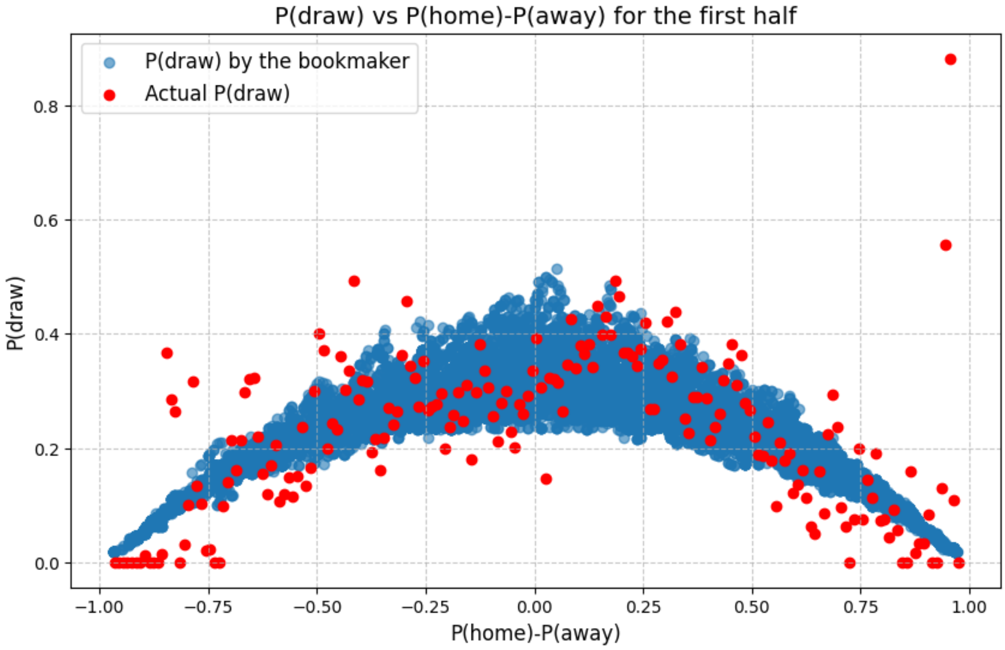
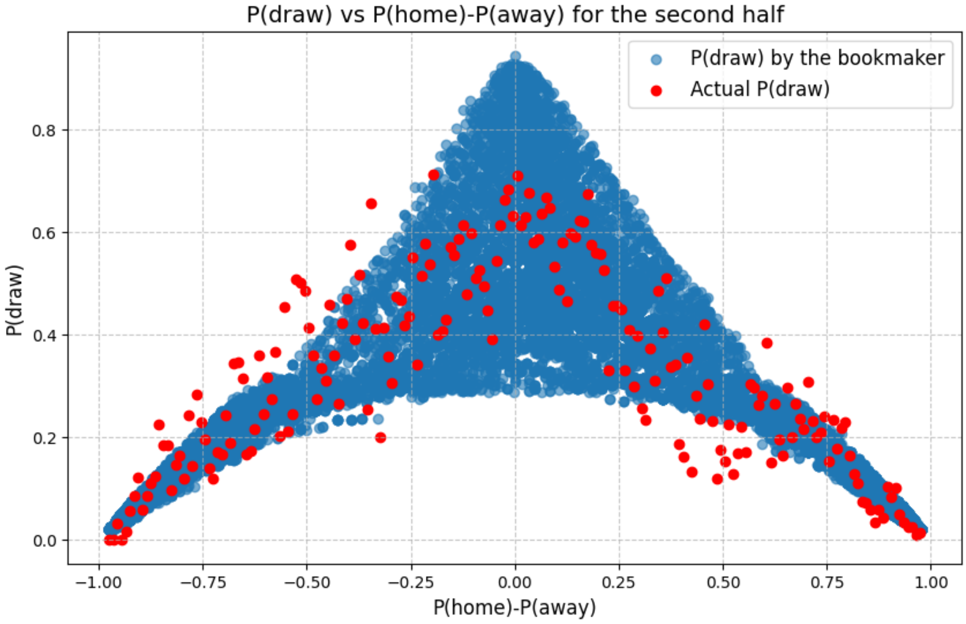
Asu Naz Yıldırım
2024708015

Homework 2

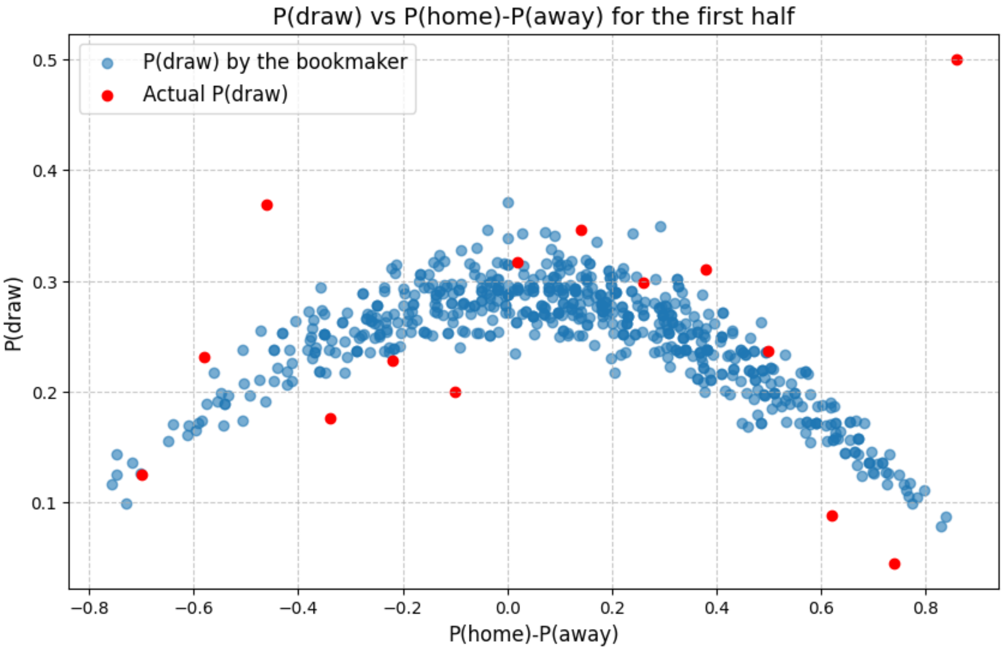
Task 1:

