Predicting Football Results Based on Odds of Betting Exchanges and Bookmakers by Logistic Regression

Module Code: BENVGSA2

Student Code: 17049324

Date: 2018.1.7

Count: 1978

Introduction

As reported by the Fédération International de Football Association (FIFA 2014), there are over 38 million professional soccer athletes and 30100 clubs. The industry of football has ushered in unprecedented development since 1990s due to the growth of television and media (Sanctis 2014). Sanctis claimed that the football lottery has played the similarly significant role on the public especially in Asia, Europe, and South America as the sport itself, which is also supported by Dobson. The industry of sports gambling is involved in psychology, economics, sociology, statistics, law and so on. The betting industry has already become a mature market since 1990s—the unprecedented growth of football gambling in UK whose customers are distributed worldwide. Companied by the flourishing of betting market, besides millions of bettors and bookmakers, some experts, statistician, economist also pay attention to the process of this game. In 1997, Dixon and Coles proposed a score-based forecasting model, and in 2005, Forrest, Goddard and Simmons stated an advanced results-based model, which are significant in the development of gambling industry, and all present excellent ability to predict. But bettors who use the tools to make decisions rarely profit because the betting is a game based on the understanding on the football match. There is a difference of understanding between the gamblers and bookmaker due to internal information or more advanced analysis (Forrest, Goddard and Simmons 2005). This article is based on the **assumption 1** that **the bookmakers can do most accurate prediction if it is possible to predict the results of the football match**. The aim of this article is to make an accurate prediction based on the information advantages of bookmakers.

Analysis of Football Betting

This article takes fractional odds, the odds are usually set by bookmakers. For example, in table1.The win, draw lose are the results of the home team, if betting on win with $1, and finally the result is win, the return is that $1 times odds of win 2.00 equals $2. The reciprocal of odds reflects a probability, everyone who is involved in betting has a series of accurate probability in their mind

Table1: Odds and Probability

Odds	Win: 2.00	Draw: 3.00	Lose: 4.00
Probability of Odds	50%	33.33%	25%
Accurate Probability	40%	50%	10%

It is significant for bettors to find a situation where the accurate probability is bigger than the probability of odds because the expected of betting profit is calculated like that if betting with $1 Where if the Accurate Probability is bigger than the reciprocal of odds – the Probability of Odds, the expected profit is positive according to (1.1) and (1.2):

$$E = (\text{Odds} * \$1 - \$1) * \text{Accurate Probability} - \$1*(1- \text{Accurate Probability}) \quad \ldots (1.1)$$

$$\downarrow$$

$$E=\text{Odds}* \text{Accurate Probability}-1>0$$

$$\rightarrow \qquad \ldots (1.2)$$

$$\text{Accurate Probability}>1/\text{odds}$$

Apparently, it is most important for bettors to get Accurate Probability which is as accurate as possible. According to **assumption 1**, it is better to get the Accurate Probability hold by bookmakers instead of making prediction by ourselves. As claimed by Forrest, Goddard, and Simmons, the bookmakers do not master all results of the football match before it, there is two situations for bookmakers to set odds depend on whether they know the real probability (to some extent very close to the truth) of results or not.

If they do not know the real probability, which is true for most situations, the bookmakers can hedge risk by setting odds because most of bookmakers know the distributions of people's bets, which is a fundamental foundation how traditional bookmakers can profit stably and long-term and it is supported by research of Dixon and Coles in 1997.It is to say that no matter which one the result is finally, the expected profit shown in (2.1) are similar and the E varies depend on how much do the bookmakers want to make.

$$E=1-\text{Swin}*\text{Owin}=1-\text{Sdraw}*\text{Odraw}=1-\text{Slose}*\text{Olose} \quad \ldots(2.1)$$

Swin, Sdraw, Slose present the percentage of betting amount betting on win, draw and lose repectively.
Owin, Odraw, Olose present the odds of win, draw and lose repectively.
E present the expected profit and the percentage of profit in betting amount.

For example, in table 2, no matter when the home team wins, loses or draws, the expected profit is 10%, which is also the return of the bookmaker.

