IE 582 Statistical Learning for Data Mining
Homework 1, due October 11th, 2018

Instructions: Please solve the following exercises using R (http://www.r-project.org/). Also, you are requested to use certain R packages for this particular homework. You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals (as you did for Homework 0).

Do not share your code (except the one in your progress journals)! As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

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.

This project is about understanding the behavior of different betting companies and leagues with the use of available information from different sources (odds from different betting companies, team status and etc.).

Background
The technical report of Mirza and Fejes (2016) 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 (2014) 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:

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

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 (2014) considers a different transformation approach based on the idea of Shin (1993) (i.e. Shin probabilities).

Data

You will find two *.rds files which will be updated every Friday night around 00:10 on the following Google Drive link: https://drive.google.com/open?id=10ubZ9Qb9j1EXQ3xKF3Sg_Bc-AP68XFgS

Match Information

In this folder, "*some_league_id*_matches.rds" stores the information about the soccer games played in a league (It is English Premier League for this assignment.) from August 2010 till today. A snapshot of the data in the file is below:

```
> data$matches_raw
                                    leagueId  matchId        home              away
   1: df9b1196-e3cf-4cc7-9159-f236fe738215 KjF6FiA6   tottenham manchester city
   2: df9b1196-e3cf-4cc7-9159-f236fe738215 ILVbJgQm aston villa        west ham
   3: df9b1196-e3cf-4cc7-9159-f236fe738215 SGIEDUvJ      wolves      stoke city
   4: df9b1196-e3cf-4cc7-9159-f236fe738215 YwL5xFHJ      bolton          fulham
   5: df9b1196-e3cf-4cc7-9159-f236fe738215 1QJAEBPC       wigan       blackpool
  ---
3078: df9b1196-e3cf-4cc7-9159-f236fe738215 dtqIoIYb    brighton       tottenham
3079: df9b1196-e3cf-4cc7-9159-f236fe738215 pK6po6ao    west-ham         chelsea
3080: df9b1196-e3cf-4cc7-9159-f236fe738215 IkpEnbJh     arsenal         everton
3081: df9b1196-e3cf-4cc7-9159-f236fe738215 pK6po6ao    west-ham         chelsea
3082: df9b1196-e3cf-4cc7-9159-f236fe738215 IkpEnbJh     arsenal         everton
      score       date   type
   1:   0:0 1281789900 soccer
   2:   3:0 1281798000 soccer
   3:   2:1 1281798000 soccer
   4:   0:0 1281798000 soccer
   5:   0:4 1281798000 soccer
  ---
3078:   1:2 1537637400 soccer
3079:   0:0 1537709400 soccer
3080:   2:0 1537718400 soccer
3081:    NA 1537709400 soccer
3082:    NA 1537718400 soccer
```

You will find the home and away teams. Each game has a unique match id. The score of the game is NA if it has not been played, otherwise a string in the form of "HomeScore:AwayScore" is provided. The date of the game is provided in "Unix Epoch Time". You can make use of "anytime" package in

R for easy conversion of epoch to dates. Since we will be interested in soccer games for now, type information is redundant and league id is provided in the first column. Upcoming homework may involve data from other soccer leagues so it might be better to keep this information during your calculations.

Odd Information

In this folder, "*some_league_id*_odd_details.rds" represent the games' odd information of different bets of multiple bookmakers stored in matches data. This data contains around 5 million rows so you are expected to use efficient data structures in R to handle such data. I would advise the use of "data.table" package in R for efficient manipulation of very large datasets like this. A snapshot of the data in the file is below:

	matchId	betType	oddtype	bookmaker	date	odd	totalhandicap
1:	004f4ING	1x2	odd1	10Bet	1420971300	1.67	NA
2:	004f4ING	1x2	odd1	10Bet	1422806160	1.65	NA
3:	004f4ING	1x2	odd1	12BET	1421025840	1.67	NA
4:	004f4ING	1x2	odd1	12BET	1422805680	1.65	NA
5:	004f4ING	1x2	odd1	188BET	1421036580	1.70	NA

