Beat the Bookie

A Preprint

Group Name: `Group I`
Department of Electronic and Electrical Engineering
University College London
London, WC1E 6BT

7th January 2019

1 Introduction

The aim of this project is to utilize past match data to make a prediction for the result of a match in the future. We will not be predicting the exact score; Instead we will be predicting if the Home team wins ("H"), the away team wins ("A") or the game ends as a draw ("D"). The league that we will be concerned with the English Premier League (EPL).

In each season of the EPL 20 domestic teams compete. For each matching, two games will be played: one at each of the two teams' home stadiums. This will result in a team always being the "home" team and the other being the "away" team.

2 Data Transformation and Exploration

The raw match data that was provided to us in the form of a csv file was imported and analysed thoroughly to identify any patterns and trends that could be exploited. Furthermore, the provided data will be manipulated to calculate metrics that describe the team's performance. These metrics will act as features in the design matrix and will eventually be used to train the classifier model.

2.1 Columns of the Provided Dataset

The provided dataset contained 22 columns. It contained data of all EPL matches from 2008 to 2019. Table 2.1 lists the columns of the provided dataset and a brief explanation as to what each quantity represents.

2.2 Number of Matches

The first thing that was observed was the number of matches played by each of the teams. It was identified that not all the teams played the same number of games from 2008 to 2019. This is most likely due to some teams getting relegated (dropping out of the EPL) and other teams being promoted (joining the EPL). This will result in the consistently strong teams who have remained in the EPL throughout having played the most games while those that have been relegated and promoted having played less games in the dataset. This is clearly visible from Fig. 2.1. The 7 teams with the most matches have played the same number of matches and are likely to have been consistently in the EPL throughout the 2008-2019 time period.

Another quantity that was confirmed was if all possible pairings have played one another at least once in the time period of the dataset. This was found to be not the case. This is not possible due to relegation and promotion changing the list of competing teams from season to season.

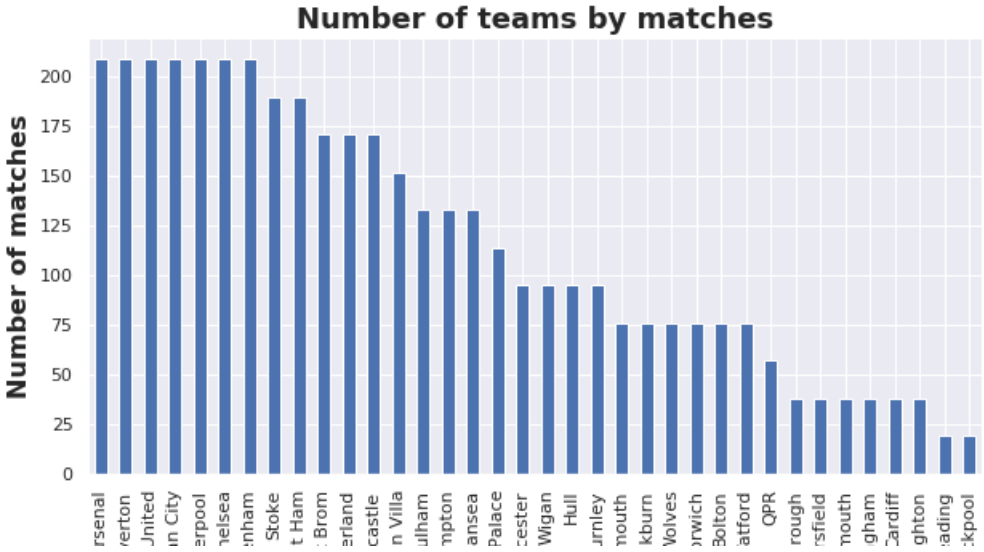

Figure 2.1: A count plot of the number of matches played by each of the different teams. It is clearly visible that not all teams played the same number of games.

2.3 Histogram Plots of Numerical Columns

From Table 2.1.1 it is evident that most columns are integers. The distribution of each of these columns was investigated by plotting the histogram for each column. In addition to this, each bar in the histogram is split according to the full-time result (FTR). The plots are presented in Fig. 2.2 for each of the 16 integer columns in the provided dataset. It is clearly visible that most of these plots have a clear maximum either side of which the frequency drops. This could possibly indicate that these quantities may be modelled using a Poisson distribution. This is possible as each of the quantities is measured per match and therefore it is measured at a constant rate. This relationship is most clearly visible with datasets that have a large variation (i.e. have a large range). An example is the HS plot. It is clearly visible that this set of data may be modelled with a mean of around 12.

TABLE HERE

2.4 Analysis of Non-Numerical Columns

As shown by table 2.1.1 the non-numerical columns of the dataset are “Date”, “HomeTeam”, “AwayTeam” and “Referee”. There are 36 unique teams listed in the “HomeTeam” and “AwayTeam” columns. As expected, all teams that appear in “HomeTeam” also appear in “AwayTeam” and vice versa. As previously discussed, one season only consists of 20 competing teams. The larger number of teams in the dataset is once again due to the relegation and promotion of teams from season to season.

There are also 36 unique referees across the dataset. Despite this number being equal to the number of unique teams, this is believed to be a coincidence as no specific rules on the assignment of referees that would result in this was found.

The “Date” column was analysed and found that matches only occurred in months January to May and August to December. This suggests that a season of the EPL runs from August to May. This can be used to split the dataset by season and thereby find variation in team statistics by season.