Table2: Odds when do not know real probability

	Win	Draw	Lose
Odds	1.8	3	4.25
Percentage of betting	50%	30%	20%
Expected Profit	10%	10%	10%

If the bookmakers know the real probability, they can make more or set odds which are more attractive for bettors to promote their business. The expected profit is calculated in (3.1):

E = A – (Owin*Pwin*Swin+ Odraw*Pdraw*Sdraw+ Olose*Plose*Slose) … (3.1)

A is the betting total.

Pwin, Pdraw and Plose present the real probability of win, draw and lose repectively.

As the E is always set positive and as big as possible. If the odds do not meet (2.1), they have to meet the (3.1). If the situation meets (2.1), there is no real probability to get, which means that we cannot get valuable clues from odds. But if the situation meets （3.1）, it is profitable to bet on situation meeting (1.2).In (3.1), the odds is open to the public, if the Swin, Sdraw and Slose are known, the relationship between Pwin, Pdraw and Plose is clear because the E keep positive and the (4.1)

Pwin + Pdraw + Plose = 1 …(4.1)

Now it is significant to get the percentage of betting amount betting on win, draw and lose which the bookmakers never public. In 2000s, as the development of the internet, a new person-to-person betting appeared called as online betting exchanges. The online exchanges work as stock exchanges who just charge processing fee instead of setting odds. The odds of betting exchanged are product like (5.1) which means the (5.2)

1-Swin*Owin = 1-Sdraw*Odraw = 1-Slose*Olose=0 …(5.1)

Swin*Owin = Sdraw*Odraw = Slose*Olose=1 …(5.2)

And because there are (5.3) and (5.2), the (5.4) is true.

Swin + Sdraw + Slose*=1 …（5.3）

Swin (= 1/Owin) + Sdraw(=1/ Odraw)+ Slose (=1/ Olose)=1 …（5.4）

As the Odds of betting exchanges are always open to public, the percentage of betting amount betting on win, draw and lose on betting exchanges is known, As the wide distribution of customers on betting exchanges, I assume that **the percentage of betting amount betting on win, draw and lose on betting exchanges reflect the percentage of betting amount betting on win, draw and lose on traditional gambling companies where the odds are set by bookmakers** and prove the assumption in the following sectors.

In addition, after getting the percentage of betting amount without the understanding of real probability. The expected profit of betting on a result can be calculated like (6.1).

$$E = 1\text{-}Swin*Owin \qquad \ldots(6.1)$$

For example, if the E is negative like -0.3, which means the bookmakers will lose 30% of total betting when the result is win. Meanwhile, the odd of win (Owin) keep increasing, which means the bookmakers will lose more as the result is win, it is true qualitatively that the bookmakers insist on that the probability of win is very small and the E in (6.1) is a significant indicator of identifying which one the result will be.

Data and Methods

Data Source

As above, the situations where the bookmakers do not know the real probability are not good for gamblers to predict results, and as stated by Dobson et.al, in most cases, the bookmakers which one keep high return rate do not have inside information and real probability. But the companies who profit by advanced analysis on prediction can provide more valuable information. This article takes odds on recent 1000 match of three bookmakers, "BET 365", "INTERWETTEN" and "Macau Lottery Co., Ltd.". And the odds on "betfair" which is an authoritative betting exchange are used to calculated the percentage of betting amount betting on win, draw and lose.

The results including win, lose and draw, the prediction of results is a classification problem, so the article takes logistic regression to do research on above analysis.

Logistic Regression

Logistic Regression was developed by David Cox in 1958. Basically, the logistic regression is used in a binary dependent variable problem, "1" represents one situation and "0" represents the other situation. The output, hypothesis function of the regression is the probability of result "1" product by the logistic function (7.1), which arrange from 0 to 1 in Fig 1.

$$\mathrm{G(t)} = \frac{1}{1+e^{-t}} \qquad \ldots(7.1)$$

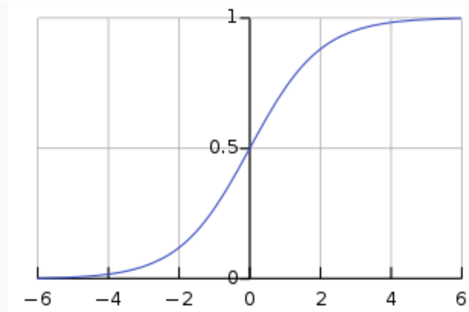


Figure1: Logistic Function $H_\theta(\mathrm{X}) = \mathrm{G(t)} = \frac{1}{1+e^{-t}}$

