AFL – HOW TO WIN

A data driven analysis of winning factors in matches

James Hooi

SPORT IS ABOUT WINNING

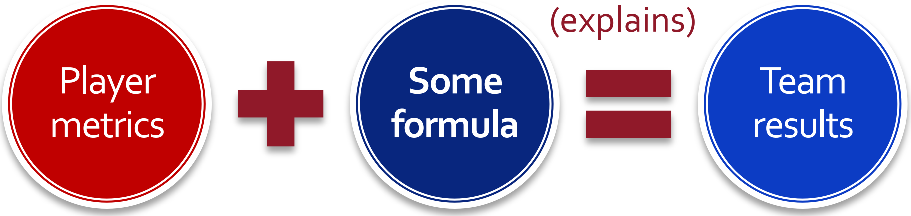
Choosing personnel and tactics in sport used to be more art than science. In some sports, it arguably still is. Nowadays, the rich commercial returns of succeeding in sport through TV rights deals, sponsorships and merchandising is too much to leave to chance – teams need to maximise their chances of "making their own luck".

The story of Moneyball and the Oakland A's signalled the beginnings of a data-driven revolution in sport

Elite sport is about winning and commercial success

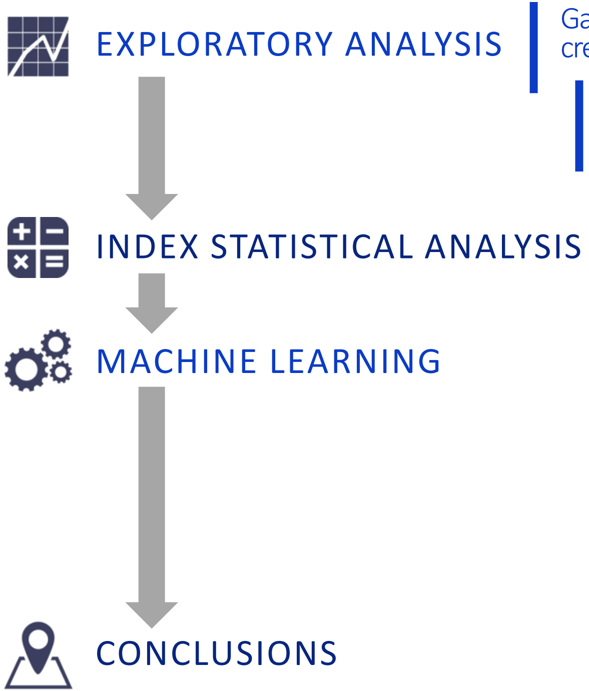
- Commercial success often follows on-field success
- Data analysis is giving teams the competitive edge in elite competition

This analysis sets out to find a link between AFL player performance and team results



ANALYSIS APPROACH

The analysis uses free data from afltables.com and comprises:



EXPLORATORY ANALYSIS

Gathering player metric and team results data, creating team-wise totals for each match

Performing exploratory visualisation to look for relationships in the raw data

INDEX STATISTICAL ANALYSIS

Creating indices to measure team vs opponent, performing visual and statistical analysis

MACHINE LEARNING

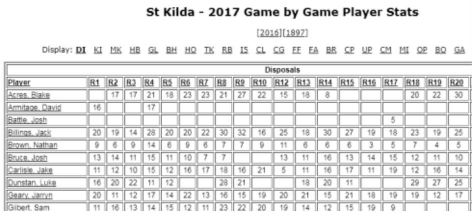
Creating additional proportion measures and clustering data for modelling

Creating Decision Tree models, testing predictive accuracy non-clustered and clustered

Creating Logistic Regression models, testing predictive accuracy non-clustered and clustered

