Applied Math 207 Final Project:
Predicting Scoring Outcomes in EPL Soccer

Akhil Ketkar,[1] Owen V. Prunskis[2]

[1]Institute for Applied Computational Science, Harvard University,
[2]School of Engineering and Applied Sciences, Harvard University

Correspondence should be addressed to:
akhilketkar@g.harvard.edu, prunskis@college.harvard.edu

The paper presents a Poisson/log-normal Bayesian hierarchical clustering model to predict scores and outcomes for English Premier League (EPL) soccer games. The model represents a significant improvement to Baio-Blangiardo[2]. We use EPL data from 2013-2014 and 2014-2015 to train and evaluate the model. We examine robustness by applying the model to different soccer leagues and different sports. We find that the model is highly effective in predicting season scoring and proves robust when applied to other athletic disciplines, outside of the EPL.

Introduction

Soccer is the most popular sport in the world, particularly in Europe. The plethora of in-depth records makes the sport a natural target for statisticians and mathematicians to flex their intellectual muscle in predicting match outcomes, scoring differentials, season champions, and far more esoteric metrics.

Dixon-Coles[3] 1997 paper uses a Poisson process to model the number of goals scored by each team and finds MLEs for their parameters. Baio-Blangiardo proposes a Bayesian hierarchical model for predicting the number of goals scored per game drawing from a bivariate Poisson distribution, which considers attack and defense strength and home field effects [2]. The Weitzenfeld expansion notably adds an intercept term to capture average goals score by the away team [4]. We implement and improve upon these models and apply it to English Premier League (EPL) soccer.

English Premier League (EPL) The EPL is the foremost soccer league in the world consisting of the best teams in England, including the world-renowned Chelsea, Arsenal, Manchester United, and Liverpool clubs. The structure of the league is as follows: there are 20 teams which play one another twice per season -yielding 380 total matches; points are awarded based on game outcomes with three points awarded for a victory and one point for a draw; and at the end of each season, the team with the most points is crowned champion, while the bottom three finishing teams are relegated to a lower league to make room for three of the highest performing teams in the lower division to enter the EPL the subsequent year. Our model incorporates sport-agnostic attack and defense ratings as well as EPL-specific nuances in efforts to accurately predict goal-scoring, and thereby fixture outcomes.

Related Work / Libraries Used

As cited in the previous section, much work has been done in the field of predicting sports outcomes, particularly for soccer. The model presented in this paper is informed by the work of Dixon-Coles [3], Baio-Blangiardo[2], and the Baio-Blangiardo-Weitzenfeld expansion[4]. We leverage this previous work to design and implement an improved Poisson/log-normal Bayesian hierarchical clustering model to predict game outcomes.

In our implementation, we leverage the following python libraries: `datetime, math,`

`numpy, matplotlib, os, pandas, pymc, urllib2, warnings.`

Methods

We implement an illustrative toy model using our own sampling procedure to provide intuition for the procedure specified by the `pymc` library. We go on to use `pymc` for subsequent additions given the conciseness of the package functions. In total, we implement six models for EPL soccer which represent successive improvements with respect to the predecessor.

Mod I. We begin by implementing the original Baio-Blangiardo model, which yields results consistent to those in the original paper.

Mod II. We then apply the Weitzenfeld expansion, which incorporates an intercept to capture average goals scored.

Mod III. Next, we stratify the intercept term based on three distinct ability cohorts, as defined by the results of the previous season.

Mod IV. Next, we stratify home effects for each team; this is met with lackluster results and is omitted from subsequent models.

Mod V. We go on to include parameters for championship-race consideration, European championship standing, and relegation risk.

Mod VI. We evaluate the significance of the additional parameters from **Mod V** and determine that only Championship-race considerations are significant. We simplify the model accordingly, to yield our final truncated Poisson/Log-Normal model.

The Poisson/Log-Normal is most appropriate due to the discrete nature of the goal-scoring process, the support over the positive integer domain (as one can't have negative or fractional goals), and the fact that our mean and variance are unknown, all of which combine to make

the Poisson/Log-Normal more appropriate than other distributions typical in rare event modelling, such as the normal-normal and normal conjugate gamma distributions. Shortcomings include the independence assumption between goals within games; however, we do not detect any significant effects caused by violation of this assumption.

