IE 582 Statistical Learning for Data Mining
Homework 2, due December 20th, 2024

1. Introduction
Sports forecasting is important for sports fans, team managers, sponsors, the media and the growing number of punters who bet on online platforms. Widespread demand for professional advice regarding the results of sporting events is met by a variety of expert forecasts, usually in the form of recommendations from tipsters. In addition, betting odds offer a type of predictor and source of expert advice regarding sports outcomes. Whereas fixed odds reflect the (expert) predictions of bookmakers, the odds in pari-mutuel betting markets indicate the combined expectations of all punters, which implies an aggregated expert prediction.

Expert forecasts of sport outcomes often come from so-called ‘tipsters’, whose predictions appear in sports journals or daily newspapers. Tipsters are usually independent experts who do not apply a formal model but rather derive their predictions from their experience or intuition. They generally provide forecasts for only a specific selection of games, often related to betting. No immediate financial consequences result from the predictions of tipsters. Empirical evidence regarding the forecast accuracy of tipsters shows that their ability is limited.

In sports forecasting, real-time information plays a significant role. Among the many factors influencing match outcomes, in-game odds stand out as a dynamic indicator of market beliefs and trends. These odds reflect ongoing assessments of match outcomes—home win, draw, or away win—as the game progresses, and their fluctuations show the collective judgment of bettors reacting to the game’s developments.

Betting behavior during live events is shaped by various factors, including emotional reactions, game statistics, and market-driven information. People place their bets based on their interpretation of game dynamics, such as changes in possession, goals scored, or player performance. This dynamic interaction between market participants drives frequent updates to the odds, which are adjusted to balance bookmaker risk while reflecting real-time market sentiment.

<u>In-Game Odds and Market Beliefs</u>

Unlike pre-match odds, which are determined based on historical data, expert opinion, and statistical models, in-game odds continuously evolve as the match unfolds. For example, a goal scored by the home team will immediately increase the probability of a home win, reducing the associated odds, while simultaneously lowering the probabilities (and increasing the odds) of a draw or away win.

This real-time adjustment reveals implicit information about the market's belief in the likelihood of different outcomes. Odds are influenced by:

- Game Events: Goals, red cards, injuries, and substitutions can significantly shift the odds.
- Betting Trends: A surge in bets on a particular outcome may prompt bookmakers to adjust the odds to balance their liabilities.
- Market Sentiment: Collective interpretation of game statistics, such as possession, shots on target, and momentum, drives the market’s expectations.

<u>Why Odds Change</u>

Odds changes are not arbitrary; they are a function of market forces and bookmaker strategies. Bookmakers aim to maintain balanced books by ensuring that the payout for any outcome is covered by the bets placed on other outcomes. Changes in odds reflect:

- Market Trends: A significant increase in betting volume on one outcome can lead to reduced odds for that outcome and increased odds for others.
- Updated Beliefs: Live data, such as statistical trends and game events, updates the implied probabilities of outcomes.
- Bookmaker Risk Management: Adjustments are made to mitigate risk and ensure profitability regardless of the match result.

In-Game Odds as Market Indicators

The movement of odds provides valuable insights beyond the match itself. For bettors and analysts, odds fluctuations serve as a proxy for market trends and collective sentiment. For example: A sudden drop in odds for the home team after an impactful event (e.g., a key player substitution or goal) suggests a shift in market belief toward a home win. Persistent odds for a draw despite game developments might indicate a balanced market perception of the teams' performance.

By analyzing these trends, bettors can make informed decisions, and analysts can use odds as a real-time signal of how the market processes and reacts to game information. This interplay between statistical trends, human intuition, and bookmaker adjustments makes in-game odds a fascinating subject for research and application in predictive modeling.

2. Background

The technical report of Mirza and Fejes [1] provides a good description of how betting odds are determined by betting companies. Based on the statistical analyses of the odd information, their aim is to predict the outcomes of the English Premier League soccer games. http://betamatics.com/ is the website they share their predictions online and details of their approaches are available both in their technical report and the website.

Here is a background information about how odds are determined:

"There are plenty of different scenarios that one can bet on when it comes to sports. In this project, only bets of the type "singles" in Premier League were analyzed. A single bet is a bet placed on just one selection. In football that yields win, draw or loss (1, X, 2), from a home team point of view. A typical single bet can look something like (1.72, 3.80, 4.50) which means one have a chance to win 1.72 times the money if betting on home win and so on.