CONCLUSIONS

Documenting findings, opportunities for further analysis and wrap up

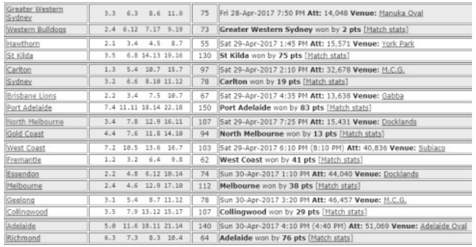


St Kilda - 2017 Game by Game Player Stats

[2016][1897]

Display: **DI** KI MK HB GL BH HO TK RB I5 CL CG FF FA BR CP UP CM MI OP BO GA

	Disposals																	
Player	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R12	R13	R14	R15	R16	R17	R18	R19	R20
Acres, Blake		17	17	21	18	23	23	21	27	22	15	18	8				20	22	30
Armitage, David	16			17															
Battle, Josh																5			
Billings, Jack	20	19	14	28	20	20	22	30	32	16	25	18	30	27	19	18	23	19	25
Brown, Nathan	9	6	9	14	6	9	6	7	7	9	11	6	6	6	3	5	7	4	5
Bruce, Josh	13	14	11	15	11	10	7	7			13	11	16	13	14	15	12	11	10
Carlisle, Jake	11	12	10	15	12	16	17	18	16	21	5	11	16	17	11	19	12	16	14
Dunstan, Luke	16	20	22	11	12			28	21			18	20	11			29	27	25
Geary, Jarryn	20	11	12	17	14	22	13	16	15	19	20	21	15	21	18	19	19	12	17
Gilbert, Sam	11	16	13	14	15	12	11	23	22	20	19	14	12	15	19	9			



Greater Western Sydney	3.3 6.3 8.6 11.9	75	Fri 28-Apr-2017 7:50 PM **Att:** 14,048 **Venue:** Manuka Oval
Western Bulldogs	2.4 6.12 7.17 9.19	73	**Greater Western Sydney** won by **2 pts** [Match stats]
Hawthorn	2.1 3.4 4.5 8.7	55	Sat 29-Apr-2017 1:45 PM **Att:** 15,571 **Venue:** York Park
St Kilda	3.5 6.8 14.13 19.16	130	**St Kilda** won by **75 pts** [Match stats]
Carlton	1.3 5.4 10.7 15.7	97	Sat 29-Apr-2017 2:10 PM **Att:** 32,678 **Venue:** M.C.G.
Sydney	3.2 6.6 8.10 11.12	78	**Carlton** won by **19 pts** [Match stats]
Brisbane Lions	2.2 3.4 7.5 10.7	67	Sat 29-Apr-2017 4:35 PM **Att:** 13,638 **Venue:** Gabba
Port Adelaide	7.4 11.11 18.14 22.18	150	**Port Adelaide** won by **83 pts** [Match stats]
North Melbourne	3.4 7.8 12.9 16.11	107	Sat 29-Apr-2017 7:25 PM **Att:** 15,431 **Venue:** Docklands
Gold Coast	4.4 7.6 11.8 14.10	94	**North Melbourne** won by **13 pts** [Match stats]
West Coast	7.2 10.5 13.6 16.7	103	Sat 29-Apr-2017 6:10 PM (8:10 PM) **Att:** 40,836 **Venue:** Subiaco
Fremantle	1.2 3.2 6.4 9.8	62	**West Coast** won by **41 pts** [Match stats]
Essendon	2.2 4.8 6.12 10.14	74	Sun 30-Apr-2017 1:10 PM **Att:** 44,040 **Venue:** Docklands
Melbourne	2.4 4.6 12.9 17.10	112	**Melbourne** won by **38 pts** [Match stats]
Geelong	3.1 5.4 8.7 11.12	78	Sun 30-Apr-2017 3:20 PM **Att:** 46,457 **Venue:** M.C.G.
Collingwood	3.5 7.9 13.12 15.17	107	**Collingwood** won by **29 pts** [Match stats]
Adelaide	5.0 11.6 18.11 21.14	140	Sun 30-Apr-2017 4:10 PM (4:40 PM) **Att:** 51,069 **Venue:** Adelaide Oval
Richmond	6.3 7.3 8.3 10.4	64	**Adelaide** won by **76 pts** [Match stats]

EXPLORATORY ANALYSIS – 1 OF 2

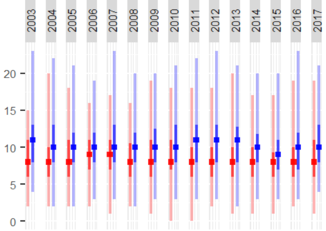
Data visualisation provides a simple and digestible method to get an overview of the data.