4908850:	zZ6f59Ue	ou	under	SBOBET	1534519020	1.64	3.25
4908851:	zZ6f59Ue	ou	under	SBOBET	1534122660	2.25	2.5
4908852:	zZ6f59Ue	ou	under	SBOBET	1534099500	1.95	2.75
4908853:	zZ6f59Ue	ou	under	SBOBET	1534395840	1.73	3
4908854:	zZ6f59Ue	ou	under	SBOBET	1534522620	1.64	3.25

For each game, multiple bookmakers (bookmaker column) provide different odd types (betType column). These can be:

- "1x2": Game result
- "ah": Asian handicap
- "bts": Both teams to score
- "dc": Double chance
- "ha": Draw no bet
- "ou": Over/under

You can do a Google search to understand what these betting types are referring to. Please note that not every bookmaker provide all types of odds. Moreover, some bookmakers may be closed for certain time periods.

Oddtype column stores the information about the decision corresponding to the odd (odd column) in the row. For example, first row provides information about the odd for "Home" win (i.e. oddtype="odd1") from bookmaker "10Bet" for the game with matchId 004f4ING. This is a bet of type "Game Result (i.e. 1x2)". For over/under bet types, there is additional parameter for the number of goals. This is given in totalhandicap column. Asian handicap type of bets also have handicap parameter.

Note that this table keeps information about the change in odds (each observation is timestamped and date column stores this info). Based on the timestamps, you can find the starting and ending odds. On the other hand, the odds change dynamically since bookmakers adjust their implied probabilities because of certain reasons. The main reason for such adjustment is to reduce the risk of losing money when the money on a certain type of bet (i.e. over 2.5) increases significantly. Moreover, this adjustment also takes place in the cases where a key player is injured before the game time. Therefore, it is an interesting problem to analyse these type of odd movements. Provided dataset allows for such analyses. For a specific matchID, bettype, oddtype and bookmaker combination, you will observe multiple odds with the corresponding time of record. The entry with minimum date refers to the initial odd whereas the one with the maximum date is the latest odd. The information in the change of the odds is also interesting since it may provide useful information about the game result.

Tasks

Task 1

This task is to understand if bookmakers are good enough in setting their odds for over/under bets. You can focus on only 2.5 threshold (i.e. totalhandicap=2.5). An empirical evidence for the probability of over or under 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 over probability is 0.5 for a specific game and your probability range is 0.48 and 0.52), you can count the games that finished over within this range. In other words, we can discretize probability of over 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 "over" in the corresponding bin. Dividing this value by the total number of games in the corresponding bin will provide the estimated probability of overs. 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 bookmakers are good enough in determining odds (in other words, if they make money), what you expect to see is that fraction of games finished over is between this implied probability range.

a) A simple way to understand if bookmakers are good at deciding over/under 2.5 odds is to draw a scatter plot of implied "over" probability average in a bin versus the fraction of games finished over for the corresponding bin. For a successful bookmaker, we would expect to see points around x=y line (also draw this line). Select at least 5 bookmakers which introduce over/under bets and plot this information (for both initial and final odds). Comment on your findings. Note that if few games reside in a bin, the statistic you compute (i.e. fraction of games finished over) may not be meaningful. You may want to merge bins in such cases.
b) Select a bookmaker to check if they are consistently good at determining "over" probabilities during the given time period. This can be achieved by calculation of the required information above over the years. This way, you will obtain the same information for each year. A line plot that shows average implied probabilities for each bin and corresponding fraction of games over time will be useful for this task. Since there would be multiple bins, you can have multiple lines on a single plot. If you observe that your graph becomes messy because of the number of bins, you can limit yourself to certain number of bins for each plot and have multiple plots to include all bins. To clarify, consider we are interested in the implied probability range of (0.4,0.5], you will find the fraction of games finished over for each year. That will make two observations for the corresponding year: average of the probabilities in the bin and the fraction of games finished over. You can plot these two on a single plot with separate lines. If you think that alternative visualizations will work better, please suggest your way of performing the same task. You are not restricted to what is proposed in this part (i.e. line plot).

Task 2

As mentioned, change in the odds are interesting. Let's focus on the change of the odds for match results (i.e. 1x2 bets) in this task. What type of visualization(s) would you suggest to see if there is any information in the odd changes and why? You can select a bookmaker for performing this task. Provide your visualization(s) and comment.

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.