2.5 Correlation between columns of raw match data

To analyse the patterns in the dataset, the correlation between each of the columns was calculated. This data was represented by using a heatmap. The colours of the cells correspond to the Pearson correlation coefficient between the two columns. The most distinct feature of this heatmap is the diagonal representing correlations of 1.0. This is due to the correlation of each of the features with itself. It also follows from this that this correlation matrix is symmetrical.

The half time home goals (HTHG) is strongly correlated with the full-time home goals (FTHG) with a correlation of 0.7. This also applies to the away case (HTAG and FTAG). The other strongly correlated columns are the half time result and full-time result columns (HTR and FTR). This strong correlation implies that the team that was winning at half time is likely to win the whole match as well.

Another interesting correlation is that between the home corners (HC) and home shots (HS) with a correlation of 0.5. This suggests that some corners result in attempted shots at the goal for the team. Furthermore, HC also correlates with shots on target (HST) with a score of 0.4 suggesting that corners also result in shots on the target. The same applies to the away team (HC, AS and AST). However, at away matches, the correlation between AC and AST is 0.3 as opposed to 0.4 at home suggesting that corners are less likely to lead to shots on the target when a team plays away.

A more obvious correlation is that the home fouls (HF) correlates with the number of yellow cards received by the home team (HY). This can easily be explained by the fact that aggressive/dangerous fouls often lead to yellow cards.

It is also clearly visible that the referee column has no correlation with any of the other columns, which is indicative that it is chosen at random and does not influence the performance of either team. This relationship is expected as the job of a referee is to preside of the game from neutral point without being biased to either team.

IMAGE HERE

3 Methodology Overview

4 Model Training & Validation

5 Results

6 Final predictions on Test Set

7 Conclusion

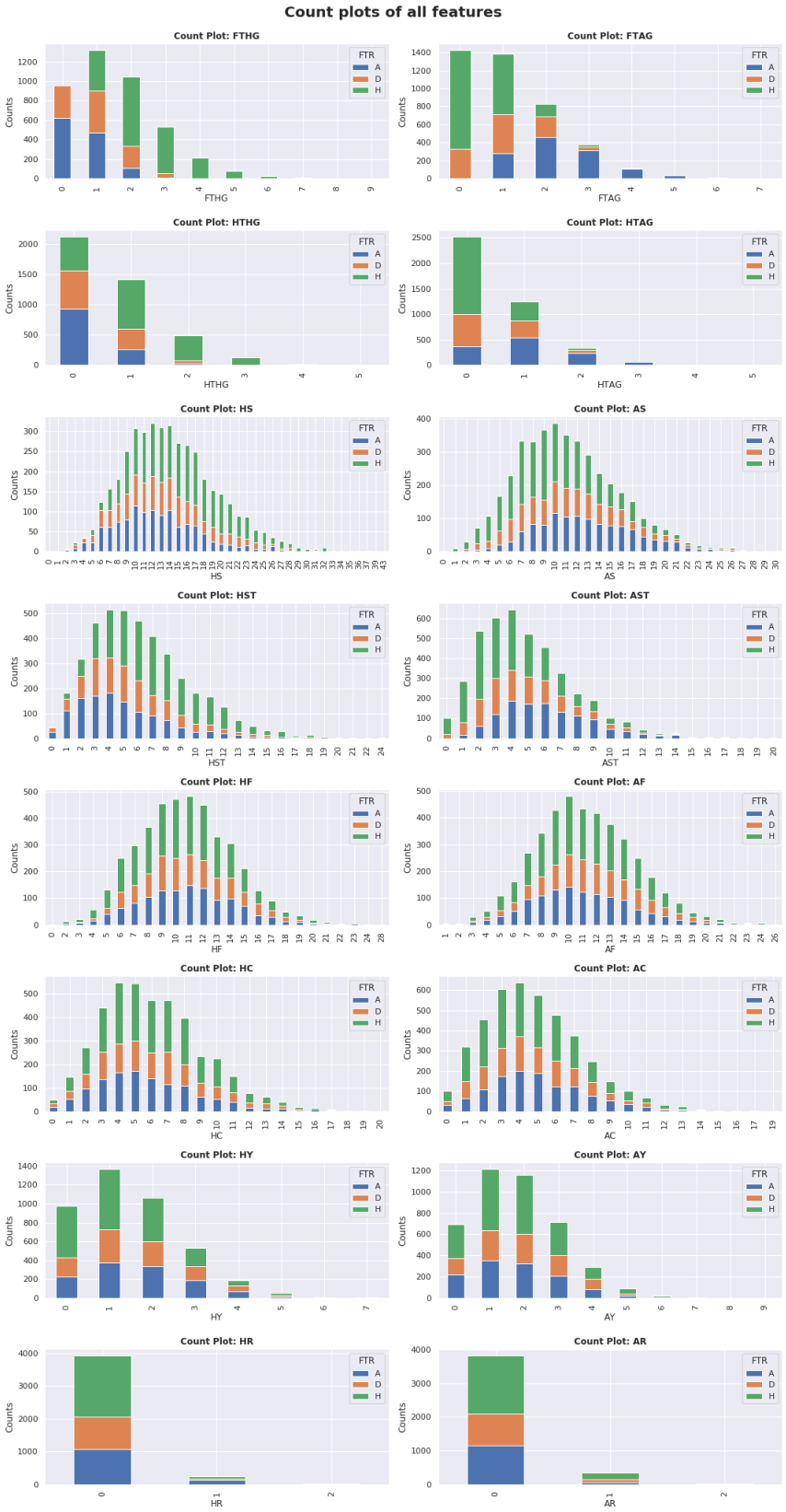

Figure 2.1: A count plot of the number of matches played by each of the different teams. It is clearly visible that not all teams played the same number of games.