The Statistical Analysis and Modelling of Australian Rules Football

The Story of MAFL (and a Badly-Bred Pomeranian)

August 2011
Can You Eat a MAFL?

- MAFL = Model AFL (courtesy Prosaic Names R’ Us)
- Commenced around 2006
- **Philosophy:** the results of AFL Australian Rules football games can be predicted “reasonably well” on the basis of a limited set of variables

**Success Criteria:**
- As well as a recognised expert
- Well enough to profit from wagering on the outcome

www.maflonline.squarespace.com (or bit.ly/maflo)
Aussie Rules is a game played between two teams of 22 players, 18 of which are on the field at any one time.

Played over 4 x 20-minute quarters and with a 30 minute half-time interval, each game lasts for about 3 hours (longer if you’ve wagered on the losing team).

Two methods of scoring:

Currently there are 17 teams in the competition and the home-and-away season lasts for 24 weeks.
Over the history of Australian Rules, winning teams have scored, on average, just over 100 points, while losing teams have scored about 70.

The average margin of victory had tended to grow for about the last century, though it is now showing signs of stabilising.
MAFL makes two types of wagers

Head-to-head wagers
• Wagering on which team will win
• Bookmaker quotes markets such as
  Carlton $1.36 / Geelong $3.00

A wager on Carlton pays $1.36 for every $1 wagered if Carlton wins
A wager on Geelong pays $3.00 for every $1 wagered if Geelong wins

Line wagers
• Wagering on which team will win after a handicap is applied
• Bookmaker quotes markets such as
  Carlton (-20.5) $1.90 / Geelong $1.90

A wager on Carlton pays $1.90 for every $1 wagered if Carlton wins by 21 points or more
A wager on Geelong pays $1.90 for every $1 wagered if Carlton wins by 20 points or fewer, Geelong wins, or the game ends in a draw
Overround – The Bookmaker’s Friend

- Consider the head-to-head market described before: Carlton $1.36 / Geelong $3.00

- At these prices no combination of wagers is capable of generating a profit for a bettor regardless of the outcome.

- This will always be the case if:

\[
\frac{1}{\text{Team 1 Price}} + \frac{1}{\text{Team 2 Price}} \geq 1
\]

- In this example this is

\[
\frac{1}{1.36} + \frac{1}{3.00} = 1.07
\]

1.07 (or 107%) is referred to as the bookmaker’s overround and is a measure of the guaranteed profitability to the bookmaker if the prices offered attract wagers in inverse proportion to those prices.

**TL;DR: You can’t make money betting both sides of the market.**

The bookmaker doesn’t (necessarily) want to predict who’ll win, just set prices that will attract the right proportion of wagers on each team.

* If \( \frac{\text{Team 1 Price}}{\text{Team 2 Price}} = \frac{\text{Wagers on Team 2}}{\text{Wagers on Team 1}} \) then the bookmaker’s guaranteed return is \( (1 - \frac{1}{\text{Overround}}) \times \text{Total Amount Wagered} \)
R Packages Used (Directly or Indirectly) To Produce The Results Shown in This Deck

base
glm
ROCR
cluster
ggplot2
relaimpo
hier.part
party
lattice
What variables seem to have information content for predicting the team that will win head-to-head?

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Variable Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert Opinion</td>
<td>• Bookmaker’s head-to-head Price</td>
</tr>
<tr>
<td>Home Ground Advantage</td>
<td>• Teams’ recent experience at the venue</td>
</tr>
<tr>
<td></td>
<td>• Whether or not the game is an interstate clash</td>
</tr>
<tr>
<td>Relative Team Quality</td>
<td>• Each team’s recent form (measured by points scored less points conceded over recent games)</td>
</tr>
<tr>
<td></td>
<td>• Team Rating</td>
</tr>
</tbody>
</table>
Expert Opinion

• The bookmaker’s head-to-head prices can be converted into implicit probabilities as follows:

  Implicit probability that Home team wins = \frac{\text{Away Team Price}}{\text{Home Team Price} + \text{Away Team Price}}

  eg \quad \frac{1.36}{3.00 + 1.36} = 31\%

• A well-calibrated bookmaker is one for whom teams with an implicit probability of victory equal to T% win about T% of the time.

• Bookmakers who are poorly calibrated are exploitable.

• Exploitable bookmakers find other jobs.

• Ergo a surviving bookmaker’s opinions about the chances of each team, as embodied in the prices offered, are reasonable Bayesian priors.

