IE 582 Statistical Learning for Data Mining

Project, due January 13th, 2025

Task

In this project, you will propose a live betting strategy based on the provided soccer match data, emphasizing the dynamic nature of in-game odds as indicators of market beliefs. The objective is to identify a single optimal time instant during a game to make a prediction about the match outcome (i.e., result after the game finishes)—home win, draw, or away win, together with a decision "bet home win", "bet draw", "bet away win" or to choose "no action". In other words, you are expected to provide a prediction for match outcome (all matches) but your decision can be "no action" based on your prediction.

A crucial aspect of this task is that you are allowed to make only one decision per match at a specific time instant. Once a decision is made, no further predictions can be made for that match. To clarify: if you decide to make a prediction at the 15th minute, you cannot use information beyond that point to revise or add predictions for the same match. This constraint reflects a real-life scenario, where decisions must be made based only on the data available up to the chosen moment. Another common pitfall to avoid is "forward-seeing"—using future data (such as odds or statistics) to inform decisions made at earlier time points. In real-world applications, future information is not accessible at the time of decision-making. Your predictions and strategy should be designed accordingly. Please use the matches starting from "2024-11-01" (included) as your test data (total of 111 games). You are free to use the matches before this date for training and tuning purposes.

Your decision-making process must be grounded in a thorough analysis of the provided data, which includes match statistics, odds, and their real-time fluctuations. The challenge lies in combining statistical insights with practical decision-making while accounting for dynamic, game-changing events such as goals, red cards, and substitutions.

Performance Measures:

The effectiveness of the proposed strategy will be evaluated based on:

- Accuracy: The proportion of the matches for which your method correctly predict the match outcome.
- Return: The profitability of the strategy, measured as the cumulative return from betting 1 unit per match across the dataset.

This task encourages students to integrate their understanding of descriptive and predictive analyses while tackling a real-world problem in sports analytics.

Report & Code Documentation

You are also expected to report your approach and findings as a document. 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.

You can work with any language you want (i.e. R, Python, Julia, Matlab and etc.). You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals. There are alternative ways to generate an html page for you work:

1. A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
    a. https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/
    b. https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps917519326e41

    Things are little easier if you install Anaconda (https://www.anaconda.com/). Please export your work to an html file. Please provide your *. ipynb file in your repository and a link to this file in your html report will help us a lot.

2. A Markdown html document. This can be created using RMarkdown for R and Python. Markdown for Python

Note that html pages are just to describe how you approach to the exercises in the homework. They should include your codes. You are also required to provide your R/Python codes separately in the repository so that anybody can run it with minimal change in the code. This can be presented as the script file itself or your notebook file (the one with *.ipynb file extension).

Academic Integrity

The last and the most important thing to mention is that academic integrity is expected! Do not share your code (except the one in your progress journals). You are always free to discuss about tasks but your work must be implemented by yourself.

Homework assignments in this course permits the use of GenAI tools. Any such use must be appropriately acknowledged and cited. It is each student's responsibility to assess the validity and applicability of any GenAI output that is submitted; you bear the final responsibility.

For all other forms of assignments, quizzes and exams use of GenAI tools is disallowed.