Beating the bookie
A look at statistical models for prediction of football matches

Helge Langseth

Norwegian University of Science and Technology

SCAI 2013

NTNU
Building a model for match outcomes

- Suppose we want to build a model to predict the outcomes of the English Premier League:
  - 20 teams, all play each other twice during a season.
  - Each team plays 38 matches, 380 games per season in total.

- The model should predict the outcome of Team $i$ meeting Team $j$, based on all games played previously in the season.

- The quality is measured by the system's ability to win bets.
  - A bet (e.g., “Liverpool to win”) is offered with odds $\omega$.
  - The model generates the corresponding probability $p$.
  - A bet is only rational whenever the expected gain is positive, i.e., $p \cdot \omega \geq 1$.
  - Accurate predictions imply a useful betting agent, thus our goal is to generate good probability estimates for upcoming games based on the history of the season so far.
An early attempt at building a statistical model:

- \( X_{ij} \sim \text{Poisson}(k \cdot \lambda \cdot \alpha_i \beta_j) \), where:
  - \( X_{ij} \) is no. goals scored by Team \( i \) vs. Team \( j \) playing at home.
  - \( k \) captures the home-team advantage.
  - \( \lambda \) is a normalization constant.
  - \( \alpha_i \) is the **attacking** strength of Team \( i \).
  - \( \beta_j \) is the **defending** strength of Team \( j \).

- \( Y_{ij} \sim \text{Poisson}(\alpha_j \beta_i) \); \( Y_{ij} \) is no. goals scored by Team \( j \).
- Crucially — and surprisingly — he assumes \( X_{ij} \perp \perp Y_{ij} | \lambda \).
- The model is under-specified, so he requires
  \[
  \text{avg}_\ell (\alpha_\ell) = \text{avg}_\ell (\beta_\ell) = 1.
  \]
An early attempt at building a statistical model:

- \( X_{ij} \sim \text{Poisson}(k \cdot \lambda \cdot \alpha_i \beta_j) \), where:
  - \( X_{ij} \) is no. goals scored by Team \( i \) vs. Team \( j \) playing at home.
  - \( k \) captures the home-team advantage.
  - \( \lambda \) is a normalization constant.
  - \( \alpha_i \) is the **attacking** strength of Team \( i \).
  - \( \beta_j \) is the **defending** strength of Team \( j \).

- \( Y_{ij} \sim \text{Poisson}(\alpha_j \beta_i) \); \( Y_{ij} \) is no. goals scored by Team \( j \).
We predict the result of the game between Team $k$ and Team $\ell$ by looking at the probability distributions for $X_{k\ell}$ and $Y_{k\ell}$.

The estimated abilities of the two best teams in the Premier league after 11 rounds are:

<table>
<thead>
<tr>
<th></th>
<th>Attack</th>
<th>Defence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsenal</td>
<td>1.42</td>
<td>0.87</td>
</tr>
<tr>
<td>Liverpool</td>
<td>1.38</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Using these parameters we can look at the joint distribution

$$P(X_{\text{Liv,Ars}}, Y_{\text{Liv,Ars}}).$$
Predictions from the model

- We predict the result of the game between Team \( k \) and Team \( \ell \) by looking at the probability distributions for \( X_{k\ell} \) and \( Y_{k\ell} \).
- The **estimated abilities** of the two best teams in the Premier league after **5 and 11 rounds** are:

<table>
<thead>
<tr>
<th>Team</th>
<th>After 5 games</th>
<th>After 11 games</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Attack</td>
<td>Defence</td>
</tr>
<tr>
<td>Arsenal</td>
<td>1.46</td>
<td>1.23</td>
</tr>
<tr>
<td>Liverpool</td>
<td>1.02</td>
<td>0.69</td>
</tr>
</tbody>
</table>

**Abilities change over time, so we need a dynamic model!!**
Adding dynamics

- We follow, e.g., Rue & Salvesen (2000) and introduce dynamics at the “strength-level”:
  - Let $\alpha_i^{(t)}$ be the attack-strength for Team $i$ at time $t$.
  - Then, $\alpha_i^{(t_1)} \mid \{ \alpha_i^{(t_2)} \} \sim N \left( \alpha_i^{(t_2)}, \frac{|t_1-t_2|}{\tau} \sigma^2 \right)$
  - Similarly for the defence-strength, $\beta_i^{(t)}$.

- **Hidden Markov** - type model: Unobserved strengths varying over time; partially disclosed through goal-model.