**Comparison of Bookmaker’s Implied Home Team Probability and Win Rate (2006-2010)**

<table>
<thead>
<tr>
<th>Implicit Probability Range</th>
<th>Average Implicit Probability</th>
<th>Win Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 10%</td>
<td>9%</td>
<td>4%</td>
</tr>
<tr>
<td>10% &lt; 20%</td>
<td>16%</td>
<td>12%</td>
</tr>
<tr>
<td>20% &lt; 30%</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td>30% &lt; 40%</td>
<td>35%</td>
<td>34%</td>
</tr>
<tr>
<td>40% &lt; 50%</td>
<td>45%</td>
<td>43%</td>
</tr>
<tr>
<td>50% &lt; 60%</td>
<td>54%</td>
<td>52%</td>
</tr>
<tr>
<td>60% &lt; 70%</td>
<td>65%</td>
<td>58%</td>
</tr>
<tr>
<td>70% &lt; 80%</td>
<td>75%</td>
<td>72%</td>
</tr>
<tr>
<td>80% &lt; 90%</td>
<td>84%</td>
<td>87%</td>
</tr>
<tr>
<td>90% or more</td>
<td>91%</td>
<td>93%</td>
</tr>
</tbody>
</table>

Evidence of a favourite-longshot bias
Home Ground Advantage

• Playing at home has always been a distinctive advantage in Aussie Rules, as in most sports

• Home teams have, in most seasons, typically won about 55-60% of contests

• The advantage appears to be worth about 5-10 points per game

• Recent trend suggests that this advantage is diminishing, probably due to:
  • Shared home grounds
  • Lesser strength of teams from outside Melbourne
Venue Experience

Home Team Win Rate for Given Difference in Venue Experience (2000-2010)

<table>
<thead>
<tr>
<th>Excess Venue Experience</th>
<th>% Win</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6 or less</td>
<td>44.2%</td>
</tr>
<tr>
<td>-5 to -1</td>
<td>54.0%</td>
</tr>
<tr>
<td>0 to +6</td>
<td>54.2%</td>
</tr>
<tr>
<td>+7 to +9</td>
<td>59.0%</td>
</tr>
<tr>
<td>+10 or more</td>
<td>67.3%</td>
</tr>
<tr>
<td>All</td>
<td>59.1%</td>
</tr>
</tbody>
</table>

• Excess Venue Experience is a more nuanced way to model Home Ground Advantage, especially given the advent of shared home grounds

• A Home team’s Excess Venue Experience is defined to be: The number of times it has played at the venue for the current game during the past 12 calendar months less the number of times its opponent has played there in this period
Recent Form

- Most teams are prone to winning and losing streaks such that a win is more likely to be followed by another win, and a loss by another loss.

- By “more likely” we mean relative to a team’s overall winning or losing rate.

- Of the 16 teams that played during the 10 seasons 2000 to 2009, 14 of them were more likely to win after a victory in the immediately preceding round, and 15 were more likely to lose after a loss in the immediately preceding round.

For modelling purposes we codify recent form as the difference between points scored and points conceded in the most recent X rounds.
Team Ratings (MARS Ratings)

ELO Style Rating System
A team’s new rating is based on its old rating plus an amount dependent on its performance relative to expectations.

A team’s Expected Outcome is a function of the difference between its team rating, its opponent’s and whether it is playing at home or away.

A team’s Actual Outcome is a function of the margin by which it won or lost.

‘Blowout’ victories shouldn’t be allowed to effect team ratings too dramatically.

\[ R_{\text{new}} = R_{\text{old}} + 37.5 (\text{Actual Outcome} - \text{Expected Outcome}) \]

\[ \text{Expected Outcome} = \frac{1}{1 + 10^{(R_a - R_b \pm 6)/550}} \]

(we use -6 for an Away Team and +6 for a Home Team)

\[ \text{Winning Team’s Actual Outcome} = 0.99 - 0.49^{(1 + \text{Margin}/130)} \]

\[ \text{Losing Team’s Actual Outcome} = 1 - \text{Winning Team’s Actual Outcome} \]

\[ \text{If Actual Margin} > 78, \text{ set Margin} = 78 \]

Initial Ratings
- All teams’ ratings were set to 1,000 at the start of the 1999 season
- In subsequent seasons, each team’s initial ratings are set to 530 + 47% x Rating at end of previous season
Together, the equation for the Expected Outcome as a function of MARS Ratings ...

... and the equation for the Actual Outcome as a function of game victory margin ...

... imply a minimum victory margin required to preserve a team’s current rating
library(lattice)

xyplot(Own_MARS_Rating ~ (Master_Round_Num - 27) | Team, data = fd_to_plot,
xlab="Master Round Number (1=R1 2000)", ylab="Team Rating", cex=0.45, type = "l", lwd=3)
A binary logit fitted to the difference in team MARS Ratings alone can be used to estimate the probability of a Home team victory.