Model Specifications Number of goals scored by home team in game g is denoted by $y_{home}^{(g)}$. The number of goals scored by the away team in game g is denoted by $y_{away}^{(g)}$. The y's are assumed to follow a Poisson distribution with an unknown rate parameter.

$$y_{home}|\theta_{home} \sim Poisson(\theta_{home})$$

$$y_{away}|\theta_{away} \sim Poisson(\theta_{away})$$

The rate parameter is assumed to depend on some underlying ability of each time. In the definitions below $h(g)$ denotes the home team for game g and $a(g)$ denotes the away team.

$$log(\theta_{home}) = home + intercept + att_{h(g)} + def_{a(g)}$$

$$log(\theta_{away}) = intercept + att_{a(g)} + def_{h(g)}$$

The att_t and def_t for each team are assumed to be drawn from a normal distribution with mean 0. τ here represents precision

$$att_t \sim N(0, 1/\tau_{att})$$

$$def_f \sim N(0, 1/\tau_{def})$$

The model defines non-informative hyper priors on τ, $home$ and $intercept$

$$\tau_{att} \sim Gamma(0.1, 0.1) \qquad \tau_{def} \sim Beta(0.1, 0.1)$$

$$home \sim N(0, 1000) \qquad intercept \sim N(0, 1000)$$

Our Updates

The updated, truncated model becomes:

$$log(\theta_{home}) = home + intercept_{c(h(g))} + att_{h(g)} + def_{a(g)} + champ_{g,h(g)} * att_{champ} + champ_{g,a(g)} * def_{champ}$$

$$log(\theta_{away}) = intercept_{c(a(g))} + att_{a(g)} + def_{h(g)} + champ_{g,a(g)} * att_{champ} + champ_{g,h(g)} * def_{champ}$$

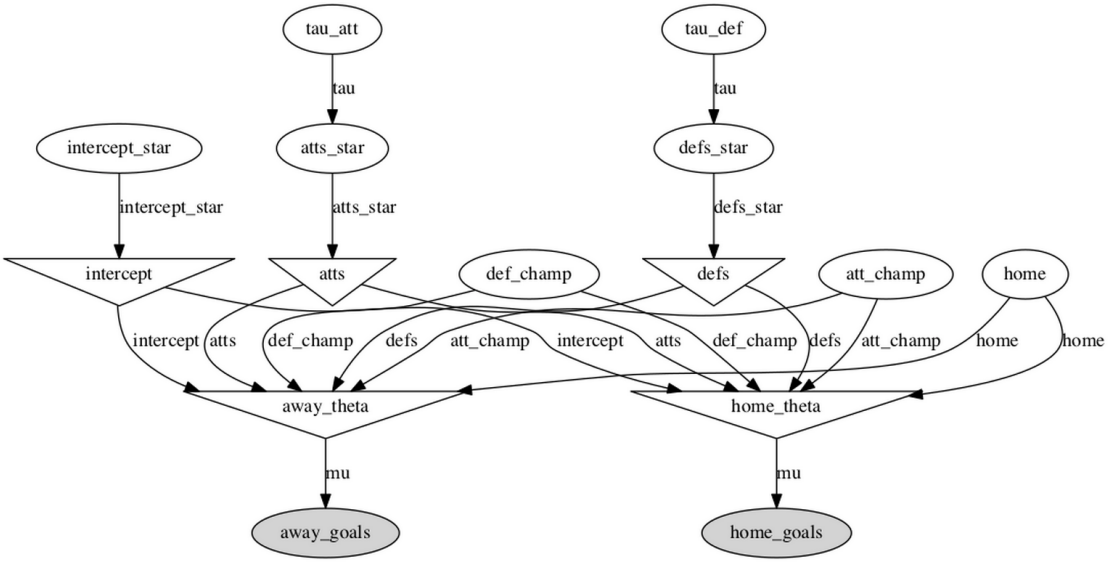
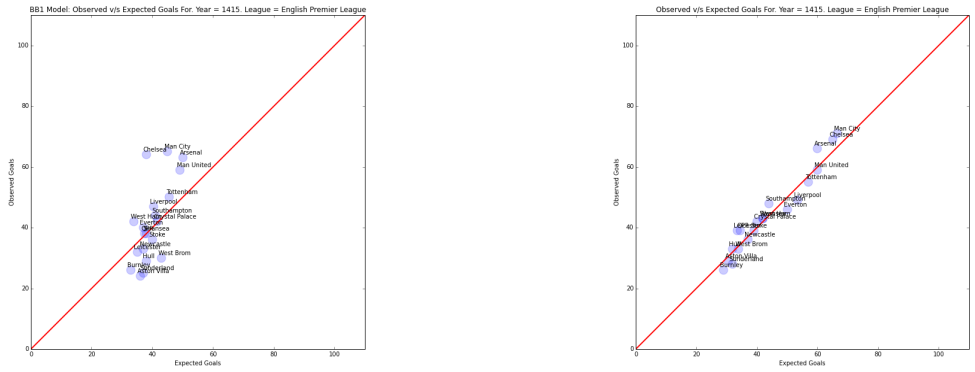