- Assume we observe the result when Team $i$ and Team $j$:
  - The chains of these teams **get correlated**.
  - Similarly, the strengths of all teams Team $i$ and Team $j$ have played previously **get correlated, too!**

- We use **Markov Chain Monte Carlo** to find estimators for the model parameters, and sample results for unseen matches.
We follow, e.g., Rue & Salvesen (2000) and introduce dynamics at the “strength-level”:
- Let $\alpha_i^{(t)}$ be the attack-strength for Team $i$ at time $t$.
- Then, $\alpha_i^{(t_1)} \mid \{\alpha_i^{(t_2)}\} \sim N \left( \alpha_i^{(t_2)}, \frac{|t_1 - t_2|}{\tau} \sigma^2 \right)$
- Similarly for the defence-strength, $\beta_i^{(t)}$. 

\[
\begin{align*}
\alpha_i^{(t)} & \rightarrow \alpha_i^{(t)} \\
\beta_i^{(t)} & \rightarrow \beta_i^{(t)} \\
\alpha_j^{(t)} & \rightarrow \alpha_j^{(t)} \\
\beta_j^{(t)} & \rightarrow \beta_j^{(t)} \\
X_{ij} & \rightarrow \alpha_i^{(t)} \rightarrow X_{ij} \\
Y_{ij} & \rightarrow \beta_i^{(t)} \rightarrow Y_{ij} \\
k & \rightarrow \lambda \rightarrow t
\end{align*}
\]
Small margins can influence the result of a football match significantly.

This inherent randomness makes the estimation of $\alpha_i(t)$ and $\beta_i(t)$ difficult, and the “signal-to-noise-ratio” is typically small.

More data, that “look behind the result”, e.g.,

- Shots on goals
- Possession statistics
- Passing accuracy

...can be useful to uncover the teams’ underlying abilities.

Estimated defensive strength for *Arsenal* over the 2010-11 season.
Data-intensive model

Here we use:

- $\lambda_i^H$: Chance creation rate; home
- $C_{ij}$: No. chances.
- $F_{ij}$: No. shots.
- $X_{ij}$: No. goals.
- $\alpha_i^{(t)}$: The attacking strength.
- $\beta_j^{(t)}$: The defensive strength.
- $\gamma_j^{(t)}$: The goalkeeper strength.
Consider a bet with offered odds $\omega$ and estimated winning probability $p$.

We require the expected gain to be non-negative, i.e., $p \cdot \omega \geq 1$.

Consider the two bet-options

- **Bet A:** $\omega_A = 11.0$, $p_A = 0.1$.
- **Bet B:** $\omega_B = 1.10$, $p_B = 1.0$.

Both bets have the same expected return of 1.1 unit per unit staked, but obviously Bet B is preferrable.

It is important to consider money management carefully!

Many strategies exist, we have considered, e.g., Fixed Bet, Fixed Return, Kelly’s Rule and Rue’s Variance Adjustment.
### Premier League 2011-2012

<table>
<thead>
<tr>
<th>Model</th>
<th>Fixed Bet</th>
<th>Fixed Return</th>
<th>Kelly</th>
<th>Var. Adjust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>17.4%</td>
<td>17.4%</td>
<td><strong>23.2%</strong></td>
<td><strong>15.6%</strong></td>
</tr>
<tr>
<td>Dynamic</td>
<td><strong>22.7%</strong></td>
<td>14.3%</td>
<td>21.3%</td>
<td>12.0%</td>
</tr>
<tr>
<td>DataIntensive</td>
<td>20.3%</td>
<td><strong>24.2%</strong></td>
<td>23.0%</td>
<td>14.3%</td>
</tr>
</tbody>
</table>

### Premier League 2012-2013

<table>
<thead>
<tr>
<th>Model</th>
<th>Fixed Bet</th>
<th>Fixed Return</th>
<th>Kelly</th>
<th>Var. Adjust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>-23.7%</td>
<td>-24.9%</td>
<td>-27.8%</td>
<td>-21.2%</td>
</tr>
<tr>
<td>Dynamic</td>
<td>-17.1%</td>
<td>-20.0%</td>
<td>-22.9%</td>
<td>-15.9%</td>
</tr>
<tr>
<td>DataIntensive</td>
<td>-6.3%</td>
<td>-0.7%</td>
<td>-3.4%</td>
<td><strong>0.4%</strong></td>
</tr>
</tbody>
</table>

**2010-2011:** Results similar, but **DataIntensive** combined with **FixedReturn** gives best result.

**2011-2012:** Only **DataIntensive** combined with **Var. Adjust** beats the bookie.
Rue’s Variance Adjustment has been the most robust money management strategy.

The goal is to minimize the expected profit minus the variance of the profit, leading to bets defined by $C \propto \frac{1}{\omega(1-p)}$.

In contrast to most other money management strategies, the amount wagered therefore is decreasing in $\omega$. 
Although we are looking at **betting agents**, and not simple **classifiers**, improving prediction is beneficial:

- Build models that **incorporate more game-information**; data can be harvested, e.g., from [http://www.whoscored.com](http://www.whoscored.com).
- Combine the **ensemble** of different candidate models into one prediction-engine.
- Utilize pre-game information about **line-ups** to enhance the predictions.

Simulate results for **more leagues** – with the aim of understanding why some leagues are easier to generate profits from than others.

Replace MCMC simulations with fast approximate Bayesian inference based on **variational approximations**.