Using a naïve cut-off of 0.5, the binary logit fitted to the period 2000 to 2008 has out-tipped the bookmaker for 2009 and 2010.
A Conditional Inference Tree using Team Ratings, Venue Experience, Interstate Status, Form and Implicit Bookmaker Probabilities provides a significantly superior fit, as measured by predictive accuracy, to the results for seasons 2000-2008 ...

library(party)

\[
\text{fd_model_cit} = \text{cforest(Result} \sim \cdot, \ \text{data} = \text{fd_use}, \ \text{controls} = \text{cforest_unbiased()})
\]
... though this superiority doesn’t carry over post sample into 2009 and 2010 (it’s still not bad though).

At a 0.5 cutoff, the most accurate model is the binary logit that **doesn’t** use bookmaker probabilities.
Say I’ve decided that I want to make wagers in the head-to-head markets.

Intuitively I might think that I need to predict which team will win.

But, imagine I’m right 75% of the time when I wager on teams paying $1.10 head-to-head.

\[
\text{Average Return} = 75\% \times 0.10 + 25\% \times (-1.00) = -0.175
\]

What I need is the ability to accurately \textbf{predict the probability that a team will win}, to which I can then apply the offered prices to calculated the expected return.

Only if the expected return is positive do I wager.
One way to assess the accuracy of probability predictions for binary events is a logarithmic probability score, defined here as:

\[
\text{Probability Score} = 1 + \log_2(\text{Probability Assigned to Winning Team})
\]

*For draws we use:

\[
1 + 0.5 \times \left[ \log_2(\text{Probability Assigned to Winning Team}) + \log_2(1 - \text{Probability Assigned to Winning Team}) \right]
\]

(It’s not hard to prove that a predictor’s probability score is optimised by assigning a probability \( p \) to a team that’s true probability of victory is \( p \))
The Binary Logit using team ratings, relative venue experience, the interstate flag and the teams’ relative recent form performs best in the testing period, 2009 and 2010.
Other Things We’ve Learned
Losing (Probably) Doesn’t Lead to Winning

Across the history of VFL/AFL from 1897 to 2008, teams that have led at the end of any quarter have been more likely to go on to win than to lose ...

... if we look only at the period from 1980 to 2008 however, teams that led at quarter time went on to win only 44.5% of the time, which is statistically significant (for $H_0: p=0.5$) at the 10% level.

In aggregate the proportion of goals and behinds scored by winning teams has been remarkably consistent across almost 100 years.
Grand Final Margins Have Been Increasing

Note how the ‘blowout’ results have been a relatively recent phenomenon and have almost exclusively been a team finishing higher on the ladder in the home-and-away season defeating a team finishing lower on the ladder.
There Are Five Types of Grand Finals …

The clustering solution displayed at left was built using the quarter-by-quarter margin data from all Grand Finals between 1897 and 2008 and the pam clustering algorithm from R.

package(cluster)
library(pam)
... and Eight Types of Home-and-Away Games

### The 8 Types of Home and Away Games (1897-2010)

<table>
<thead>
<tr>
<th>Winning team's typical lead at end of</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>% of Games</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quarter 1 Press</td>
<td>26</td>
<td>17</td>
<td>17</td>
<td>18</td>
<td>8.4%</td>
</tr>
<tr>
<td>Quarter 2 Press</td>
<td>-8</td>
<td>27</td>
<td>25</td>
<td>39</td>
<td>7.3%</td>
</tr>
<tr>
<td>Quarter 2 Press Light</td>
<td>-5</td>
<td>11</td>
<td>6</td>
<td>15</td>
<td>19.1%</td>
</tr>
<tr>
<td>Quarter 3 Press</td>
<td>19</td>
<td>10</td>
<td>40</td>
<td>43</td>
<td>10.4%</td>
</tr>
<tr>
<td>2nd-Half Revival</td>
<td>-1</td>
<td>-4</td>
<td>18</td>
<td>19</td>
<td>12.9%</td>
</tr>
<tr>
<td>Coast-to-Coast Nail-Biter</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td>15.8%</td>
</tr>
<tr>
<td>Coast-to-Coast Comfortably</td>
<td>15</td>
<td>27</td>
<td>36</td>
<td>46</td>
<td>19.3%</td>
</tr>
<tr>
<td>Coast-to-Coast Blowout</td>
<td>13</td>
<td>39</td>
<td>63</td>
<td>86</td>
<td>6.9%</td>
</tr>
</tbody>
</table>

Principal co-ordinates analysis on proportion of each game type by season:

```r
era_profile = t(prop.table(table(solution$clustering,fd$Era),margin=2))
-era_diss = daisy(era_profile,stand=FALSE)
era_plot = cmdscale(era_diss)
x = era_plot[,1]; y = era_plot[,2]
plot(x, y, type="n", xlab="", ylab="", main="Eras 1897 to 2010")
text(x, y, rownames(era_plot), cex=0.8)
```