Below are plots showing team totals of player metrics in wins vs losses

Illustrated by year for last 15 years

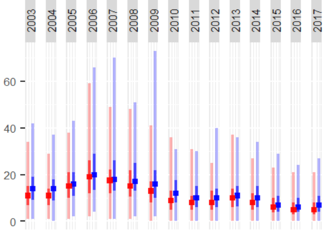
Annotated with statistical confidence interval of win / loss difference being positive – test of whether metric is significant to winning

Behinds



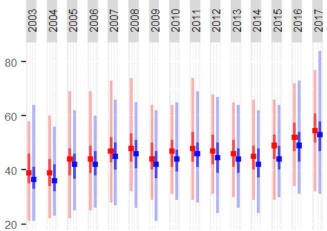
65.9% confident win/loss difference positive

Running Bounces



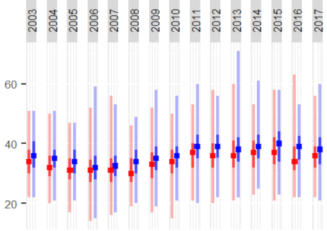
57.1% confident win/loss difference positive

Clangers



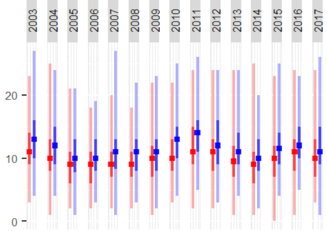
58.7% confident win/loss difference negative

Clearances



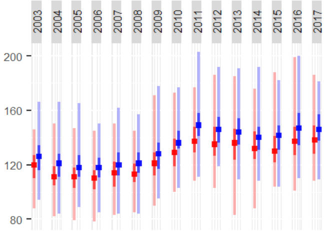
59.9% confident win/loss difference positive

Contested marks



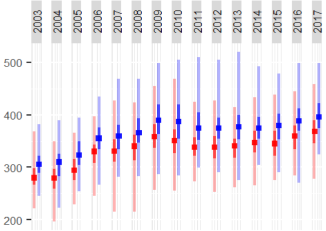
62.2% confident win/loss difference positive

Contested possessions



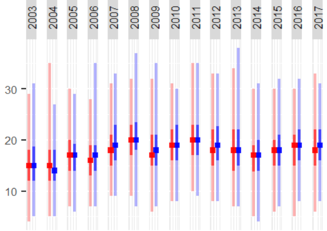
62.9% confident win/loss difference positive

Disposals



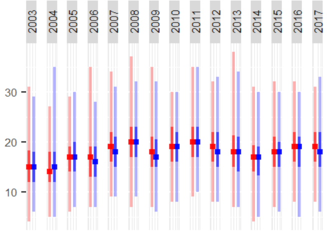
69.9% confident win/loss difference positive

Free kicks awarded



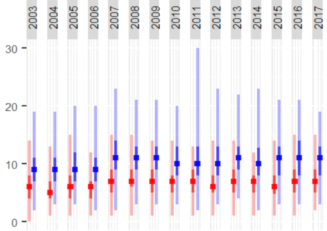
52.4% confident win/loss difference positive

