International Journal of Forecasting 30 (2014) 934–943

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

On determining probability forecasts from betting odds

Erik Štrumbelj *

University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, 1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Sports forecasting
Probability forecasting
Fixed-odds
Betting exchange
Shin's model
Betfair
Calibration

ABSTRACT

We show that the probabilities determined from betting odds using Shin's model are more accurate forecasts than those determined using basic normalization or regression models. We also provide empirical evidence that some bookmakers are significantly different sources of probabilities in terms of forecasting accuracy, and that betting exchange odds are not always the best source, especially in smaller markets. The advantage of using Shin probabilities and the differences between bookmakers decrease with an increasing market size.

1. Introduction

There has been interest in the scientific literature in the accuracy of betting odds-based probability forecasts both directly, by comparing them to other sources of probability forecasts, and indirectly, through their use in betting strategies and as explanatory variables in statistical models. The probabilities from betting odds are also used in research into issues such as market efficiency and the competitive balance of sports competitions. For reviews, we refer the reader to Humphreys and Watanabe (2012), Stekler, Sendor, and Verlander (2010) and Vaughan Williams (2005).

The widespread use of betting odds is not surprising, as there is substantial empirical evidence that betting odds are the most accurate publicly-available source of probability forecasts for sports. With the growth of online betting, betting odds are also readily available for an increasing number and range of sports competitions. However, we believe that the following two issues with using betting odds as probability forecasts have not yet been addressed sufficiently:

(a) Which method should be used to determine probability forecasts from raw betting odds?
(b) Does it make a difference as to which bookmaker or betting exchange we choose, when two or more are available?

We address these two issues in the context of fixed-odds betting, with an emphasis on evaluating the most commonly used methods for determining probability forecasts from odds. Empirical evaluation is performed using data from several different online bookmakers across 37 competitions and five different team sports (basketball, handball, ice hockey, soccer, and volleyball).

1.1. Related work

As a matter of brevity and convenience, we focus on the most relevant results for fixed-odds betting, which is prevalent in team sports.[1]

The empirical evidence suggests that betting odds are the most accurate source of sports forecasts. Odds-based probability forecasts have been shown to be better than, or

* Tel.: +386 1 4768459.
E-mail address: erik.strumbelj@fri.uni-lj.si.

1 We omit a substantial subset of the literature on racetrack betting that focuses primarily on parimutuel markets and their efficiency (see Hausch, Lo, & Ziemba, 2008, for a review).

http://dx.doi.org/10.1016/j.ijforecast.2014.02.008
0169-2070/

at least as good as, statistical models using sports-related input variables (Forrest, Goddard, & Simmons, 2005; Song, Boulier, & Stekler, 2007; Štrumbelj & Vračar, 2012), expert tipsters (Song et al., 2007; Spann & Skiera, 2009), and (aggregated) lay predictions (Pachur & Biele, 2007; Scheibehenne & Broder, 2007).

A special subset of betting odds are odds from betting exchanges. Unlike fixed-odds, which are formed by bookmakers, betting exchange odds are formed by bettors. That is, betting exchanges facilitate both backing and laying bets, and can be considered a form of prediction market.

In many different domains, forecasts from prediction markets are more accurate than those produced by traditional forecasting approaches and single forecasters (Arrow et al., 2008; Graefe & Armstrong, 2011; Tziralis & Tatsiopoulos, 2007). In sports forecasting, the term 'betting exchange' in most cases means Betfair, the world's largest betting exchange. There is substantial empirical evidence that the probabilities determined from Betfair odds are more accurate forecasts than those from fixed-odds bookmakers (Franck, Verbeek, & Nuesch, 2010; Smith, Paton, & Williams, 2009; Spann & Skiera, 2009; Štrumbelj & Vračar, 2012). Štrumbelj and Robnik-Šikonja (2010) also showed that there are significant differences between online fixed-odds bookmakers in terms of forecasting accuracies.