... same analysis using proportion of each game type by era
The model is based on Brownian motion (no kidding) and the exact fitted equation is:

\[
\Pr(\text{HomeTeamWins}) = \frac{\left(e^{-0.1569 + \frac{0.0567L}{\sqrt{1-t}} + 0.5184\sqrt{1-t}}\right)}{\left(1 + e^{-0.1569 + \frac{0.0567L}{\sqrt{1-t}} + 0.5184\sqrt{1-t}}\right)}
\]

\(L = \text{Home Team Lead, and}\)

\(T = \text{Time Elapsed in Game (as a proportion of the total game, so } 0 \leq T \leq 1)\)

See: A Brownian Motion Model for the Progress of Sports Scores
Modelling Victory As a Function of Lead and Time Remaining

We can restate the fitted model to define isopros – lines joining leads (deficits) representing the same probability of victory for the Home team.

The model fit is excellent.
Win Production Function

- Using just four summary scoring statistics for a team we can explain 89% of the variability in its winning percentage for an entire season.
- A single “Win Production Function” was fitted to all 114 seasons.

\[
E(TeamWin\%) = \frac{e^{0.164(SSF-SSA)+6.18(OppConv-OppConv)}}{1+e^{0.164(SSF-SSA)+6.18(OppConv-OppConv)}}
\]

SSF = Own Average Scoring Shots (ie Goals + Behinds) per Game
SSA = Opponents’ Average Scoring Shots per Game
OwnConv = Own Goals Scored / (Own Goals Scored + Own Behinds Scored)
OppConv = Opponents’ Goals Scored / (Opponents’ Goals Scored + Opponents’ Behinds Scored)
Reflections on the Journey
MAFL as a Microcosm of Modelling

• Maybe if I had one more variable ...
  • Weather/ key player availability /

• I should be using something simpler/different/more robust instead of a Conditional Inference Tree / Binary Logit / <insert model type here> ...

• Maybe if I expressed that variable in a different way ...

• Things have changed so much this year compared to last ...
  • There’s a new team
  • The rules have changed
  • Now there are byes
  • Home ground advantage is dissipating
  • It’s a new bookmaker

• I’ve just been lucky when I’ve made money ...

• I really shouldn’t start wagering until I’m sure the models have had a chance to adapt to the current season ....

• I’m missing out on opportunities by not wagering this week ...

• Why is/isn’t the model doing that ...

Maybe It Is All Just Chance

Even exceptional probability assessment prowess doesn’t guarantee profit ...

Someone who predicts line results at a stunningly good rate of (on average) 55% can expect to make a profit across 100 bets paid out at $1.90 fewer than 7 seasons in 10.
WHAT’S DIFFERENT

Analysing and modelling football

• provides an opportunity to create well-defined problems with clear and objective measures of success.

• requires answers that are probabilities; here such answers aren’t treated as statements of ignorance and timidity.

WHAT’S SIMILAR

In analysing and modelling football, as in business

• there is accepted wisdom that is amenable to data-driven validation (and that’s often wrong).

• it’s more than finding the right variables, it’s finding the right way to express those variables that provides the most predictive power or the greatest insight.
## R Packages Used in MAFL

<table>
<thead>
<tr>
<th>R Package (Function)</th>
<th>Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>glm</td>
<td>Create binary logits for game result</td>
</tr>
<tr>
<td></td>
<td>Create binary logit for in-running probability assessment</td>
</tr>
<tr>
<td>ROCR (performance)</td>
<td>Plot performance of binary logit models</td>
</tr>
<tr>
<td>cluster (pam)</td>
<td>Cluster game types, season types and epoch types</td>
</tr>
<tr>
<td>ggplot2</td>
<td>Chart Grand Final margin history</td>
</tr>
<tr>
<td>relaimpo (calc.relimp)</td>
<td>Measure variable importance in regressions in the presence of correlation</td>
</tr>
<tr>
<td>hier.part</td>
<td>Measure variable importance in binary logits in the presence of correlation</td>
</tr>
<tr>
<td>party (cforest)</td>
<td>Create Conditional Inference Trees for predicting game results</td>
</tr>
<tr>
<td>BradleyTerry2 (BTM)</td>
<td>Assessment of team strength based on head-to-head results</td>
</tr>
<tr>
<td>earth</td>
<td>Investigating the existence of non-linearities in the regression relationships</td>
</tr>
<tr>
<td>RWeka (JRip, M5Rules, PART)</td>
<td>Generating rules to summarise the results of simulations (eg to forecast the final competition ladder)</td>
</tr>
<tr>
<td>lm</td>
<td>Create regression models to fit margin of victory</td>
</tr>
<tr>
<td>caret</td>
<td>Fit and assess the performance of a range of regression or classification algorithms using in-built or customised performance measures (eg profit in-use)</td>
</tr>
</tbody>
</table>