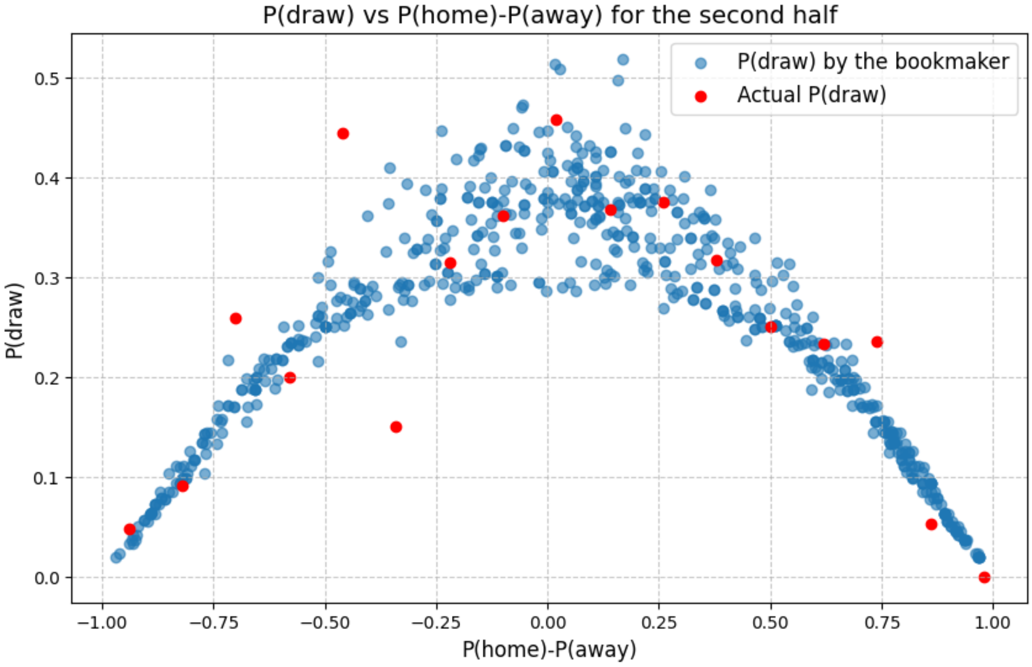




I used all the match data including in-game odds.
In the first half, as the difference P(home)-P(away) approaches near +1 or -1, the graph shows that the probability of a draw set by the bookmaker deviates from the actual probability. In other words, when one team is heavily favored, the bookmaker tends to overestimate or underestimate the actual probability of a draw: At the edges, the red points cluster either above or below the blue curve.

For the second half, when P(home) exceeds P(away), the bookmaker's P(draw) often lies above the actual P(draw) (overestimating P(draw)) especially between 0.30 and 0.60. On the other hand, when P(away) exceeds P(home), there are red points which sit below blue points, showing that actual draws happen more frequently than the odds indicate.
These trends may indicate that the bookmaker's draw probability under different match conditions show a bias.

