Milestone 3 Writeup

How the data was preprocessed and predominant techniques used

The first data preprocessing step we performed was turning the date column in our dataset into a datetime object. We did this not only to make the formatting of each of the dates less confusing (the dates were originally formatted as DD/MM/YYYY) but also so that we will be able to use all of the functionality that the Pandas library has for dates. Changing the date column to a datetime object will allow us to look at events like winning and losing streaks in our analysis. The next change we made was adding a column named “HighScoring” to our dataset which is a binary variable telling us if a given match could be considered high scoring or not. If the total goals scored in a match was greater than or equal to 5, the “HighScoring” value will be 1, otherwise it will be 0. We thought that adding this column could be useful because we are curious if it is harder to predict the outcome of a match when lots of goals are scored. Another data preprocessing step we took was converting our categorical variables to discrete numerical values. The columns “FTR” and “HTR” tell us the full-time and half-time result of each match, so if the home team won the “FTR” would be “H”, if the away team won the “FTR” would be “A”, and if it was a draw the “FTR” would be “D”. To ensure that these columns could be included in any model we build in future parts of this project, we replaced these letter values with number values (2 for “H”, 1 for “A”, 0 for “D”). Another important thing we looked at was the importance of each feature (column) in predicting the outcome of a match. We were able to rank the importance of each of the features in predicting match outcome, which helps give us an idea of what features will be important and which ones won’t if we end up building a predictive model. The techniques involved for these data preprocessing steps include using the pd.to_datetime() function, adding columns based on the conditions of other columns, finding and replacing values, and using the ExtraTreesClassifier package.

The next main step we took in preprocessing our dataset was deleting any extra unnecessary columns to ensure our data set only had important values. The first column we deleted was the division column. This column was redundant as all the teams are playing in the same division, the Premier League. The next deletion we did was deleting certain betting statistics. We uncovered in our research that there are two different types of betting, and thus two categories of statistics: win-loss betting and spread betting. The former looks at the odds of a team winning, losing, or drawing, while the latter looks at how accurate a certain part of the event is. For example, in spread betting, you would be betting on the amount of total goals scored. While spread betting can be an interesting insight, it is rather extraneous to what we are trying to predict, which is a teams chances of winning and losing. Thus, we decided to delete all the statistics pertaining to spread betting. With this, we were able to trim off 17 columns of statistics for all 380 games. The next data preprocessing step we took was using feature engineering to enhance the data set. We took the average of all the different betting organizations’ odds for the chances of a home team win, away team win, and match draw and put those averages into new columns. From there, we were able to calculate the percentage probability of each event occurring, which we also added to a new column. Lastly, we added some stat differential columns for the different features in our dataset. We believe adding these feature differentials as well as the winning percentages and match results will help us determine the best way to predict wins.

New insights uncovered from preprocessing steps

One new insight uncovered during data preprocessing was the amount of matches in the season that would be considered high scoring. Around 55 matches had 5 or more goals scored and this should be a large enough sample size to help us determine if predicting the outcome of high scoring matches is more difficult to do. Another new insight uncovered during data preprocessing was the importance of each of the features in predicting match outcome. This gave us a good initial idea of what our important features should be in any future model, and the ranking of the features can be found in our code. The last main insight that we uncovered during our data preprocessing were the average odds, probability of winning, and feature differentials. Adding these columns allows us to get some concise statistics that can help predict a potential match win or loss.