Introduction to Math for DS Group Mini-project

IMDS Group 24
Zehao Qian, Mohammad Jamshaid Iqbal, Chloe Mendez

November 20, 2023

1 Problem

Looking to find out which team is more likely to win the league given their current form. The focus will be on the English Premier League and the target is that the model can be used for other leagues too.

2 Variables

The variables that we will be taking in are the following:

1. The number of matches won by the nth game week, v1.
2. The number of matches lost by the nth game week, v2.
3. The number of matches drawn by the nth game week, v3.
4. The number of goals scored by the nth game week, v4.
5. The number of goals conceded by the nth game week, v5.
6. The number of points gained by the nth game week, v6.
7. The number of matches in the last 5 games, v7

3 Method

We will change these variables into a 7-dimensional vector:

$$V = [v1, v2, v3, v4, v5, v6, v7]$$

3.1 METHOD 1

The first method that I can think of is using vector projection to see which team will be performing the best by the end of the season given current statistics.

We will then normalize the above vector for each team using the following formula:

This will help us to compare the team's performance with each other. We can then define a reference vector; a reference vector can be calculated using the performance of previous champions by the nth game week for the season they were championed it. If we have the data for the past 10 seasons, then we will have 10 reference vectors and we can take the average out for those reference vectors. Let's call this reference vector Vr.

Next, we will project each team's normalized vector onto the reference vector. This will give us how each team's performance is compared to an ideal performance. The team with the largest projection is the team that is most likely to win.

3.2 METHOD 2

This vector is basically representing the team's performance till the nth game week. We can use this vector to predict team's performance for the remaining matches by applying linear transformation.

The linear transformation will basically transform the 7-dimensional vector to a 1-dimensional vector which will be the total number of points by the team. Fv : R7 -> R

The linear transformation F will take in the vector V which is the vector of the team's statistics in the nth game week and map it to the predicted point at the end of the season. $\mathrm{F(V)} = \mathrm{wV} + \mathrm{b}$

w is the 7-dimensional weighted vector which is basically a weighted vector for each statistic. b is the scalar bias. Scalar bias is a systematic error in a model and the importance of scalar bias will be to make the value of the points closer to the accurate value.

To take out the weighted vector and scalar bias, we will use the data of previous seasons (we can use the data of 10 years) and then use gradient descent in Python. After inputting the historical data, we will be able to take out the weighted vectors and scalar bias which then can be applied to the formula above to take out the points.

4 Limitations

The model just takes in the statistics such as wins till a certain game week, the number of goals scored, etc. as input. This might not be an accurate representation of the prediction for who will win the Premier League. This could be because there are other factors that can come into play such as the number of quality players bought by a team, the amount of money spent on transfers, the number of players injured during the season, the experience of managers, etc. There are many more factors that could affect a team's performance throughout the season, but our model uses the on-field statistics to predict who will win the

league. A linear regression model can be used to check the accurate representation of who will win the league by using more data.

5 Data

https://www.premierleague.com/tables