IE 582 Fall 2019 Project, Due final exam

Objective and Evaluation

Your first homework provides the details regarding the sports forecasting and the data. The aim of the project is to provide better forecasts for 1x2 bets (i.e. home, tie, away) compared to bookmakers using the odds data and other type information you can come up with (i.e. features related performance of the teams for the last 5 games, day of the week for the match and etc.). One can model this as a classification problem with three labels and use any classification approach to generate predictions for probability of home win, tie and away win. Although classification approaches can work, this problem is an ordinal regression problem because of the nature of the target class. Think of a match where the score is 1:0 (i.e. home teams is ahead with one goal margin), the result of the game cannot be "away win" without being "tie" (i.e. in order to be an away win with a score of 1:2, we need to have 1:1). Hence, taking the ordinal nature of the class into account can improve the models (although it is not directly covered in class).

Use the data on the Google Drive link (same link as your first homework) which will be updated regularly (everyday around 00:00). The link is below:

https://drive.google.com/open?id=10ubZ9Qb9j1EXQ3xKF3Sg_Bc-AP68XFgS

You can remember the details of what is provided in the document of first homework. Basically, the folder contains:

- One file for the matches and the other one for odd details.
    - Matches data contains the information for both matches from the past and upcoming matches (for which score is NA)
    - Odd details data contains the odd information for different type of bets from multiple bookmakers changing over time.
- Submission schedule in Excel format.

In this project, you are expected to build a model based on the methods in the course to provide predicted probabilities for the outcomes of the upcoming matches. You are also free to use extended versions of the approaches covered in the lectures. Note that there are three classes: "home", "tie", "away" which makes this problem multi-class classification problem (or an ordinal regression problem as discussed).

You are expected to provide your predictions based on the schedule provided as Excel file on Google Drive link (name of the file is submission_schedule.xlsx). Matches to be forecasted is also listed for each round (8 rounds in total). The submission format is described in the later sections.

Performance measure:

Ranked Probability Score (RPS): "is a measure of how good forecasts, expressed as probability distributions, match with observed outcomes. Both the location and spread of the forecast distribution are taken into account in judging how close the distribution is to the observed value".

$$\mathrm{RPS} = \frac{1}{r-1}\sum_{i=1}^{r}\left(\sum_{j=1}^{i}p_j - \sum_{j=1}^{i}e_j\right)^2,$$

where r is the number of outcomes, p_j is the forecasted probability of outcome j and e_j is the actual probability of outcome j. $r = 3$ in the case of 1x2 bets. Note that this score is defined for each match. The primary measure to be used for evaluation is "Average RPS" of the predicted games of the corresponding week. Below you can find examples of RPS calculation for two different predictions for a game results with "home" win. The second prediction is better in terms of RPS as "away" win probability was predicted to be smaller for a game finished as "home" win. This measure captures the ordinal nature of the target.

P(Home)	P(Draw)	P(Away)	RPS
0.5	0.2	0.3	$(0.5 - 1)^2 + (0.7 - 1)^2 + (1 - 1)^2 = 0.34$
0.5	0.3	0.2	$(0.5 - 1)^2 + (0.8 - 1)^2 + (1 - 1)^2 = 0.28$

Prediction and Report

You will submit your predictions through Google Forms from the link below:

https://goo.gl/forms/G60HutYl66AX4oIs1

You will enter your ID, token and class probability predictions to relevant place.

ID: Your team id (sent via email)

Token: Your team token (sent via email)

Predictions: Class probability predictions. Note that there are 3 classes in this task. Suppose the number of test observations is NTEST, then your class probability predictions are expected to be stored in a "NTEST x 3" matrix. Because of the restrictions of Google Forms, you are required to provide your predictions in a single row where each entry is separated with comma. Please provide your predictions as in the following form:

MatchId for 1st Observation, 1st Observation Prob. for "home" Class, 1st Observation Prob. for "tie" Class, 1st Observation Prob. for "away" Class, MatchId for 2nd Observation, 2nd Observation Prob. for "home" Class, 2nd Observation Prob. for "tie" Class, 2nd Observation Prob. for "away" Class,…….

Any submission not following this format will return an error and will not be evaluated. We have provided sample codes for RPS calculation and submission formatting in the repository on the following link: https://github.com/BU-IE-582/fall18-instructor. If you fail to provide the predicted probabilities for certain matches, your RPS for the corresponding match will be defaulted to one. Moreover, if your predicted probabilities do not sum up to one, we will follow the same strategy and set RPS to one for the corresponding matches.

The schedule provides the time of the match and you have to provide your predictions at least 1 hour before the first game of the round. Otherwise, your submission will not be evaluated. Note that provided date/time is for GMT (+1). Our time zone is GMT (+3). If you make more than one submission (it is possible with Google Forms), the latest submission will be counted as your final submission.

Your final score will be determined by the performance of your proposed approach (i.e. average RPS over the games). The evaluation will be done using the performance of six rounds with best score. Hence, you are free to skip submission for 2 rounds out of 8. However, this may deteriorate your final score as more submission will increase your chance to obtain better final RPS. Note that 30% of your project grade will be determined by your final rank in this competition. First place will get full points (30 points) and this will decrease to a minimum of 15 points proportional to your deviation from the top performer. Note that you must submit for at least 6 rounds. If you fail to satisfy this condition, 3 points will be deducted from your final score for each skipped round. You are allowed to work as a group of at most 3 members.

Your report should have the following format:

1. *Introduction:* Problem description, summary of the proposed approach, descriptive analysis of the given data.
2. *Related literature:* Summarize relevant literature if there is any
3. *Approach:* Explain your approach to this problem.
4. *Results:* Provide your results and discussion.
5. *Conclusions and Future Work:* Summarize your findings and comments regarding your approach. What are possible extensions to have a better approach?
6. *Code:* Provide the Github link for your codes at the end of your report.