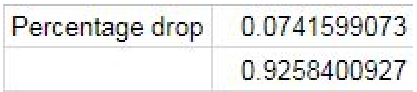
Parameter Estimation

Due to the relative lack of modern research into the granular events within a football match, we were forced to estimate certain variables that would affect the outcome of these micro-events. More research needs to be done in these areas, including fitting relevant models to current data.

One of the first things we noticed during our data analysis stage was that home teams win about 45.8% of games. That's a lot! Thus, we felt that there was an inherent need to scale the probabilities of each micro-event based on whether a team was playing home or away. However, we could not simply use some percentage value and multiply it with the calculated result, since we were simulating smaller events within a match. After some research on the effects of home advantage[1][2], we found that there were certain significant factors that helped home teams win more often. Amongst them is a statistically significant decrease in possession lost and a statistically significant increase in number of goals scored.

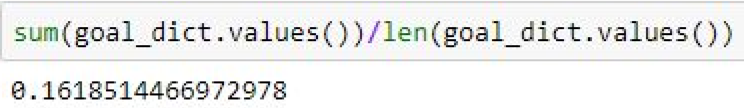
As such, we sought to estimate the parameters that scale the above probabilities based on the data we had. For the possession lost data, we retrieved data from Whoscored.com and calculated the ratio of completed passes to attempted passes for both home and away games. This data was processed by team, to avoid bias.



Percentage drop	0.0741599073
	0.9258400927

After processing, we found that there was an average of 0.0074 percentage drop in pass completion rates in away games compared to home games. Thus, we can use 0.92584 as a metric to scale the probability of a home team losing the ball, thus ensuring they keep the ball more often.

Next, we attempted to measure the effects of home advantage for number of goals scored. We found from the Kaggle dataset that on average, teams had 16.2% more chance of scoring a goal from any shot. Thus, we used this value to scale the probability of a home team scoring a goal when the opportunity arises.



```
sum(goal_dict.values())/len(goal_dict.values())
```

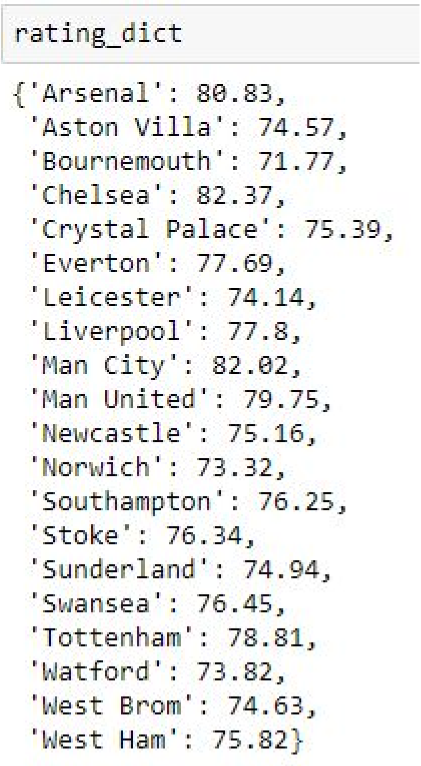
```
0.1618514466972978
```

Finally, we considered perhaps the most important factor in simulating a football match - the relevant strength of each team. From our analysis, it was clear that better players cause better results (most of the time). As such, we needed to estimate a metric to evaluate the strength of the

[1] "Modelling home advantage for individual teams in UEFA Champions ...." https://www.sciencedirect.com/science/article/pii/S2095254615001325.

[2] "(PDF) Home advantage in soccer: A retrospective analysis." 11 Dec. 2015, https://www.researchgate.net/publication/20272586_Home_advantage_in_soccer_A_retrospective_analysis.

teams involved in a match. For these values, we relied on the popular soccer game FIFA, which has a player rating system, where players are given rating ranging from 40-95. By joining these values to the line-ups in each match in the 2015/16 Premier League, we were able to derive the average rating of each team in every match. As expected, we found that better teams tend to win more often. As such, we decided to scale each micro-event in a simulated game by the ratio of the average team ratings. This helped ensure that the better team succeeds in what they are trying to do more often than the weaker team, but due to the laws of probability it is still possible for the weaker team to pull off an upset.



```
rating_dict
```

```
{'Arsenal': 80.83,
 'Aston Villa': 74.57,
 'Bournemouth': 71.77,
 'Chelsea': 82.37,
 'Crystal Palace': 75.39,
 'Everton': 77.69,
 'Leicester': 74.14,
 'Liverpool': 77.8,
 'Man City': 82.02,
 'Man United': 79.75,
 'Newcastle': 75.16,
 'Norwich': 73.32,
 'Southampton': 76.25,
 'Stoke': 76.34,
 'Sunderland': 74.94,
 'Swansea': 76.45,
 'Tottenham': 78.81,
 'Watford': 73.82,
 'West Brom': 74.63,
 'West Ham': 75.82}
```

In conclusion, we are reasonably satisfied with the estimations made, although we feel that more work needs to be done in future to prove the validity of these estimations, and to build models that better fit the available data.