Free kicks given away



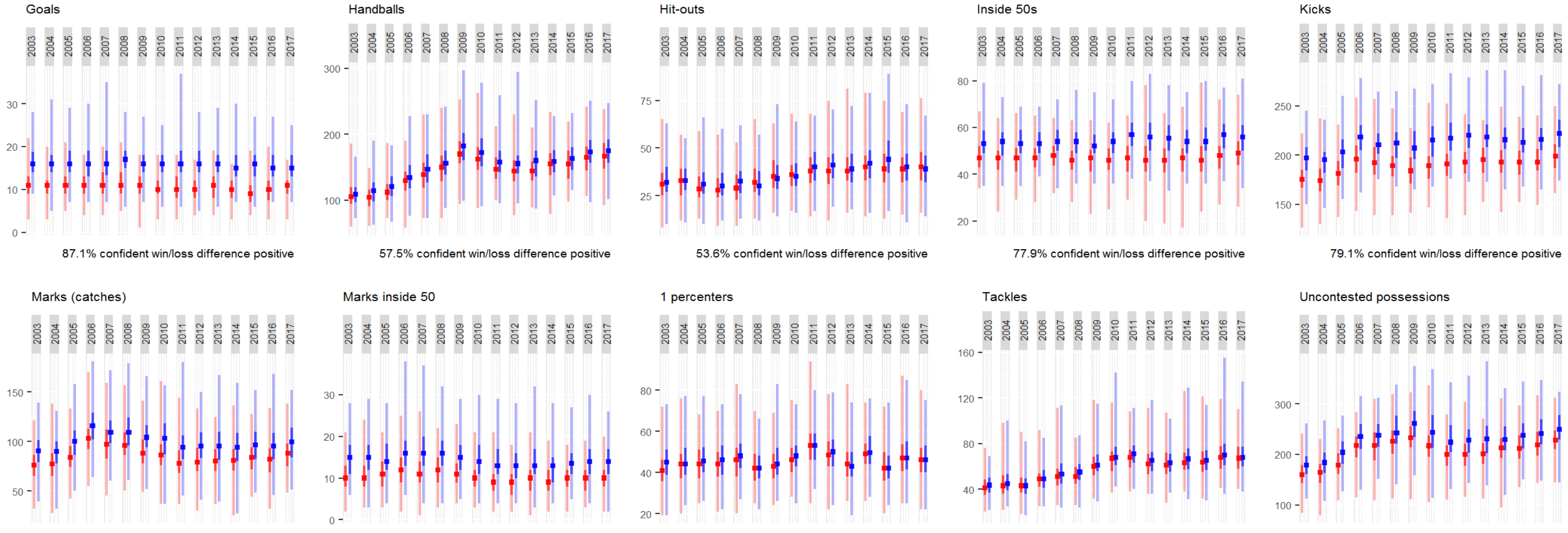
52.0% confident win/loss difference negative

Goal assists



81.3% confident win/loss difference positive

EXPLORATORY ANALYSIS – 2 OF 2



Goals: 87.1% confident win/loss difference positive

Handballs: 57.5% confident win/loss difference positive

Hit-outs: 53.6% confident win/loss difference positive

Inside 50s: 77.9% confident win/loss difference positive

Kicks: 79.1% confident win/loss difference positive

Marks (catches): 68.1% confident win/loss difference positive

Marks inside 50: 76.1% confident win/loss difference positive

1 percenters: 53.2% confident win/loss difference positive

Tackles: 53.6% confident win/loss difference positive

Uncontested possessions: 65.9% confident win/loss difference positive

While some relationships are evident, they aren't useful

The raw data tells us to score, by possessing the ball, from winning it in contests (obvious)

Deeper analysis may uncover more useful insights

INDEX STATISTICAL ANALYSIS – 1 OF 5

In sports analytics, it's critical to remember the game is adversarial – opponents will also influence the result.