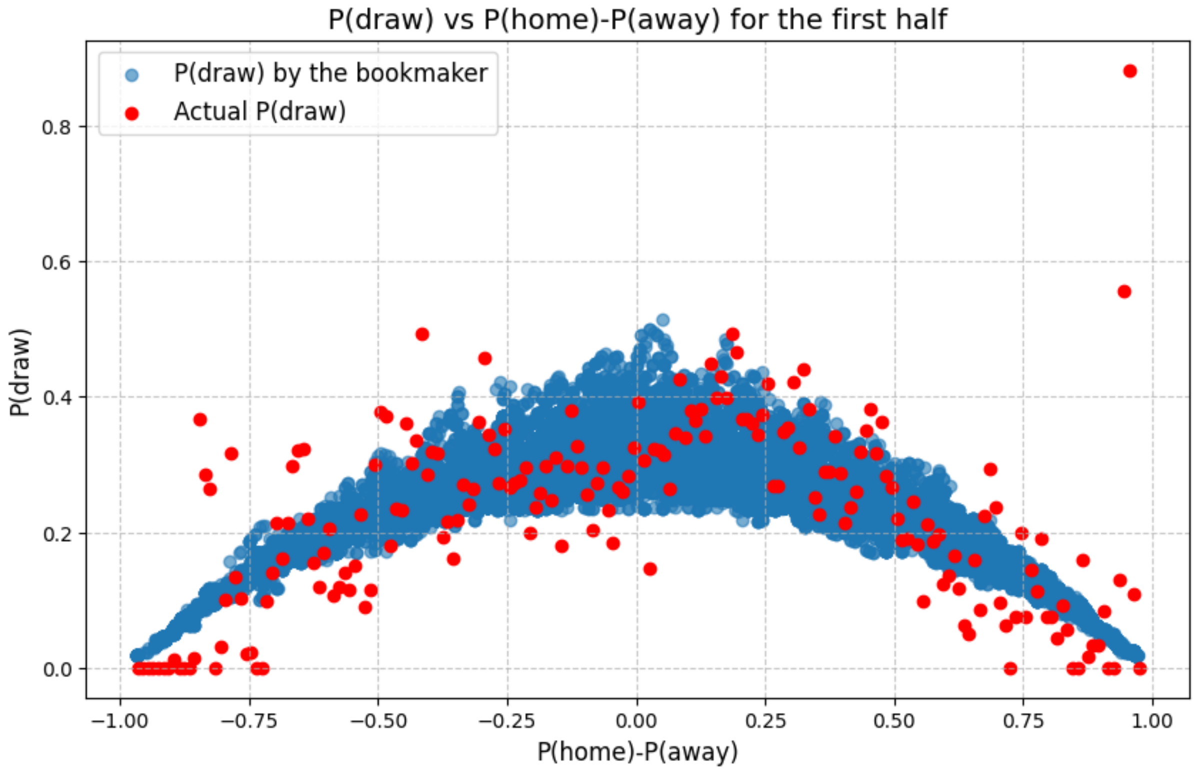


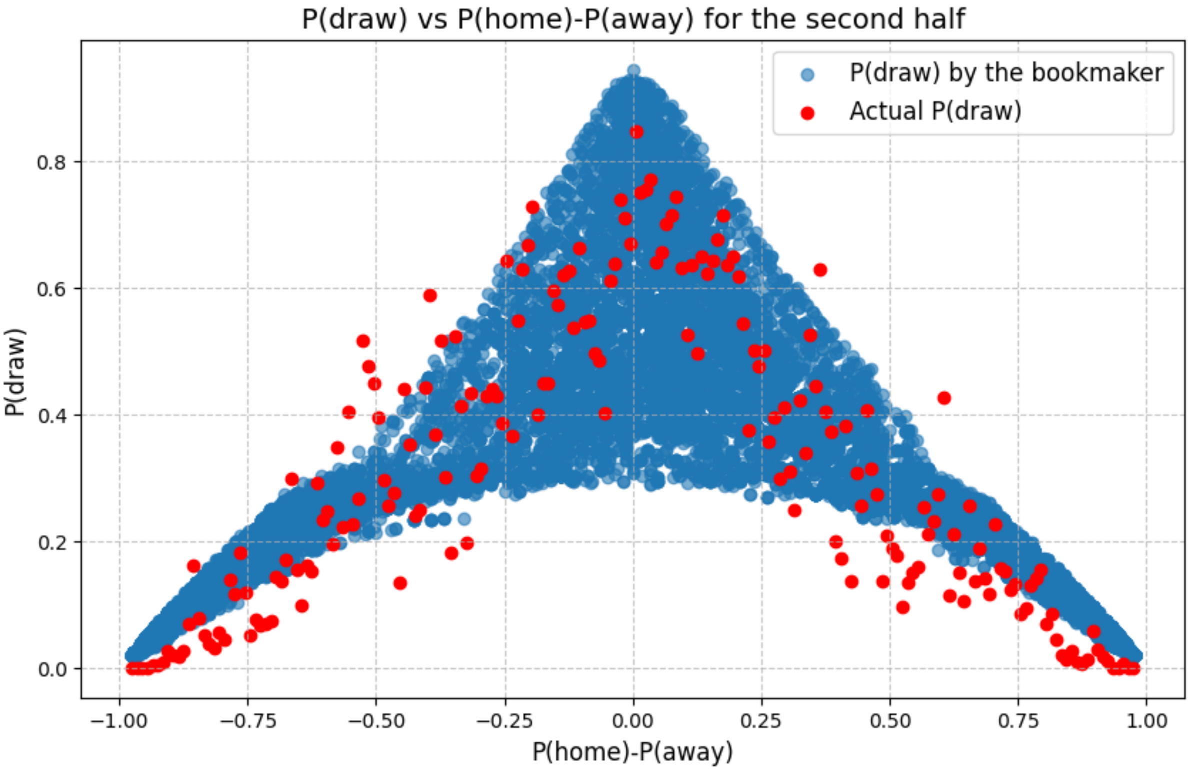


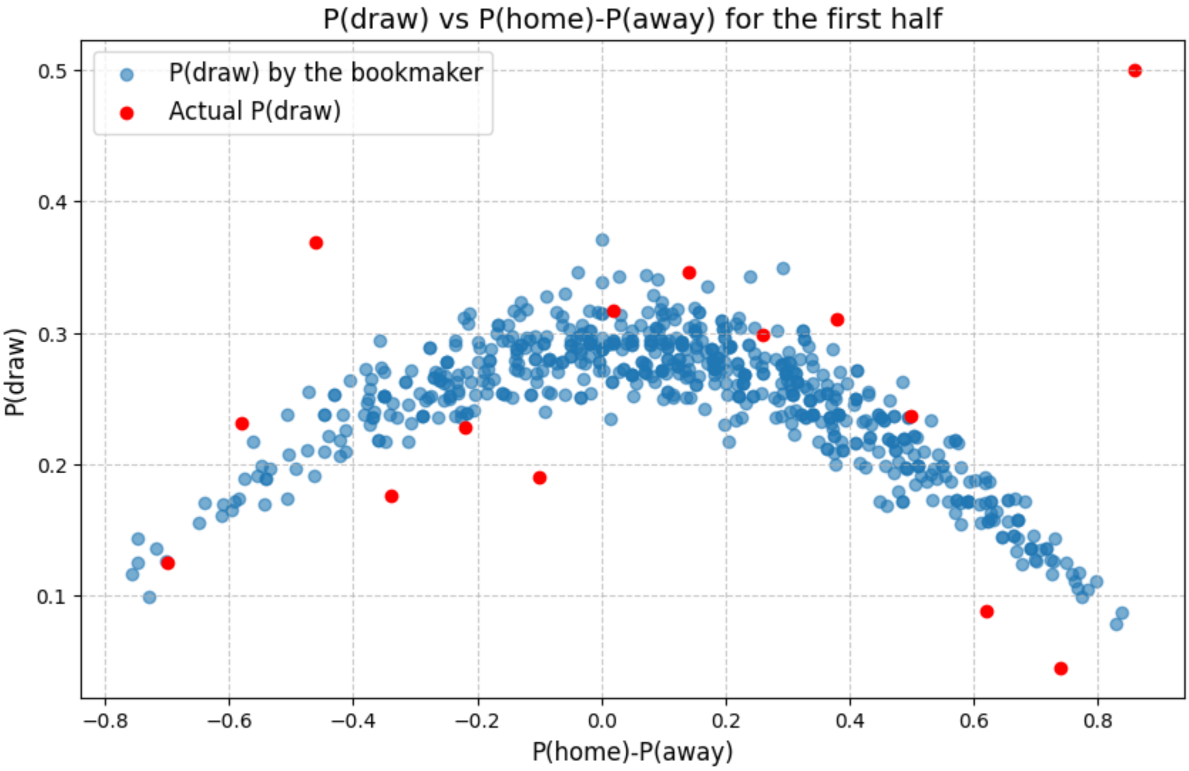
For these plots, I only used the first bet from each game to compare results. Both the first and second half plots (the blue dots) show a pattern similar to what we observed previously.

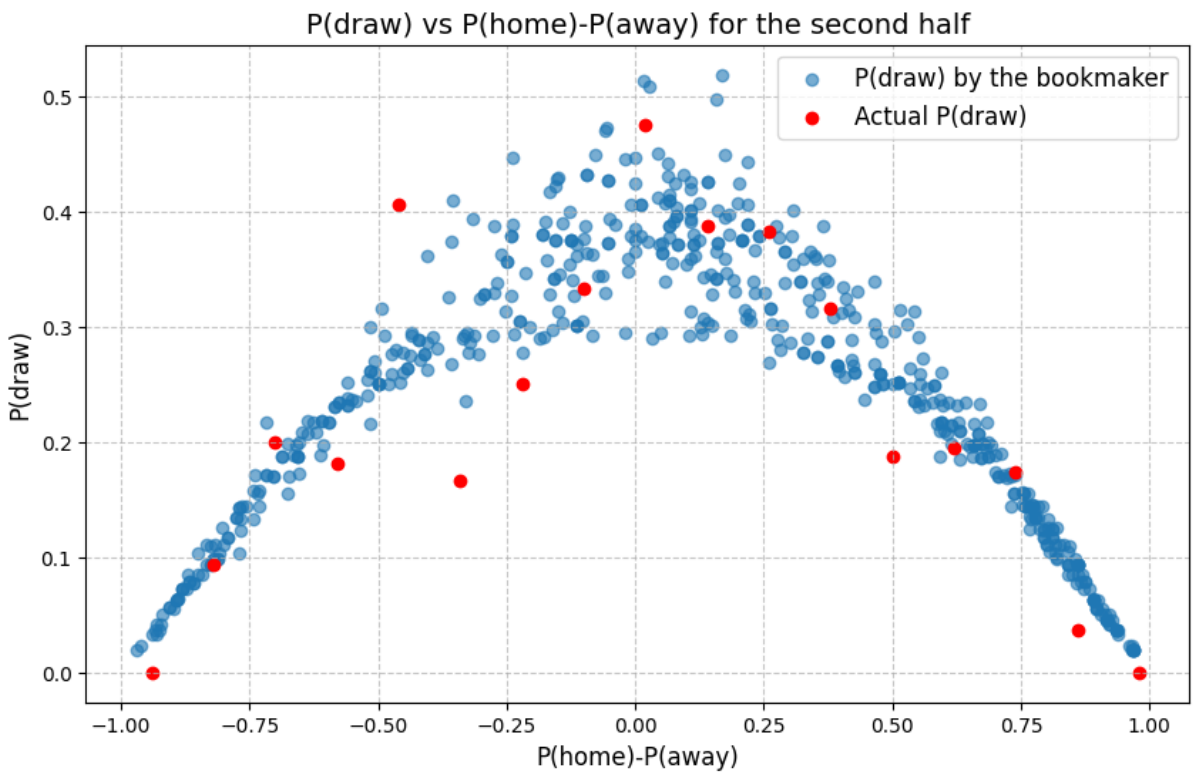
In the first half, the actual probabilities indicated by red points seem to fluctuate compared to blue dots' shape. Red dots either fall outside the bookmaker's predicted range (blue dots) or at its extremity. This may show the inefficiencies in the P(draw) set by bookmakers.
In the second half, although the red dots appear more aligned with the blue dots, some red points still lie outside the bookmakers' predicted range.

