Predictive Modeling

Submission Deadline

Please submit your completed results within 7 days of receiving this assignment. Your results can be submitted to the HR representative who provided you the assignment.

Assignment

Sports betting is a 500 billion dollar market, is played by 250 million players in over 200 countries (most popular sport globally). The English Premier League is the most popular domestic team in the world. It contains 20 years of EPL matches dataset.
The dataset includes train (train.csv), test (test.csv), submission template (sample_submission.csv), and results (train_target_and_scores.csv).

Dataset Description

Descriptive columns

- target - The variable you have to predict the probabilities only available in the train set.
- home_team_name - The name of the **Home** the team. Hidden in test set, see this discussion
- away_team_name - The name of the **Away** the team. Hidden in test set, see this discussion
- match_date - The match date (UTC).
- league_name - The league name.
- league_id - The league id. Note that league names can be identical for two differents id.
- is_cup - If the value is 1 the match is played for a cup competition.
- home_team_coach_id - The id of the **Home** team coach.
- away_team_coach_id - The id of the **Away** team coach.

Historical columns

Historical columns contain the key world history and are indexed with numbers from 1 to 10 for the last 10 matches played by the two teams. The number **1** means the most recent while **10** is the oldest. Note that the historical matches can have been played in different leagues.
For instance home_team_history_match_date_1 is the date of the last match played by the home team while home_team_history_match_date_2 is two matches ago. Another example, home_team_history_is_play_home_3 tells you if the team played home 3 matches ago and home_team_history_opponent_rating_3 what was the opponent rating.
Historical home team features

- home_team_history_match_date_{i} - The date of the last i-th match played by **Home** team.
- home_team_history_is_play_home_{i} - If 1, the **Home** team played home.
- home_team_history_is_cup_{i} - If 1, the match was a cup competition.
- home_team_history_goal_{i} - The number of goals scored by the **Home** team on its last i-th match.
- home_team_history_opponent_goal_{i} - The number of goals conceded by the **Home** team on its last i-th match.
- home_team_history_rating_{i} - The rating of the **Home** team on its last i-th match (pre match rating).

- home_team_history_opponent_rating_{i} - The rating of the opponent team on **Home** team last i-th match (pre match rating).
- home_team_history_coach_{i} - The coach id of the **Home** team on its last i-th match.
- home_team_history_league_id_{i} - The league name id by the **Home** team on its last i-th match.

Historical away team features

- away_team_history_match_date_{i} - The date of the last i-th match played by **Away** team.
- away_team_history_is_play_home_{i} - If 1, the **Away** team played home.
- away_team_history_is_cup_{i} - If 1, the match was a cup competition.
- away_team_history_goal_{i} - The number of goals scored by the **Away** team on its last i-th match.
- away_team_history_opponent_goal_{i} - The number of goals conceded by the **Away** team on its last i-th match.
- away_team_history_rating_{i} - The rating of the **Away** team on its last i-th match (pre match rating).
- away_team_history_opponent_rating_{i} - The rating of the opponent team on **Away** team last i-th match (pre match rating).
- away_team_history_coach_{i} - The coach id of the **Away** team on its last i-th match.
- away_team_history_league_id_{i} - The league name id played by the **Away** on its last i-th match.

What are the rating features?

The rating features are calculated by Octosport. Ratings are meant to give information on the team's relative strength for a given match. For instance, we would expect that a team with a rating of 10.5 beats a team with a rating of 2.3.

Reference

[1] Football Match Probability Prediction. https://www.kaggle.com/competitions/football-match-probability-prediction/overview

Question

1. Please train a model to predict the results of matches in test.csv. The output format is the sample in sample_submission.csv
2. Based on the historical data, write code to include results of last 5 matches between two teams.
3. We need to predict the probability of scores. For example, given a match data, output the probability of 3-1 score. Describe your idea/code to sovle the proble.

Deliverables

The following deliverables must be submitted to Thinkprompt:

- Your code Repo (Optional)
- Prediction result (for Task 1)
- Task 2 and 3 in PDF format