So how do the bookmakers set the odds? If gambling had been a fair game the odds should correspond to the estimated probability for the outcome they represent. In this case home win will give 1.72 the money and therefore the probability for it would be its inverse 0.58. However, this is not the case and a simple example can show why. If one takes the inverse and sums up the probabilities for all the outcomes in one game one expects the sum to be equal to one, but for the bets stated above the sum is 1.07 which means there is a 7% margin added by the bookmakers. Further on, the bookmakers have no real interest in predicting the outcome themselves."

Štrumbelj [2] also provides some insights into how odds are useful.

Odds and Probabilities

The odds are generally given in a format so called "European style" in the gambling community, which for a fair (no-margin) bet is given as odds = 1/P(win) as described in the background. Bookmakers generally set their odds based on the expert opinion or using a statistical model. Therefore there is always possibility that the odds may not be the best possible prediction of the match outcomes. Assuming that the odds represent those given by a naive bookmaker who has predicted the match outcomes to her best, the odds can be set as the reciprocal of the probability, and scaled them down by some percentage to take a revenue only on the winning bets. Then the implied probabilities become:

$$\begin{bmatrix} \mathrm{P(home)} \\ \mathrm{P(draw)} \\ \mathrm{P(away)} \end{bmatrix} = \begin{bmatrix} 1/\mathrm{odds}_1 \\ 1/\mathrm{odds}_X \\ 1/\mathrm{odds}_2 \end{bmatrix} \cdot \frac{1}{\sum_{i \in \{1,X,2\}} 1/\mathrm{odds}_i},$$

where the normalization (second term where we divide probabilities by the sum of probabilities) is needed to remove the margin from the odds. If the match results were to be distributed exactly by these probabilities, we would always lose in the long run due to the bookmaker's margin. On the other hand, Štrumbelj [2] considers a different transformation approach based on the idea of Shin [3] (i.e. Shin probabilities).

3. Data

You can find "match_data.zip" on the Moodle course page. The data includes match details, statistics, and odds for the final result of the match. Each row represents an instant in a match. Data is mostly available in minute granularity. However, there might be missing minutes (even a missing halftime of data). It is also possible to see multiple rows for the same minute and the same match while the seconds of the rows will be different.

Column Descriptions

- 'fixture_id': Unique ID of the specific match.
- 'halftime': Half of the match in which the minute occurs.
- 'current_time': Datetime of the row.
- 'half_start_datetime': Halftime start datetime of the row.
- 'match_start_datetime': Match start datetime of the row.
- 'minute': Minute of the row. The minutes are counted from the beginning of the halftime. Here are some examples to better understand:
  - If 'halftime' is "1st-half" and minute is 49, then it is 45+4th minute of the half and the match.
  - If 'halftime' is "2nd-half" and minute is 7, then it is the 7th minute of the second half or 52nd minute of the match.
  - If 'halftime' is "2nd-half" and minute is 50, then it is 45+5th minute of the second half or 90+5th minute of the match.
- 'second': Second of 'minute'
- 'latest_bookmaker_update': Last time the odds were updated by the bookmaker.
- 'suspended': Whether the odds are suspended or not. It might have several reasons (penalties, fouls, goals, bookmaker's adjustments etc.). Odds provided in the rows in which "suspended" values are true must be disregarded.
- 'stopped': Whether the odds are stopped. Odds provided in the rows in which "'stopped'" values are true must be disregarded.
- '1': Odd for "home wins".
- '2': Odd for "away wins".
- 'X': Odd for "draw".
- 'name': Names of the home and away teams.
- 'ticking': Indicates whether the match is still in progress at the given time.
- 'current_state': Current status of the match ("1": Home team is winning, "2": Away team is winning, "X": Draw).
- 'final_score': Final score of the match (home_score-away_score). This value is the same for all minutes for a given match.
- 'result': Final status of the match ("1": Home team is winning, "2": Away team is winning, "X": Draw).

The rest of the columns are the statistics of the match for the given time. The column names are self-explanatory.

4. Tasks

The aim of this homework is to get you familiar with the data you will deal with in your project. It involves certain descriptive analyses to understand the data. Please note the following before performing tasks:

- You should not use odd columns (“1”, “X”, and “2”) in rows where “suspended” values are true or “stopped” values are true.
- The data might contain missing values or rows. Don't let these lead you to misleading conclusions in your analysis.

