IE 256 Statistics for Industrial Engineers
Project Part 1, due March 27th, 2018

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."

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 probabilities become:

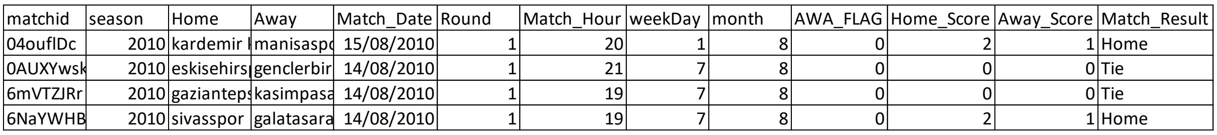
$$\begin{bmatrix} \mathrm{P(home)} \\ \mathrm{P(draw)} \\ \mathrm{P(away)} \end{bmatrix} = \begin{bmatrix} 1/\mathrm{odds}_1 \\ 1/\mathrm{odds}_X \\ 1/\mathrm{odds}_2 \end{bmatrix} \cdot \frac{1}{\sum_{i \in \{1,X,2\}} 1/\mathrm{odds}_i},$$

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.

Data

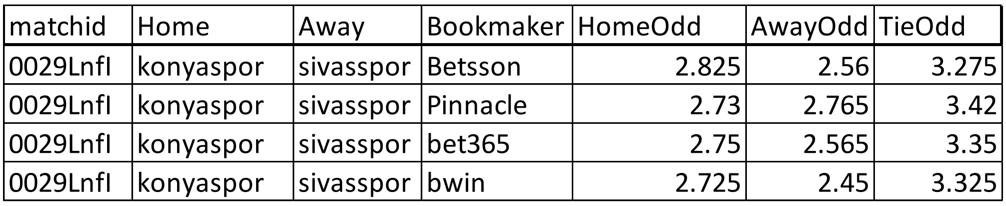
You will find two .csv files on the moodle: "tr_super_league_matches.csv" and "tr_super_league_odd_details.csv".

"tr_super_league_matches.csv" represents the Turkish Super League match results in terms of scores and the match ended in which team's favour. A match can result with the win of home or away team or a tie (in case of scores of teams are equal). "Home_Score" column represents the score (number of goals) of home team; "Away_Score" column represents the score (number of goals) of away team. "Match_Result" column represents the match result in terms of "Home", "Away" or "Tie". The other columns are as follows: First column is specific to each match which represents a unique match id. Second column contains the season information in terms of year. You can find the home and away team information under "Home" and "Away" columns. From columns 5th to 9th, date, round, hour, week, day and month information of the match can be found. "AWA_FLAG" column takes a value of one if the game is suspended. In other words, there was an event during the game and the game was not finished regularly.



matchid	season	Home	Away	Match_Date	Round	Match_Hour	weekDay	month	AWA_FLAG	Home_Score	Away_Score	Match_Result
04ouflDc	2010	kardemir k	manisaspo	15/08/2010	1	20	1	8	0	2	1	Home
0AUXYwsk	2010	eskisehirs	genclerbir	14/08/2010	1	21	7	8	0	0	0	Tie
6mVTZJRr	2010	gaziantep	kasimpasa	14/08/2010	1	19	7	8	0	0	0	Tie
6NaYWHB	2010	sivasspor	galatasara	14/08/2010	1	19	7	8	0	2	1	Home

"tr_super_league_odd_details.csv" represent the odd information of four bookmakers. "Home" and "Away" columns contain the name of teams; "Bookmaker" column gives the name of bookmaker office. For each match there are four different odds from four different bookmakers. You can see that from "matchid" column. "Home_Odd", "Tie_Odd" and "Away_Odd" columns refer to odds provided for each match by each bookmaker.



matchid	Home	Away	Bookmaker	HomeOdd	AwayOdd	TieOdd
0029Lnfl	konyaspor	sivasspor	Betsson	2.825	2.56	3.275
0029Lnfl	konyaspor	sivasspor	Pinnacle	2.73	2.765	3.42
0029Lnfl	konyaspor	sivasspor	bet365	2.75	2.565	3.35
0029Lnfl	konyaspor	sivasspor	bwin	2.725	2.45	3.325

Tasks for Project Part 1

Task 1

1. By using “tr_super_league_mathches.csv” data set; 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. What distribution do you think home and away goals are coming from? 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?

Task 2

In this task, you will use “tr_super_league_odd_details.csv” in addition to match information (“tr_super_league_mathches.csv”). Please read the **“Odds and Probabilities”** part again before this task.

1. Calculate the P(home win), P(draw) and P(away win) probabilities by using normalization formula at “Odds and Probabilities” part for each bookmarker. “Draw” in the formulation corresponds to “Tie” in the data.

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

   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.

3. You will do this for each book marker separately (You will construct 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 bookmarker at the top of each plot.

References

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

Instructions

Please solve the following exercises using R. One easy way to present your solutions is to copy all of your code and results (plots, numerical answers, etc.) into a Word document, which you should submit in class. **There is no online submission.** You can work as groups of three.

The last and the most important thing to mention is that academic integrity is expected! **Do not share your code! As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.**