Players are not in total control of their performance outcomes due to opponents' actions

Indices can be used to compare a team versus its opponent in key outcome measures (log used to standardise results)

Win contested ground ball

$$idx\_win\_ground\_ball = \log \frac{\text{contested possessions}}{\text{opposition contested possessions}}$$

Win contested aerial ball

$$idx\_win\_aerial\_ball = \log \frac{\text{contested marks}}{\text{opposition contested marks}}$$

Assisted goal

$$idx\_goal\_assist = \log \frac{\text{goal assists/goals}}{\text{opposition goal assists/opp goals}}$$

Maintain possession

$$idx\_mark\_kick = \log \frac{\text{marks/kicks}}{\text{opposition marks/opp kicks}}$$

Accurate 50m entry

$$idx\_50m\_entry = \log \frac{\text{marks inside 50/inside 50s}}{\text{opposition marks i50/opp i50s}}$$

Less clangers

$$idx\_less\_clangers = \log \frac{\text{opposition clangers}}{\text{clangers}}$$

Tackling

$$idx\_tackle = \log \frac{\text{tackles}}{\text{opposition tackles}}$$

Clear ball to advantage

$$idx\_clear\_ball = \log \frac{\text{clearances}}{\text{opposition clearanced}}$$

One percenters

$$idx\_one\_pct = \log \frac{\text{one percenters}}{\text{opposition one percenters}}$$

Give away less frees

$$idx\_less\_frees = \log \frac{\text{opposition frees against}}{\text{frees against}}$$

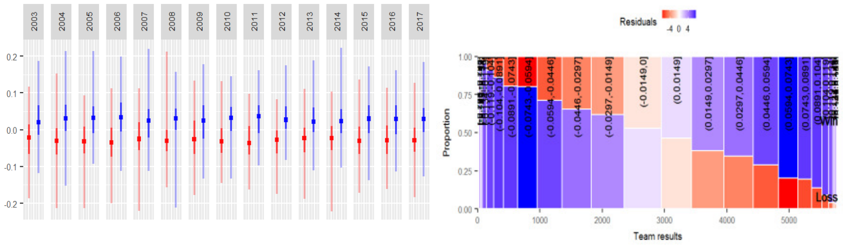
INDEX STATISTICAL ANALYSIS – 2 OF 5

Below are plots showing team indices in wins vs losses

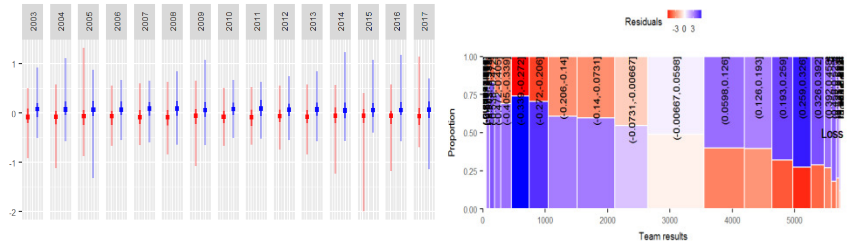
Next to the line plots are mosaics coloured by Chi-squared residuals: blue for over-represented, red for under-represented