Task 2:
Number of matches removed due to early red cards: 2
Number of matches removed due to late goal affecting the result: 66



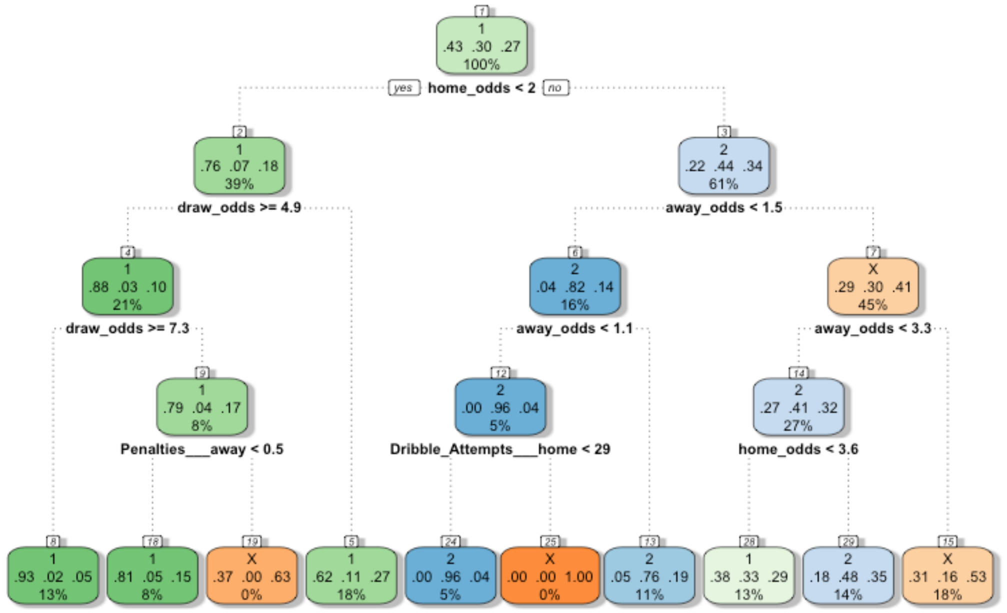

P(draw) vs P(home)-P(away) for the second half
P(draw) by the bookmaker
Actual P(draw)
P(draw)
0.0
0.2
0.4
0.6
0.8
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
P(home)-P(away)


P(draw) vs P(home)-P(away) for the first half
P(draw) by the bookmaker
Actual P(draw)
P(draw)
0.1
0.2
0.3
0.4
0.5
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
P(home)-P(away)



I remove the matches when there is a red card in the first 15 minutes or when there is a late goal affecting the bet result. Total of 68 matches are removed. Since this number is small compared to total number of instances, there is no significant change in the observation.

Task 3:



Rattle 2024-Dec-18 23:28:40 asunazyildirim

First, I include all the match related features (e.g. fouls, penalties, shots..) to train a decision tree. The splitting variables are home odds, draw odds, away odds, penalties given to away teams, dribbles attempts by home teams. Most of the nodes use odds as splitting features, which suggest market provides strong prediction for match results. Therefore, although a wide range of match performance metrices are used, odds arise as powerful indicators to predict the result. This makes sense because professional bookmakers, who have extensive knowledge, set these odds to closely reflect the actual likelihood of each game.

If we evaluate the decision tree as a whole, it produces reasonable results. The first splitting variable-value pair is home odds-2. If the home team is the favorite (<2), the home team is likely to win (1) as a result of nodes 8, 18 and 5. However, there is a splitting variable that is worth examining: penalties given to away teams. If the away team awarded a penalty in a game where a home team is the favorite, it significantly change the distribution of the result: the match is expected to end in a draw (X) in the node 19.