Figure 1: DAG Representation of Hierarchical Model VI

with each of the k intercepts $intercept_m \sim N(0, 1000)$ and championship-race parameters $(champ_{g,a(g)} * att_{champ})$ and $(champ_{g,h(g)} * def_{champ})$

Results

We note that the model represents a significant improvement:



(a) BB Predictions for 2014-2015 EPL (b) Improved Predictions for 2014-2015 EPL

Figure 2: Side-by-Side Comparison of Model Predictions

The model also allows week-by-week, match-level predictions, as exhibited:

Home	Away	Prob. Home Win	Prob. Away Win	Prob. Draw	Mean Goals Home	Mean Goals Away
Arsenal	Swansea	0.634	0.157	0.209	2.117	0.900
Aston Villa	West Ham	0.316	0.412	0.272	0.963	1.161
Chelsea	Liverpool	0.605	0.183	0.212	1.931	0.960
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Stoke	Tottenham	0.379	0.358	0.263	1.374	1.309

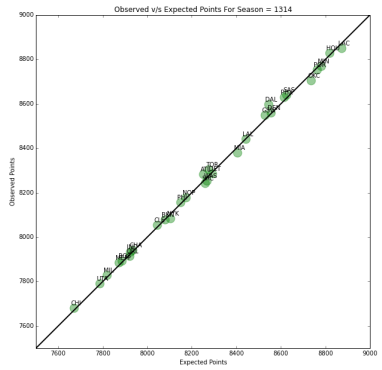
Table 1: Illustrative Match Predictions for May 9, 2015

Discussion

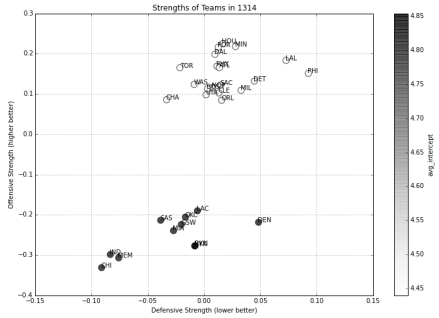
The improved model addresses the shortcomings of the BB model:

1. shrinkage to the mean alleviated $\rightarrow$ observed vs. expected goal plot reveals less clustering around mean
2. missing variables $\rightarrow$ addition of several significant parameters
3. sub-optimal approach to mean goals scored $\rightarrow$ stratification of intercept term
4. static ability for a season $\rightarrow$ incorporation of championship-race parameter

Additionally, the model is evaluated for robustness across other soccer leagues (Serie A, La Liga) and other sports (MLB, NBA), with encouraging results. Surprisingly, basketball, despite having a significantly different scoring system, exhibits an extremely high level of predictability with respect to our Poisson model, despite having extremely frequent scoring events. Analysis of the offensive and defense strength matrices reveals an extreme degree of separation between ”good” and ”bad” teams, which has direct correspondence to the stratified intercept term. This effective delineation coupled with the large sample size of training data allows the model to be robust in a completely different sport with a seemingly disparate scoring system.



(a) Model Predictions for 2013-2014 NBA



(b) Relative Strength Matrix for 2013-2014 NBA

Figure 3: Evidence of robustness of model in NBA basketball

Conclusion

Overall, our model seems very robust to different leagues and across different sports but there are various ways to improve upon our model such as:

1. Accounting for changes within the team intra-season
2. Accounting for changes in coaching
3. Accounting for different substitutions or line-ups that play in any particular game

We suspect that further improvements would allow for more accurate predictions of match-level results, as the model has fairly high predictive power in aggregate.

Acknowledgements

We thank Patrick Ohiomoba for his enthusiastic support and guidance in every step of the project and Verena Kaynig-Fittkau for her helpful feedback.

References

[1] Aitchinson, J. Ho, C. (1989), The multivariate poisson-log normal distribution, *Biometrika* 76, 643653.

[2] Baio,G. Blangiardo M.A (2006) Bayesian hierarchical model for the prediction of football results *Statistica*

[3] Dixon, M. Coles, S. (1997), 'Modelling association football scores and inefficiencies in the football betting market', *Journal of the Royal Statistical Society* C 46, 265280.

[4] Weitzenfeld, D. 2014. A hierarchical bayesian model of the premier league. *Pass the ROC*.

Data

www.baseball-reference.com
www.basketball-reference.com
www.football-data.co.uk