Stronger slopes and shades indicate a more significant relationship to wins

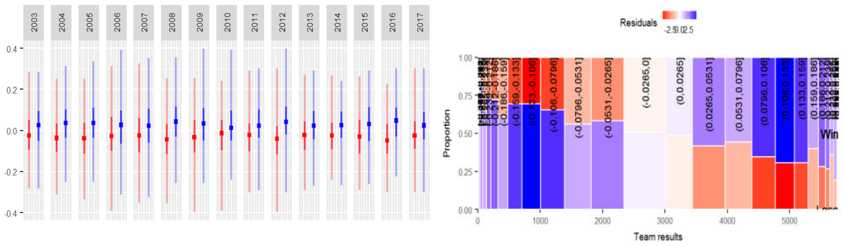
idx_win_ground_ball: Win ground ball



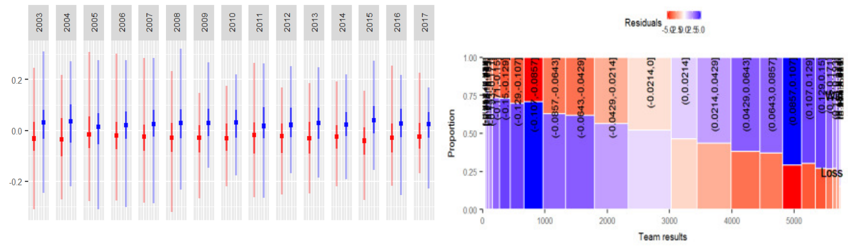
idx_win_aerial_ball: Win aerial ball



idx_clear_ball: Win first use

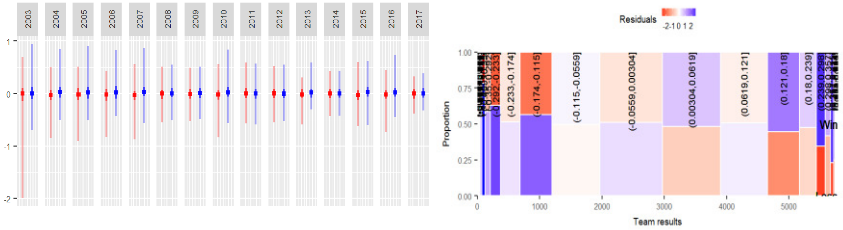


idx_less_clangers: Make less errors

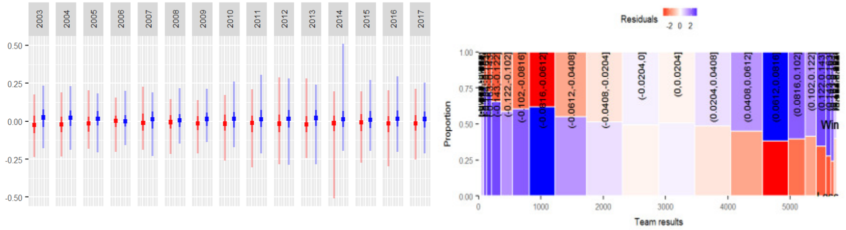


INDEX STATISTICAL ANALYSIS – 3 OF 5

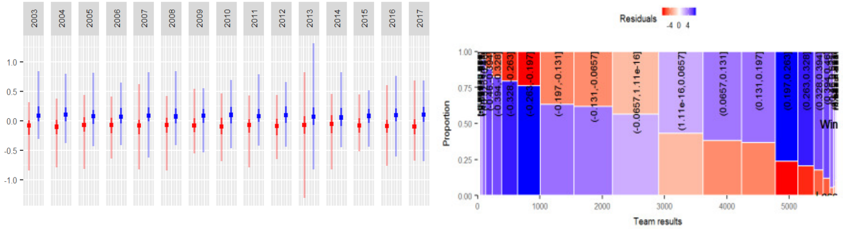
idx_goal_assist: Teamwork for goal



