiere League and might also be new referees coming to the league, we have some things we can’t rate. Since we rate the teams anyway with many features like market value, last few games etc., this won’t influence the rating unless at the beginning of a saison. But since every team performs different in a new saison, it’s anyway unsafe to start betting at the beginning of the saison. The new/unknown referees, we could handle like missing data and replace referees with e.g. less than 5 games with the mean of the referee score.

*Models*

To train our model, we should probably use so many data as possible, since we wanna figure out the influence of e.g. yellow cards or corners to the score of the game. But we have to be careful since football changed over the years (the way how the teams play and also some rules have been adjusted), so we have to take care, which data we can use.

We have to figure out then, which model we finally use to predict the features and the outgoing of the game. We should give a try to XGBoost, LightGBM, Naive Bayes, SVM and Random Forest. To find a rating for our models, we can e.g. use the data of 2016/17 as Test-Data and the data from some years before as training data.

**Important Data in our Dataset [4]**

|  |  |
| --- | --- |
| Date | Match Date |
| HomeTeam | Home Team Name |
| AwayTeam | Away Team Name |
| FTHG / HG | Full Time Home Team Goals |
| FTAG / AG | Full Time Away Team Goals |
| FTR / Res | Full Time Result (H=Home Win, D=Draw, A=Away Win) |
| HTHG | Half Time Home Team Goals |
| HTAG | Half Time Away Team Goals |
| HTR | Half Time Result |
| Attendance | Crowd Attendance |
| Referee | Match Referee |
| HS | Home Team Shots |
| AS | Away Team Shots |
| HST | Home Team Shots on Target |
| AST | Away Team Shots on Target |
| HHW | Home Team Hit Woodwork |
| AHW | Away Team Hit Woodwork |
| HC | Home Team Corners |
| AC | Away Team Corners |
| HF | Home Team Fouls Committed |
| AF | Away Team Fouls Committed |
| HO | Home Team Offsides |
| AO | Away Team Offsides |
| HY | Home Team Yellow Cards |
| AY | Away Team Yellow Cards |
| HR | Home Team Red Cards |
| AR | Away Team Red Cards |
| HBP | Home Team Bookings Points |
| ABP | Away Team Booking Points |
| Many Odds | Odds for H, D, A from different bookmakers like bet365 etc. |

**Approach**

The first thing we will do is to rate the teams, including the scraped news. Then we will include the odds from as many odd makers as possible. After this we can start to predict features like yellow cards, fouls etc. and guess their influence on the game.

**References**

[1]“Annual Review of Football Finance 2017 | Deloitte UK.” *Deloitte United Kingdom*, 11 July 2017, www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance.html.

[2]*England Football Results Betting Odds | Premiership Results & Betting Odds*, football-data.co.uk/englandm.php.

[3]“Football transfers, rumours, market values, news and statistics.” *Transfermarkt*, www.transfermarkt.co.uk/.

[4] *Football Betting*, football-data.co.uk/notes.txt.