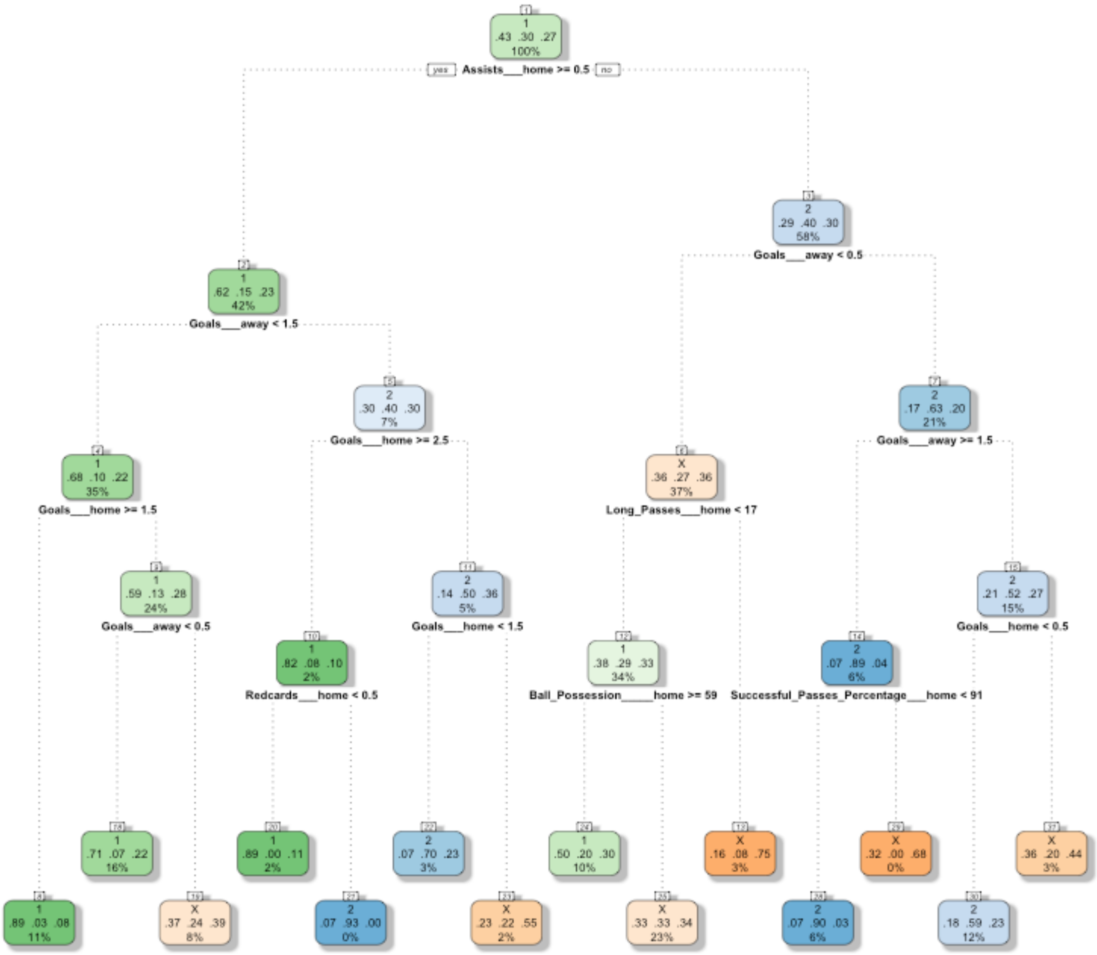
If the home odds are higher than 2 and away odds are lower than 1.5, the away team seems to be the favorite; therefore, the away team is likely to win (2). However, if the home odds are higher than 2 and the away odds are higher than 1.5, the game is open to all possible outcomes: 1, 2, and X.

To gain insight if odds are efficient or not, the values from the decision tree are examined. Some nodes show that the model's predictions are different from the odds' implied probabilities.

At node 8 where the home odds are lower than 1.975 and draw odds is higher than 7.25, the decision tree predicts a probability of 93.3% for home win, 1.5% for away win, and 5.2% for a draw. The implied probability can be estimated from the splitting variables: Draw odds of 7.25 corresponds to probability of 13.5%, while home odds of 1.975 to 50.6% (before adding bookmarker margin). Therefore, there exists a gap between model's probabilities of node 8 and estimated implied probabilities. In this case, the market might underestimate how dominant the home team is.

When the home odds are lower than 1.975 and draw odds is lower than 7.25 with at least one penalty awarded to the away team (>0.5), the decision tree predicts a probability of 36.7% for home win, 0% for away win, and 63.3% for a draw. Although the home team is the favorite (home odds < 1.975 corresponding to the probability of 50.6%), the award of a penalty to the away team significantly changes the model’s prediction towards a draw. In this scenario, the home odds appear inefficient.

As a result, certain nodes may indicate potential inefficiencies in the odds.



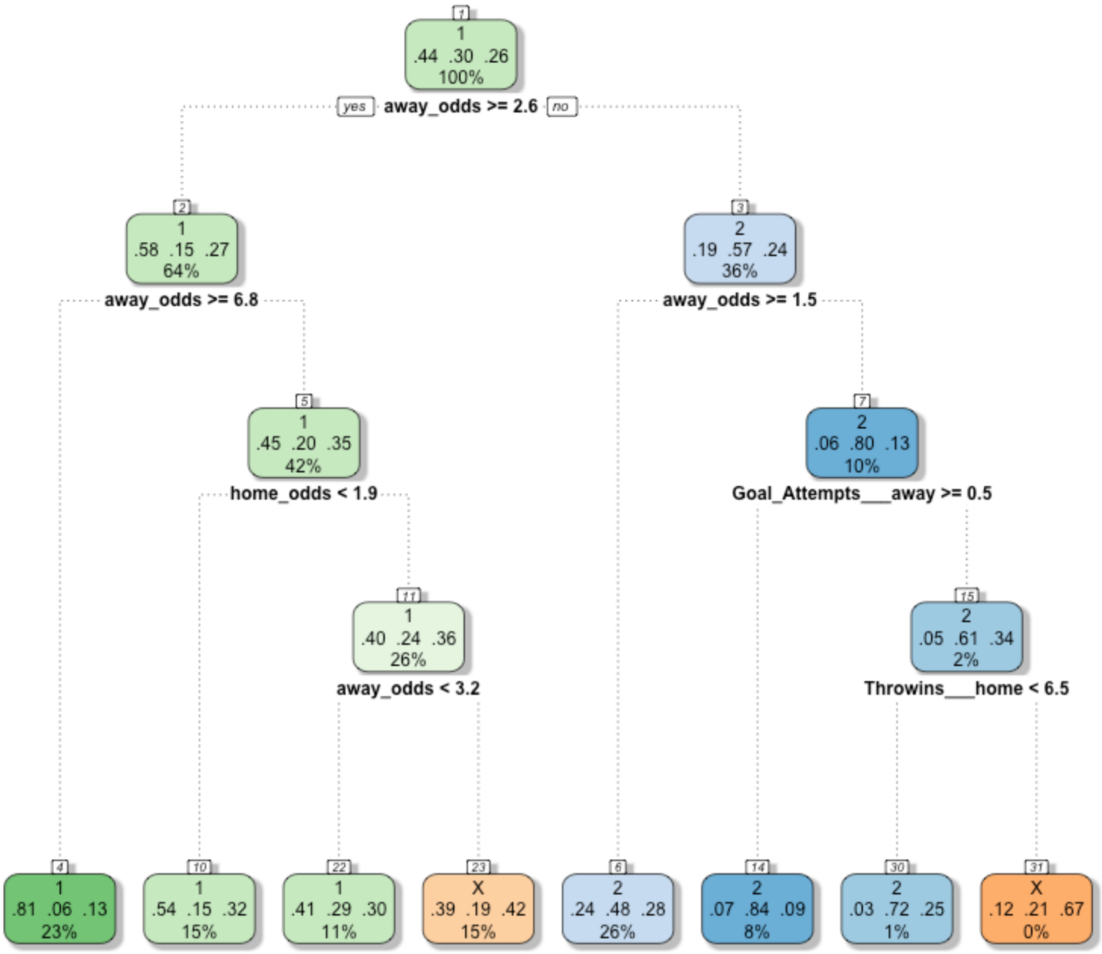
Rattle 2024-Dec-16 02:11:48 asunazyildirim

