Manchester United vs Chelsea: Score Line Predictions

MATH 321: Mathematical Statistics I
Tuan Dang
November 15th 2021

1 Overview:

On November 28th 2021, Manchester United will face off against Chelsea in the English Premier League. This project will aim to predict the winner and score line of this match.

At the time that the analysis for this project was conducted, each team has played 11 games, and this will be the first time they play against each other this season in the EPL. After 11 weeks into the season, Chelsea sits comfortably in 1st place with 8 wins out of their 11 games while Manchester United falls behind in 6th place with 5 wins. The match will take place in the Stamford Bridge stadium, so Chelsea will be the home team.

2 Data sets and Methodology:

I found the datasets for past seasons of the EPL on Kaggle [1]. There is a csv file for each season, where each entry is a match and the columns are the information, results, and statistics for that match. For the purpose of this project, we will use two datasets from that Kaggle post, the 2018/2019 and 2019/2020 seasons.

I downloaded the datasets and kept a few of the features that I think will be usable for our models. I also added some data on my own which I think will increase the accuracy of the models.

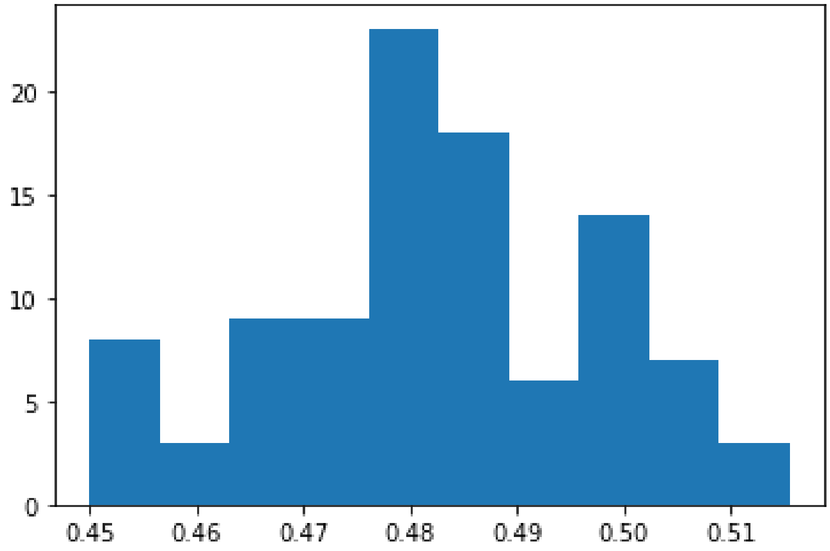
Modified dataset features:
Features from the original dataset:
HT Goals: the number of goals the home team scored in that match
AT Goals: the number of goals the away team scored or the home team conceded
Match Result: H = HT wins, D = Draw, A = AT wins

Added features: [2]
Home Team: the final rank of the home team at the end of that season
Away Team: the final rank of the away team at the end of that season
HT GF: goals HT scored in the entire season
HT GA: goals HT conceded in the entire season
HT Avg Scored Goals: average goals scored in that season (HT GF / 38 matches)
HT Avg Conceded Goals: average goals conceded in that season (HT GA / 38 matches)
AT GF
AT GA
AT Avg Scored Goals
AT Avg Conceded Goals

3 Part 1: The winner *probably* is...

The dependent variable that we need to predict is the match result (H, D, A), so I decided to go with a classification model. For simplicity, I decided to also go with the Decision Tree classification model, although I think a logistic regression model could have worked here.

The parameters are similar to the previous model. The accuracy comes out to be around 50%, and the model predicted that Chelsea would win.



Histogram of the accuracy score of the match result model

4 Part 2: Score Line Model

After a few times of trial and error, I decided to go with a classification model over a regression model for this problem. I wanted to simplify the problem by having the model produce the prediction in the form of a whole number.

The current 2021/2022 EPL season has not been that long yet, so I thought we would not have enough data points if we use all the matches so far, especially when Man United and Chelsea have only played 11 games each this season. So instead, I decided to train and test the model using all the games from the 2018/2019 and 2019/2020 seasons.

The next problem we need to figure out is to decide what features we want to be in the model. This proved to be a problem because each year the three bottom teams of the league are relegated and replaced by three other teams. So we don't get consistent data team-wise from year to year. The way I decided to address this is instead of considering each team individually, which would be very hard since there are so many variables to account for (players, coach and staff, injuries, substitutions, strategies, weather, locations, etc), I would use the team's rankings as a variable of how strong of a team they are. I'm unsure if this approach is valid or not considering we are making a lot of assumptions for this to work, but it does simplify the problem quite a bit.

With the first Decision Tree model that I built, I used the home team and away team rankings as independent variables in order to predict the number of goals the home team and the away team would score individually. The accuracy came out to be only about 25% for each, which we will try to improve later on.

For making predictions, we can plug in 1 and 6 (the current rankings of the home team (Chelsea) and away team (Man Utd) respectively). With the current results of this model, I predict that Chelsea will win 3-1.

5 Part 3: Improved Score Line Model

After working on the first model previously, I looked at all the possible variables again in the original dataset to see which one is viable to include in the model to increase accuracy. I didn't really find anything useful since most of the information will not be available until the match actually starts (number of shots on target, number of fouls, etc). And if we analyze more specific details like each player on the teams and their individual performance, well, that just seems like diving deeper into the rabbit hole and introducing even more variables into the mix.

While looking at the records for the 20 teams of each season, I noticed two more features that might help us: the total goals a team scores or concedes in the entire season. This feature wouldn't be useful by itself since each team from previous seasons played 38 games, while Man Utd and Chelsea have only played 11 games each this season, so comparing the total goals wouldn't be reasonable. Then I came up with the idea to take the average of the goals that they scored or conceded in all of their games each season. I figured this would help improve our model, especially in terms of score line since this can be a way of measuring if a team is good at attacking (scoring more goals) or good at defending (conceding fewer goals).

I decided to manually enter this information into the data set and include these features in the previous model to see if it would improve anything. Surprisingly, as it turns out, it barely changed the results of the model at all! The accuracy is still the same at around 25%, and if I were to use this to make a prediction, I would still say that Chelsea will win 3-1.

6 Accuracy, Uncertainty, and Future Improvements:

There are some things that could be added to improve the analysis. Graphs probably would have been good to show if there were correlations between the variables, as well as hypothesis testing (I didn't find a good way to do Wald tests in Python).

Personally, I don't think the modelling part is especially impactful in solving this problem. The data set and variables actually seem much more important, as I can't find better features to include in the model.

Originally, before thinking about using classification models, I first considered using the Poisson distribution in some shape or form, which made sense for this problem (the number of goals a team scores or concedes in a game would follow a Poisson). I decided to disregard the idea since there seems to be too many variables for the Poisson to handle. If it was possible to apply the Poisson somehow, we could get a joint probability distribution of the goals a team scores and concedes. Once we have that distribution, we can get our predictions from the most probable case when we plug in the numbers.

7 References:

Data sources:

[1] $https://www.kaggle.com/saife245/english-premier-league/version/3$

[2] $https://www.footballcritic.com/premier-league/season-2021-2022/2/50885$