# IE 582 Statistical Learning for Data Mining
## Homework 1, due November 20th, 2020

Instructions: Please solve the following exercises using R (http://www.r-project.org/) or Python (https://www.python.org/). 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:

- A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
  - https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/
  - https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps-917519326e41

  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.

- 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).

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. 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 homework is about understanding the behaviour 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 [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).

### DATA

The website, https://www.football-data.co.uk/data.php, provides historical match betting odds data from up to 10 major online bookmakers are available back to 2000/01. Additionally, the data for 16 other worldwide premier divisions, with fulltime results and closing match odds (best and average market price, and Pinnacle odds) are also provided dating back to 2012/13.

The information about the soccer games are provided as csv files (i.e. https://www.football-data.co.uk/mmz4281/2021/E0.csv for the latest season of English Premier League). Each row corresponds to a match and related information is provided as columns. There are several columns in the data each of which is described in the Google Sheet on https://docs.google.com/spreadsheets/d/16aZYItIHmGmqaLnRaW0XUMcwFFNxhcLB6nKqpzog06c/ . Please note that the definitions for the first 56 columns are provided, you can discard the rest of the variables.

**TASKS**

The aim of the first homework is to get you familiar with data manipulation. It involves certain descriptive analyses to understand the data. Please use English Premier League data of Season 2020/2021, Season 2019/2020 and Season 2018/2019 for the following tasks:

***Task 1***

There two related subtasks:

1. Plot the following histogram diagrams
   a. Home Score(goals)
   b. Away Score(goals)
   c. Home Score(goals)– Away Score(goals)

   Name all y-axes "Number of Games", and each x-axis "Home Goals", "Away Goals" and "Home goals – Away Goals" for each plot respectively.

2. To which probability distribution do home and away goals fitting well? Does the distribution look like Poisson distribution? Calculate the expected number of games corresponding to each quantile (number of goals) with Poisson distribution by using sample means as distribution mean and plot these values on the histogram. Is this consistent with Poisson distribution claim? In other words, compare the actual outcomes with the theoretical distribution on a plot. It is expected to obtain something similar to the third plot on the following link: https://www.statmethods.net/graphs/density.html

***Task 2***

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 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 as "draw" is between this implied probability range. Select at least 4 bookmakers for this task.

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 for each bookmarker.

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.** You will do this for each bookmaker separately (You will construct at least 4 plots in total). Comment on if there is a bias in odds representing the probabilities? Name the x and y axes accordingly. Write the name of bookmaker 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 3***

There can be some events during the matches that create noise in the outcomes. To be more specific, let's consider a specific case. 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 2*** again after removing the matches fitting well to the case above. Is there any significant change in the observations you have for ***Task 2***? Comment on the results.

### 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.