The independent variable is the coefficients time variables in (8.1)

$$\mathrm{t} = \theta^{T} \cdot X$$

$$H_{\theta}(\mathrm{X}) = \mathrm{G}(\theta^{T} \cdot X) \ldots(8.1)$$

Cost function is used to get coefficients θ, the cost of logistic regression is J(θ) in (9.1)

$$\mathrm{J}(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(\ h_{\theta}\left(x^{(i)}\right)\ , y^{(i)})$$

$$\text{If } y = 1 \qquad \mathrm{Cost}(h_{\theta}(x), y) = -\log\left(h_{\theta}(x)\right. \qquad \ldots(9.1)$$

$$\text{If } y = 0 \qquad \mathrm{Cost}(h_{\theta}(x), y) = -\log\left(1 - h_{\theta}(x)\right)$$

m presents the amount of samples

The θ – coefficients are set after minimizing the J(θ), this article takes "gradient descent" to minimize the J(θ). The process of gradient descent likes that some find a fast way to down mountain in (10.1). It is necessary to repeat the process (10.1) for many times to prevent from getting local minimizing and set a small α as 0.01to make the minimizing accurate.

Repeat: {

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \qquad \ldots(10.1)$$

}

α presents the length of a "step".

The above is for binary classification, but the football results including 3 categories win, lose and draw. After repeating the logostic regression for three times to set win, lose and draw as result "1" respectively, and the highest $H_{\theta}(\mathrm{X})$ is the predicted result which has the greatest probability.

Results

Odds of "BET365" and Percentage of "betfair"

80% of sample data is drawn to set as train data for regression and 20% is used to test. Coefficients of logistics regression when the win is set as result "1" are shown in table3. 1/ the original odds are set as variables instead of odds itself.

Table 3: Coefficients when the win is set as result "1"

ariables	1/Owin	1/Odraw	1/Olose	Ewin	Edraw	Elose	OwinC	OdrawC	OloseC	Intercept
oefficients	-3.5336	-14.506	-9.0104	0.3730	0.1949	-0.0611	-0.2348	0.421	-0.3978	8.9293

Owin, Odraw and Olose present the original odds set by the bookmaker
Ewin, Edraw and Elose are calculated as in (6.1)
OwinC, OdrawC and OloseC present the change of odds from the original odds to the odds before the match begin. (Odds keep varying before beginning)

Coefficients of logistics regression when the draw is set as result “1” are shown in table3.

Table 4: Coefficients when the draw is set as result “1”

riables	1/Owin	1/Odraw	1/Olose	Ewin	Edraw	Elose	OwinC	OdrawC	OloseC	Intercept
efficients	6.6183	12.226	8.1001	-0.0615	2.9817	-0.2909	0.1154	0.1676	0.4568	-10.5591

Coefficients of logistics regression when the lose is set as result “1” are shown in table3.

Table 5: Coefficients when the lose is set as result “1”

riables	1/Owin	1/Odraw	1/Olose	Ewin	Edraw	Elose	OwinC	OdrawC	OloseC	Intercept
efficients	-3.7606	0.9411	2.3341	0.1376	-2.3072	2.5573	0.0882	-0.4671	-0.0047	-0.2752

The coefficients of the three regressions are used to predict the result in the 20% data – testing data to test the ability of prediction. To improve the accurateness, instead of $H_{\theta}(X)$ >50% that usually shows the result “1” will appear, higher and varied thresholds of the highest $H_{\theta}(X)$ in the three regressions predicting are set in predicting results. For example, if the threshold is set at 0.7, the three $H_{\theta}(X)$ are all below 0.7, the predicting is classified as invalid. The accuracy of predicting in different thresholds is shown in Figure 2 and Table 4.