I train a second decision tree by excluding home, away, draw odds to examine how match-related variables affect the outcome of the game. I want to see which match-related metrices arise as splitting variables when the market information is absent. Under this condition, the splitting variables are Assists - home, Goals - away, Goals - home, Red cards - home, Long - Passes - home, Ball Possession -home, Successful Passes Percentage - home.

If we evaluate the decision tree as a whole, it produces reasonable results. Assist and goals are key features that directly influence the outcome of the game; therefore, it makes sense that they are key predictors in the decision tree. The tree uses Goals - away and Goals - home multiple times as splitting variables, demonstrating that goals are the most straightforward indicator of a game's progress.

The decision tree first uses goals and assists as indicators to determine whether the home team is creating good scoring chances. To illustrate, if the home team has one or more assists (Node 1), the home team has a strong chance of winning (Nodes 8,18 and 20). After looking at the home team's offensive performance, it checks if the away team can respond. Multiple away goals or home red cards may result in an away win (Node 21) or a draw (Node 23). Without odds, factors such as ball possession, successful passes percentage, and red cards become more significant. These factors affect match outcomes in cases where the results are more balanced or less obvious.

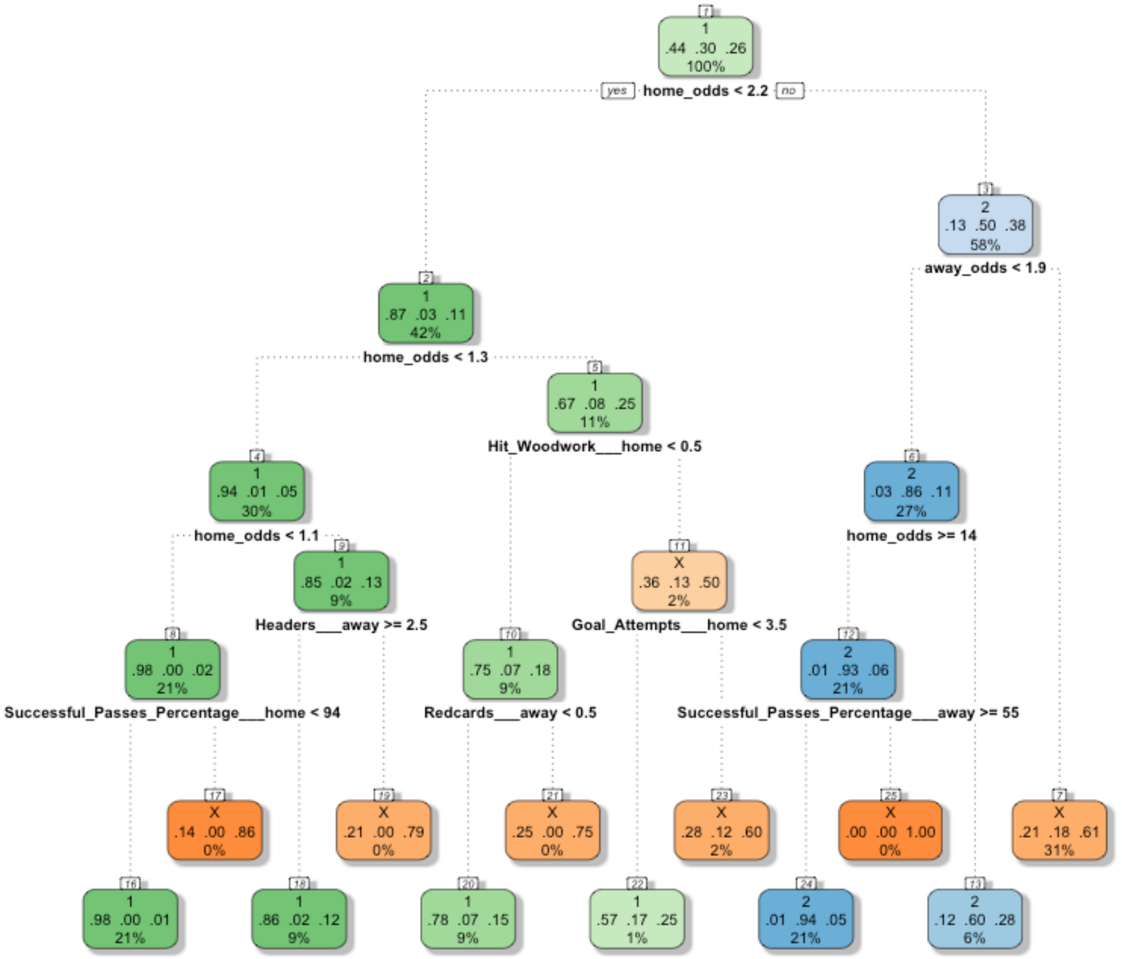
Since I removed the odds, it has become difficult to determine whether they are efficient or not. When the odds are absent, the model only considers match conditions and ignores how the market assessed such conditions.