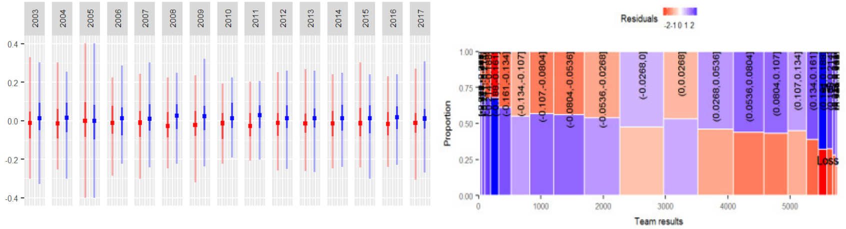
idx_mark_kick: Maintain possession



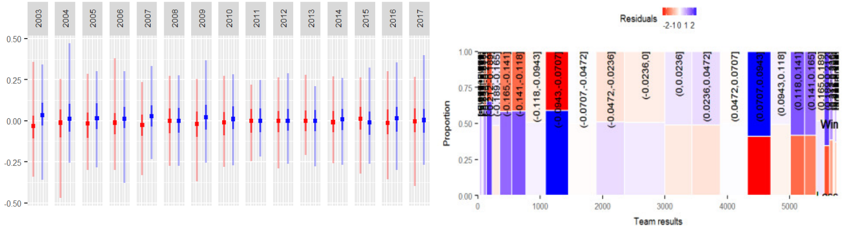
idx_50m_entry: Accurate 50m entry



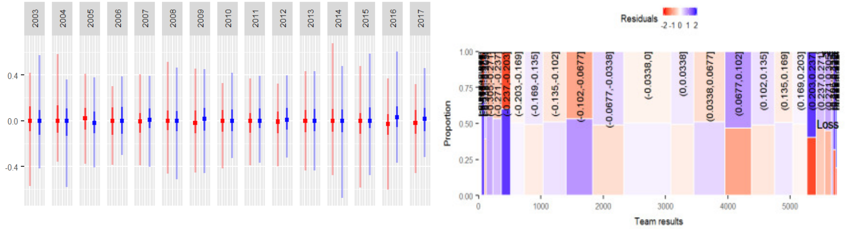
idx_tackle: Win ball back (tackles)



idx_one_pct: Exceptional effort (1 percenters)



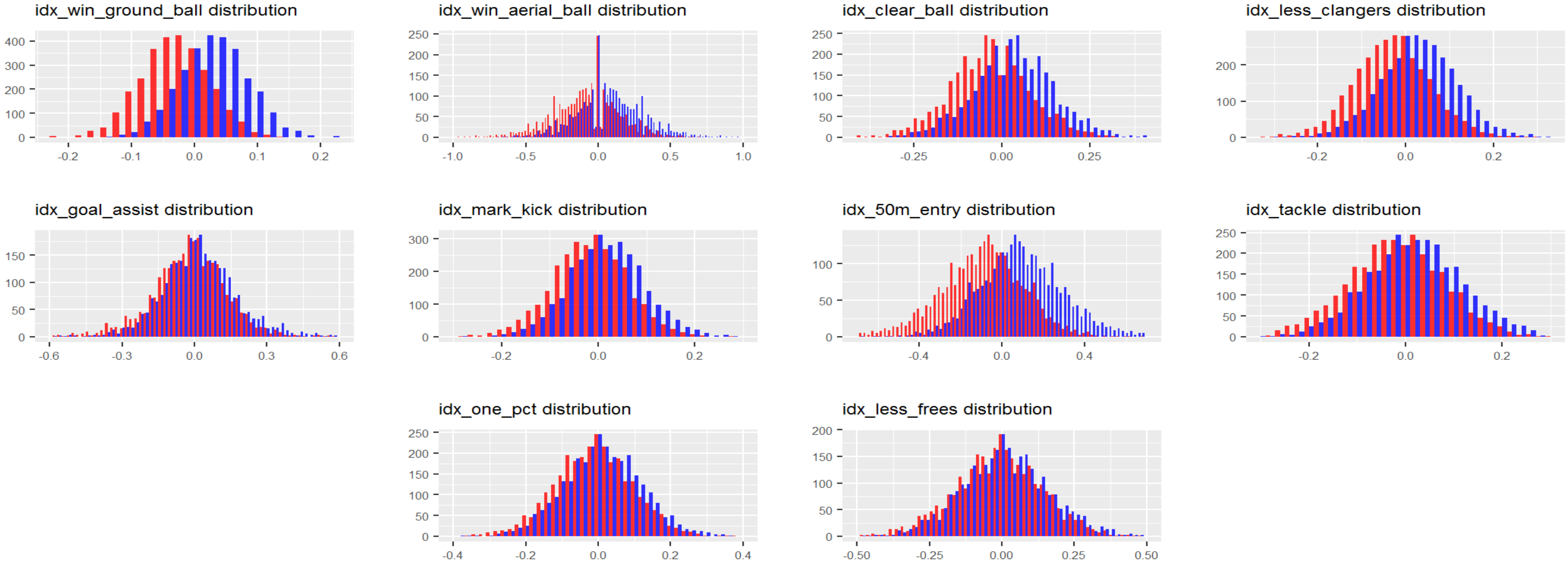
idx_less_frees: Discipline



INDEX STATISTICAL ANALYSIS – 4 OF 5

Distribution histograms of index values by win vs loss help confirm observations from index and mosaic plots