Table 6: Accuracy with varied thresholds

Threshold of $H_{\theta}(X)$	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
Accuracy	57%	60%	61%	62%	63%	62%	62%	62%	64%	64%

Figure2: Threshold and Accuracy

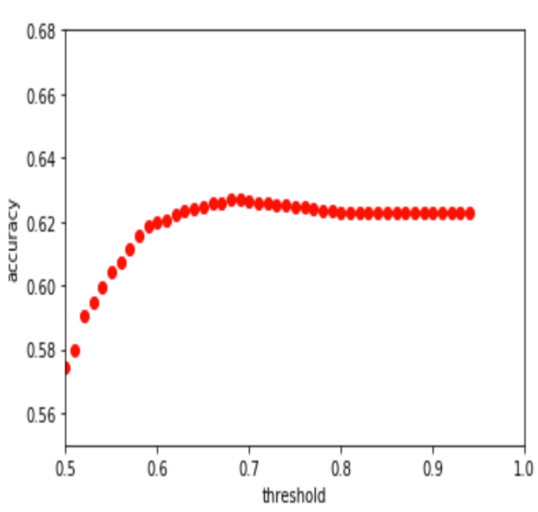
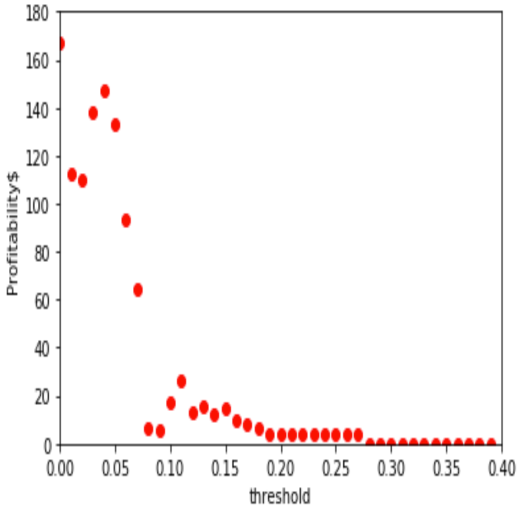


Figure 3: Profitability



Profitability (Fig3) of betting based on the predicting results varied with increased thresholds (there is $1 at beginning, if meet the threshold, bet $1, the profitability is the remaining money after 1000 match) of difference between accurate probability and the reciprocal of odds – the Probability of Odds according to (1.1) and (1.2). And Accuracy of “INTERWETTEN” and “Macau Lottery Co., Ltd” are shown in Fig4 and Fig6, the profitability of the two firms are shown in Fig5 and Fig7.

Odds of “INTERWETTEN” and Percentage of “betfair”

Figure4: Threshold and Accuracy of INTERWETTEN

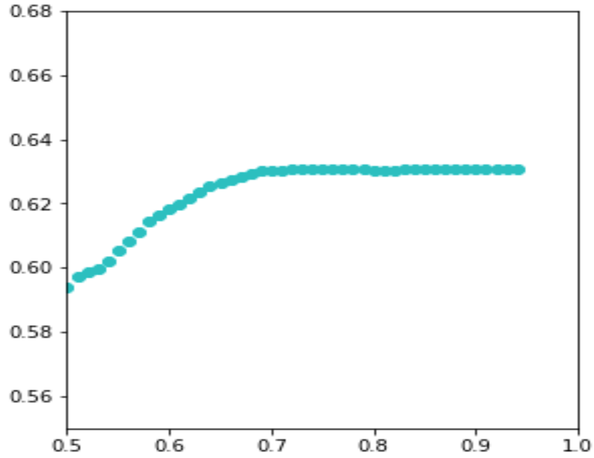
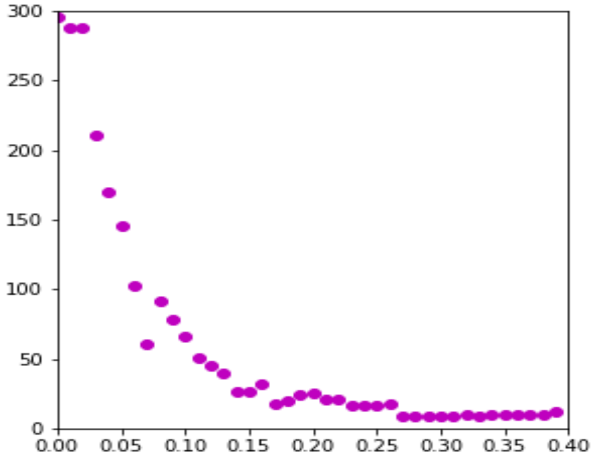