Task 1

The aim of this task is to understand if bookmakers are good enough in setting their odds for “draw” bets. An empirical evidence for the probability of “draw” can be calculated by determining the certain probability intervals on the implied probabilities by the bookmakers for the specific result. Once you determine a probability range (i.e. a bookmaker’s implied draw probability is 0.4 for a specific game and your probability range is 0.38 and 0.42), you can count the games that finished as draw within this range. In other words, we can discretize probability of draw values into bins (i.e. (0.00,0.05], (0.05, 0.10], …, (0.95,1.00]) and calculate the number of games ended as “draw” in the corresponding bin. Dividing this value by the total number of games in the corresponding bin will provide the estimated probability of “draws”. Please note that implied probabilities may not be larger than a certain value (since it is not reasonable), modify your bins accordingly if this is the case. Aforementioned bins are provided for illustration purposes. If people are good enough in in their beliefs, what you expect to see is that fraction of games finished as “draw" is between this implied probability range. You will perform the following tasks for each half of the game.

1. Calculate the P(home win), P(tie) and P(away win) by P(x) = 1/odd.

2. Then calculate these probabilities again using normalization formula at “Odds and Probabilities” part.

3. First construct a plot of P(home win) – P(away win) on x-axis and P (tie) on y-axis with first probability calculation; then plot the actual probabilities calculated using the results.

   In other words, we can discretize P(home win) – P(away win) values into bins (i.e. (-1,-0.8], (-0.8, -0.6], …, (0.8,1]) and calculate the number of games ended as “Draw” in the corresponding bin. Dividing this value by the total number of games in the corresponding bin will provide the estimated probability of draws. If this probability (calculated from the sample) is larger than the probability proposed by the bookmaker, one can potentially make money in the long run by betting on “Draw” for the games whose odds reside in the corresponding bin.

4. Comment on if there is a bias in odds representing the probabilities? Name the x and y axes accordingly. Write the half at the top of each plot.

Please read [1] if you have difficulty in understanding this question. Section 3.3 discusses relevant topics and Figure 6 (a) is a nice representation.

Task 2

There can be some events during the matches that create noise in the outcomes. To be more specific, let’s consider two specific cases:

- Think of a match in which a team wins with a goal towards the end of the game (i.e. the team scores the winning goal after 90th minute) or similarly match ends draw (tie) or away because of the same reason.

- Bookings can affect the game result. A red card in the first few minutes of a game can change the outcome of the match drastically. Playing with few players is always a disadvantage for the teams.

Perform third and fourth subtask of ***Task 1*** again after removing the matches fitting well to the cases above. Please clearly mention about your decisions in removing the games (i.e. match is removed if there is a red card in the first 15 minutes) and provide removed match counts for each cases. Is there any significant change in the observations you have for ***Task 1***? Comment on the results.

Task 3

Suppose you decided to fit a decision tree model to the outcome of the game (i.e. your target is the result column) and use the statistics for each minute as feature (or some derived feature you end up with). Each row is an instance in that case. It is about trying to foresee the result of the game given the current situation.

Set your decision tree learning parameters such that it allows for commenting on the rules you have identified. Is there any interesting observation regarding the match result?

Your tree may provide insights into if odds are not efficient. In other words, you will obtain a predicted probability of an outcome from your model and if the deviation of the predicted probabilities are not close to the implied probabilities, that might hint for an inefficiency in the market. Do you observe such cases? If so, are they reasonable in terms of what model is telling you?

Note that this part is open-ended and the analyses you will make can provide insights into your project work.

5. Report & Code Documentation

Combine your results and visual aids into a comprehensive report. Your report should:

5. Discuss your methodologies and findings.
6. Draw conclusions based on the patterns observed and their possible real-world implications.
7. Offer insights into any challenges faced and how you overcame them.

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.

6. Conclusion

Data mining is both an art and science. As you journey through this homework, remember that the process is as valuable as the outcome. Your ability to connect the dots between seemingly unrelated pieces of information will shape your growth in the realm of data science.

We're eager to see the unique insights each of you will bring to this project, especially given the diverse academic backgrounds of our cohort. Best of luck, and happy mining!

References

[1] Jonas Mirza and Niklas Fejes,2016, "Statistical Football Modeling A Study of Football Betting and Implementation of Statistical Algorithms in Premier League", available online: http://www.it.uu.se/edu/course/homepage/projektTDB/ht15/project16/Project16_Report.pdf

[2] Štrumbelj, E., 2014. On determining probability forecasts from betting odds. *International journal of forecasting*, *30*(4), pp.934-943.

[3] Shin, H.S., 1993. Measuring the incidence of insider trading in a market for state-contingent claims. *The Economic Journal*, *103*(420), pp.1141-1153.