This report will present the main results found in the prediction exercise.

The accuracy reported by the google team (around 77%) is not the real accuracy of a real-life experiment. Google reported the accuracy of its model after dropping draws from the sample. Therefore, their performance is overestimated. When we included draws in our sample, we achieve an accuracy of around 53% in our test sample. Our performance is better than a random guess (1/3 with a big enough sample), but not enough. Figure 1 displays the ROC curves obtained through google’s algorithm –excluding draws–.

Figure 1. ROC curves from google’s algorithm

|  |
| --- |
| ROC curve with initial data |
| F:\prediction\output\graphs\ROC_power.png |
| Notes: the blue curve (ROC 3) was generated using (i) the basic dataset and (ii) a “power” attribute. The orange curve (ROC old) includes only the former. |

The dataset used to get the performance reported includes all games (i) balanced[[1]](#footnote-1), (ii) included in Premier League (Top division from England) and La Liga BBVA (Top division from Spain). The first filter is common practice. We must have two observations per match. The second filter was implemented because odds data has information about these leagues, but not the MLS (Top division from USA) or the World Cups.

The confusion matrix obtained by our model with a threshold of 0.5 is presented in Figure 2.

Figure 2. Confusion matrix (threshold = 0.5)

|  |  |  |
| --- | --- | --- |
| Confusion matrix | | |
|  | True | False |
| Positive | 533 | 450 |
| Negative | 188 | 189 |

We find that it is easier to predict the Spanish League than the English League. For Spain, we get an accuracy of 54.3%; for the latter, 51.7%. This is reflected in the average payout. For Spain, we obtain, on average, 0.97 for each 1 dolar invested. For England, we receive 0.93. For both leagues, the average payout is 0.95.

The distribution of the payouts received is shown in Figure 3. We only display the payouts that we managed to get by correctly predicting the outcome of a match. As expected, most of our payouts are lower than 2. These kind of payouts are often allocated to the favorite –stronger– team to win the match. Nonetheless, we still manage to predict games where the favorite does not win.

Figure 3. Distribution of the payouts from matches correctly predicted

|  |  |
| --- | --- |
| 1. Distribution of the payouts | 1. Distribution of the payouts higher than 0 |
|  |  |

The predicted probability of home team winning is highly correlated with the implicit probability of the home team winning (rho = 0.89). The former value is obtained through our model and equal to the outcome predicted (y\_hat). The second probability is the inverse of the home payout, assuming it is a fair game. We know that it is not a fair game, but the loss of precision is little. We can see the evolution of these two variables for the first 50 played games.

Figure 4. Implicit and predicted probability of winning

|  |
| --- |
| 1. Probability of winning |
| F:\prediction\output\graphs\performance.png |

There we only 10 extreme events in the sample. An extreme event is defined as a match where the less likely team to win (or the one with the highest payout) actually wins. Of these 10 matches, we only predicted 1 correctly. The payout received was 4.2.

1. A balanced observation means that there is complete information about both teams participating. [↑](#footnote-ref-1)