Figure 5: Profitability of INTERWETTEN



Odds of “Macau Lottery Co., Ltd” and Percentage of “betfair”

Figure6: Threshold and Accuracy of Macau

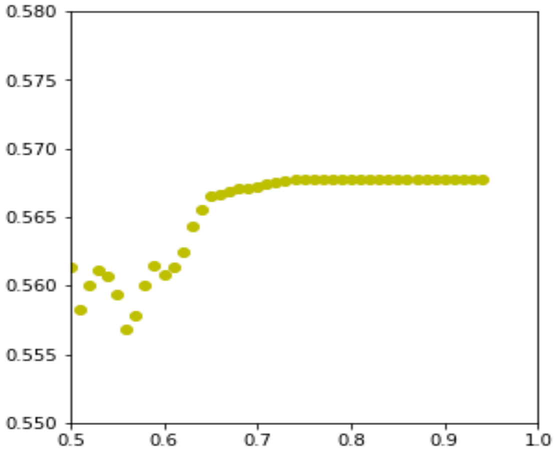
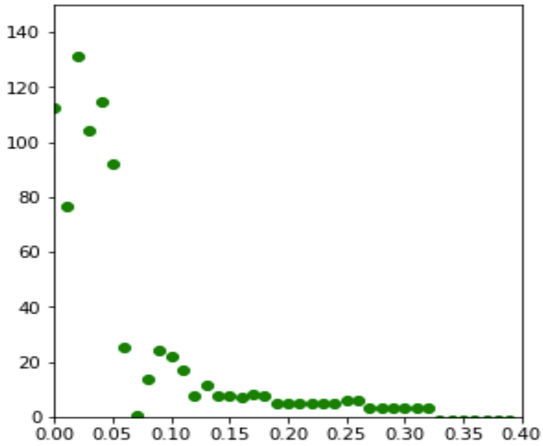


Figure 7: Profitability of Macau



Discussion

As we see from Table4,5,6, the probability reflected by odds (1/Owin, 1/Odraw, 1/Olose) have an influence on results, the higher probability reflected by odds may represent there is a higher real probability of the result. The E-results (calculated from (2.1)) also have a positive influence on the results. The transformation of odds set by bookmakers has little influence on results and the increase of odds show a decrease of probability on the result to some extent.

As we see from Figure2,3,4,5,6,7, the accuracy of predicting keeps around 60% except for Macau that just reaches less than 57%, which are all not accurate prediction. But the results of profitability are positive, the profitability betting on Macau and "bet365" can reach $140 and $170, betting on INTERWETTEN can get return around $300. The thresholds should keep less than 0.1, otherwise, there will be no qualified match to bet on and the profit will become $0 soon.

Conclusion and Limitation

The predicting for the results are not so good because the output is just a probability, but the results of profitability can keep positive, which approves there are some matches where the bookmaker know the accurate probability and the formula (3.1) can work, if the distribution of how public bet on different results is known, the results of the kind of match can be predicted or it is obvious that which result hardly appear or which result there is a great chance appearing of depend on the (6.1).

In fact, the distribution of people betting on betting exchanges and bookmakers is different. The distribution of how people bet on bookmakers should be calculated or observed in better approach, which will make predicting and probability more reasonable.

Reference:

De Sanctis, F.M., 2014. *Football, gambling, and money laundering: A global criminal justice perspective*. Springer.

Dobson, S., Goddard, J.A. and Dobson, S., 2001. *The economics of football* (pp. 106-130). Cambridge: Cambridge University Press.

Dixon, M.J. and Coles, S.G., 1997. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), pp.265-280.

Forrest, D., Goddard, J. and Simmons, R., 2005. Odds-setters as forecasters: The case of English football. *International journal of forecasting*, *21*(3), pp.551-564.
Cox, D.R., 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.215-242.

McCall, A., Davison, M., Andersen, T.E., Beasley, I., Bizzini, M., Dupont, G., Duffield, R., Carling, C. and Dvorak, J., 2015. Injury prevention strategies at the FIFA 2014 World Cup: perceptions and practices of the physicians from the 32 participating national teams. Br J Sports Med, 49(9), pp.603-608.