Rattle 2024-Dec-19 18:01:27 asunazyildirim

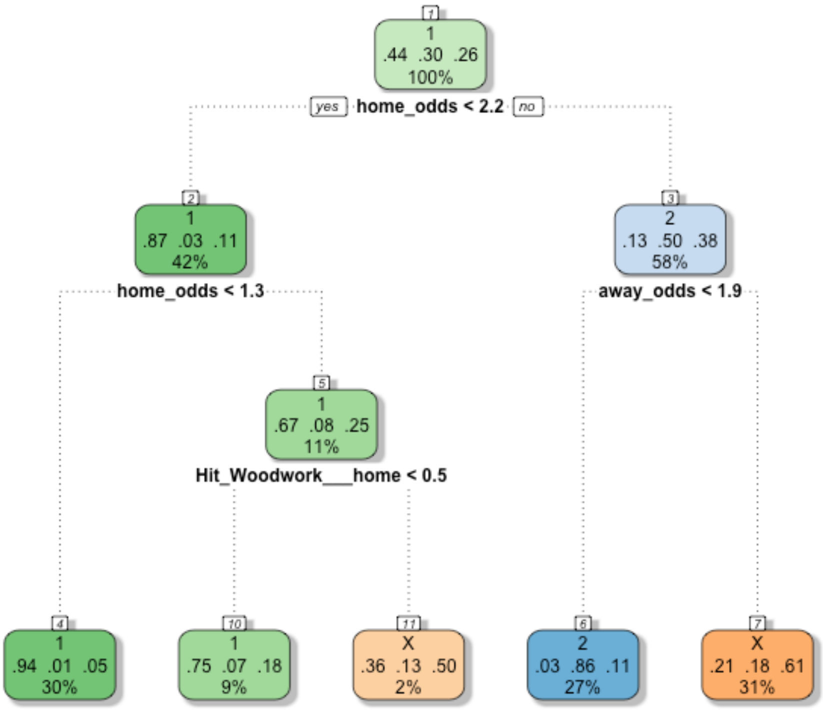
I trained a decision tree for the first half of the data which excludes the matches where a red card was given in the first 15 minutes or where a late goal influenced the outcome. In this decision tree, the odds (specifically, away team odds and, to a lesser extent, home team odds) remain the most important splitting variables. The first splitting variable (away odds) provides insight into which team is the favorite. In the left part of the decision tree, when away odds are very high (>6.75), the model predicts a home win, which is reasonable. When away odds are between 2.61 and 6.75 (Node 5), there is a possibility in the draw depending on the home odds (Node 23). In the right part of the decision tree, where away odds are lower than 2.6, the match most often results in an away win, which makes sense. However, in more balanced matches,

Goal Attempts – Away and Throw-ins - Home arise as key indicators, with the model predicting a draw at Node 21.
To examine the inefficiencies in the odds, Node 4 is examined. The decision tree predicts home win probability of 80.95%, away win probability of 5.81% and draw probability of 13.24%. However, the implied probability of away win from the away odds of 6.75 correspond to 14.81%. Therefore, model's prediction is lower than the implied probability, which may create inefficiencies.



Rattle 2024-Dec-19 20:42:34 asunazyildirim

I trained a decision tree for the second half of the data which excludes the matches where a red card was given in the first 15 minutes or where a late goal influenced the outcome. Since the are several nodes with 0% of the data. I reset the maximum depth to 3 to obtain a simpler model.



Rattle 2024-Dec-19 20:45:11 asunazyildirim

In this decision tree, the first splitting variable is once again home odds. When home odds are lower than 2.15, the model strongly favors the home team to win. For home teams that are slightly less favored, home odds are between 1.29 and 2.15, a new splitting variable arises to predict whether the game will result in a home win or a draw: hit woodwork - home. In this decision tree, if the home team hits the woodwork multiple times, the model interprets it as an unlucky scenario and predicts the result as a draw (X).

On the other hand, if home odds are higher 2.15, which shows that the home team is not strongly favored, the model uses away odds to predict the result. When away odds are lower than 1.92, the away team is the favorite and likely to win. Otherwise, the match is balanced and the result is draw.

Reference:

I used ChatGPT in my code. In case of need, I can provide my prompt.