2. Determining outcome probabilities from betting odds

Fixed-odds bookmakers post betting odds, which indicate how much a bet placed with the bookmaker at that time would pay if it were to win. An online bookmaker posted the following betting odds for the 2012 Champions League Final match regular time outcome: Bayern Munich 1.80, Draw 3.75, Chelsea 4.33. Given that regular time ended in a 1:1 draw, we now know that for every unit we had bet on a draw, we would have won 3.75 units (2.75 + the unit we bet). Money bet on either of the other outcomes would have been lost.

The inverse odds are an indication of the bookmaker's underlying probabilistic beliefs. In our case, they suggest that Bayern had at most a $\frac{1}{1.80} = 0.56$ chance of winning, Chelsea 0.23, and that there was at most a 0.27 probability of a draw. However, bookmakers do not offer fair odds, so the sum of the inverse odds (also known as the *booksum*) will always be greater than 1 ($0.56 + 0.27 + 0.23 = 1.06$). In order to use the inverse odds as probability forecasts, we therefore have to account for the excess 6% (also known as the *bookmaker take* or bookmaker margin).

Most studies use basic normalization (dividing the inverse odds by the booksum).[2] In fact, this approach has become almost synonymous with the use of betting odds, although it is not clear whether bookmakers really do add their take proportionately across all possible outcomes. The widespread use of basic normalization can be attributed to its simplicity.

Alternatively, we can view the outcome as a categorical variable and model the probabilities using a historical data set of betting odds and corresponding match outcomes (see for example Forrest et al., 2005; Forrest & Simmons, 2002; Goddard, Beaumont, Simmons, & Forrest, 2005). Due to the categorical nature of the dependant variable, either logistic (probit) regression or multinomial regression is used, depending on the number of outcomes. An ordered model is preferred if there is a natural order to the outcomes.

There are only a few studies that have used an alternative to basic normalization or regression modeling. Smith et al. (2009) used a theoretical model of how bookmakers set their odds that was originally proposed by Shin (1993). Shin's model can be used to reverse-engineer the bookmaker's underlying probabilistic beliefs from the quoted betting odds. For earlier uses of Shin's model, see the works of Cain et al. (for example Cain, Law, & Peel, 2002, 2003; Smith et al., 2009, and references therein). They show that Shin's model-based approach improves on basic normalization. We adopt their term *Shin probabilities* to refer to probabilities determined from betting odds by using Shin's model.

Surprisingly, Shin probabilities have not been adopted widely, and the little use they have seen has focused almost exclusively on racetrack betting. A logical question that follows, therefore, is, can normalization based on Shin's model improve on basic normalization in individual and team sports?

2.1. Basic normalization

Let $\mathbf{o} = (o_1, o_2, \ldots, o_n)$ be the quoted decimal odds for a match with $n \geq 2$ possible outcomes, and let $o_i > 1$ for all $i = 1 \ldots n$. The inverse odds $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$, where $\pi_i = \frac{1}{o_i}$, can be used as latent team strength variables, but do not represent probabilities, because they sum to more than 1.

Let $\beta = \sum_{i=1}^{n} \pi_i$ be the booksum. Dividing by the booksum, $p_i = \frac{\pi_i}{\beta}$, we obtain a set of values that sum to 1 and can be interpreted as outcome probabilities. We refer to this as *basic normalization*.

2.2. Shin's model

Shin (1993) proposed a model which is based on the assumption that bookmakers quote odds which maximize their expected profit in the presence of uninformed bettors and a known proportion of insider traders.

The bookmaker and the uninformed bettors are assumed to share the probabilistic beliefs $p = (p_1, p_2, \ldots, p_n)$, while the insiders are assumed to know the actual outcome before the actual experiment (race, match, etc...). In sports, such superior information can be due either to a better aggregation of publicly available knowledge, or to private information, such as match-fixing.

Without loss of generality, we can assume that the total volume of bets is 1, of which $1 - z$ comes from uninformed bettors and z from insiders. Conditional on outcome i occurring, the expected volume bet on the ith outcome is $p_i(1 - z) + z$. If the bookmaker quotes $o_i = \frac{1}{\pi_i}$

[2] For our example, basic normalization gives Bayern $\frac{0.56}{1.06} = 0.528$, Chelsea 0.217, and Draw 0.255. Using Shin's model, we get 0.535, 0.215, and 0.250, respectively.