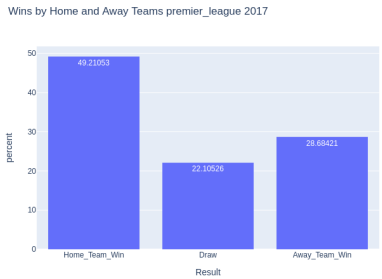
PREDICTING FOOTBALL MATCHES

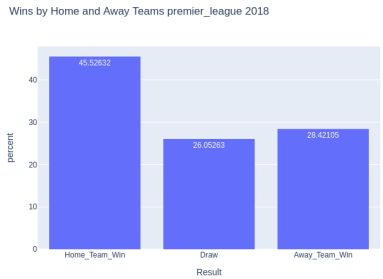
ADAM H WISE

1. Considerations in predicting the outcome of a football match

1.1. **Home Team Advantage.** Either due to familiarity with the playing field, a psychological boost from fans, or some unthought of other reason, data indicates that teams are more likely to win when playing at their home stadiums than away. As the graphs below attest, win percentage for a home side is significantly higher than for an away side. In recent years, playing on a team's home field yields wins between 45% and 50% of the time. This compared with a 15% to 30% chance of a draw and 25% to 35% of a loss.



(A) 2017



(B) 2018

(C) 2019

(D) 2020

Figure 1. plots of win/draw/loss percentages in the previous 4 premier league seasons

1.2. **Recent Form/Results.** Logically and by examination of fact, a team is more likely to perform well after a string of good performances especially against a side which has been performing poorly. This extends as well to the previous season in many cases as teams tend to have a final position in the league table which is similar to where they finished the last season. This is why it's a good idea to calculate both winning and unbeaten streaks for football clubs as well as the number of points they have won in the current season as an indicator of the likelihood of winning a game.

1.3. **Elo Rating.** While technically this measure is calculated from performance, a team's elo rating is a good standard factor in whether they will be able to beat another club. This system works by assigning each club a base value and raising or dropping it depending on whether games are won or lost and to whom they are won or lost. This system is optimized to show recent strength of a club however with modern computing a Glicko score would perhaps improve the accuracy.

1.4. **League Position.** While also technically this is just a measure of previous performances, a team's position in the league will also factor in to the likelihood of them beating a team lower or higher in the current table.

1.5. **Stadia.** This is a seemingly irrelevant factor however it is perhaps a good measure of a side's historical strength. As good teams have larger stadiums than other sides and sides which win games are often the stronger, perhaps a machine learning model would benefit from the inclusion of the capacity of a team's home ground.

2. Other factors for which data is unavailable

2.1. **Team Sheets and Injuries.** A telling indicator for a winning side is whether they are playing their A or B team and how many players are available on the roster. Logically a team marred by injuries cannot play as well as one with full choice of players. However, this data is unavailable to us so it my not be factored into the algorithms.