Inefficiencies in the Soccer Betting Market

Owen Prunskis   Akhil Ketkar

April 10, 2015

1   Introduction

Soccer is the most popular sport in the world, particularly in Europe. The popularity of the sport combined with the legal status of gambling has led to the formation of a very deep and active betting market for soccer. A number of different bookmakers allow people to bet large sums of money on various aspects of the game. One can obviously bet on the outcomes of games but also other scenarios like the score at half-time, number of penalties, etc. The objective our project is to find actionable inefficiencies in the soccer betting market.

2   Previous work

A fair amount of work has been done in this area. But it pales in comparison to the amount of research done on American sports, especially baseball. The most important work for the purposes of our project is the Dixon-Coles 1997 paper [1]. Additional models and expansions are included in the bibliography as well. [4] [5] [6]

3   Methodology

Dixon and Coles propose a bayesian model that uses a bi-variate Poisson distribution for number of goals scored by each team. We will use this model as the basis of our inquiry and improve upon it as needed. In particular we might look at differences in scoring in the first half of a game compared

to the second half, the effect of yellow and red cards, the effect of multiple games within the same week, of injuries to certain players, of coaches etc.

We will train the model on available data and then use MCMC methods to sample from the distributions of the model parameters. Finally we will test the model against bookmakers' odds to see if they can be systematically beaten.

4   Data

The data we will use for this project will primary come from the various divisions of English Football focusing on the Premier League which is the top tier of competition. Although one might suspect the betting market in smaller teams to be less efficient. The data comes from the website Football Data [2] which has data for all divisional games for multiple years along with the bookmakers odds for each game. Additionally, we will leverage the existing EPL ipython notebook, which implements the basic model and will allow for more flexibility in improving upon the existing work in the field. [3]

5   Expansions and Improvements

Existing models typically generate a clustering effect, which tends to normalize outcomes. We look to incorporate additional parameters to address this shortcoming. We consider the following additions:

1. *Game Importance*, as defined by a threshold point distance away from winning the championship or the lower bound for relegation.

2. *Midweek Games*, to capture the effects of player fatigue.

3. *Weather*, binary as 'clear' or 'rain', which has a marked effect in other sports litrerature (i.e. baseball)

4. *Roster Cost*, amount spent on players and its correlation to improved results (QPR being an obvious outlier)

5. *Player Strength*, allows for more granularity in assessing player performance

We also hope to assess the validity of the model by exploring intergame data. We hypothesize that modeling goal scoring as poisson is not supported by data, and we hope to fit a more appropriate distribution to the model (such as one distributed over a more reasonable support). Lastly, we hope to investigate the accuracy of the validity of the FIFA league coefficients.

References

[1] Mark J. Dixon, Stuart G. Coles *Modeling Association Football Scores and Inefficiencies in the Football Betting Market* Applied Statistics, Volume 46, Issue 2 (1997) 265-280

[2] http://www.football-data.co.uk/englandm.php

[3] EPL iPython Notebook

[4] Constantinou, A., N. E. Fenton and M. Neil (2013) *Profiting from an Inefficient Association Football Gambling Market: Prediction, Risk and Uncertainty Using Bayesian Networks.*

[5] Constantinou, A., N. E. Fenton and M. Neil (2012). *pi-football: A Bayesian network model for forecasting Association Football match outcomes.*

[6] Constantinou, A. , Fenton, N.E., *Solving the problem of inadequate scoring rules for assessing probabilistic football forecasting models*, Journal of Quantitative Analysis in Sports, Vol. 8, Article 1, 2012.