The more distinctly the win curve (blue) differs from the loss curve (red), the more confidently we can visually confirm the relationship of index to wins



INDEX STATISTICAL ANALYSIS – 5 OF 5

Quantifying the results statistically with a significance test helps give them more meaning.

The null hypothesis is an index has no bearing on a win or loss result, so the true mean of the index is zero

If sport is random, a team should win 50% of the time

The calculated z-score from the population mean and standard deviation corresponds to probability of a win if the team scores higher on the index than its opponent

Index	Z-score	Significance test result interpretation
idx_win_ground_ball	0.52	The team leading this index wins approximately **70%** of the time **(Strong influence)**
idx_win_aerial_ball	0.33	The team leading this index wins approximately **63%** of the time **(Fair influence)**
idx_clear_ball	0.27	The team leading this index wins approximately **61%** of the time **(Fair influence)**
idx_less_clangers	0.30	The team leading this index wins approximately **62%** of the time **(Fair influence)**
idx_goal_assist	0.12	The team leading this index wins approximately **55%** of the time **(Weak influence)**
idx_mark_kick	0.19	The team leading this index wins approximately **58%** of the time **(Weak influence)**
idx_50m_entry	0.45	The team leading this index wins approximately **67%** of the time **(Fair influence)**
idx_tackle	0.16	The team leading this index wins approximately **56%** of the time **(Weak influence)**
idx_one_pct	0.10	The team leading this index wins approximately **54%** of the time **(Weak influence)**
idx_less_frees	0.06	The team leading this index wins approximately **52%** of the time **(Weak influence)**

MACHINE LEARNING – 1 OF 5

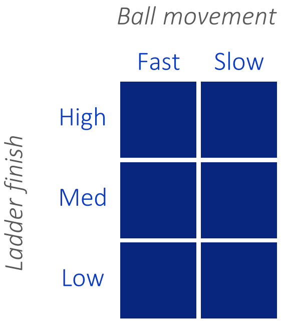
The exploratory and statistical analyses show relationships between indices derived from player metrics and team results. The next step is to model these relationships seeking to predict match results from player performance, and more importantly interpreting the models to gain insights. When modelling, the data used to "train" the model must be quarantined from the data used to "test" the model's accuracy.

Calculated proportion of top ten players belonging to each team for key stats

- E.g. Of players with top ten most inside 50 entries, team A had six (60%)
- Top ten expanded for ties, e.g. if players ranked 10 and 11 for a stat were the same, the proportion was of the top 11

Created six data clusters, thought of as ladder position (H / M / L) and game plan (fast / slow)

- Future analyses could explore different cluster stratification



Ball movement

Fast Slow

Ladder finish

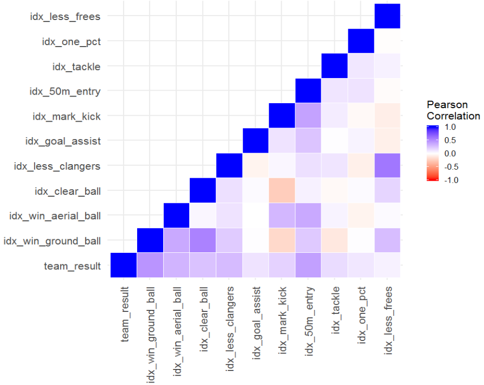
High

Med

Low

Tested for variable impact on each other (multicollinearity), which affects model validity

- Correlations below indicate some care in modelling is needed



MACHINE LEARNING – 2 OF 5

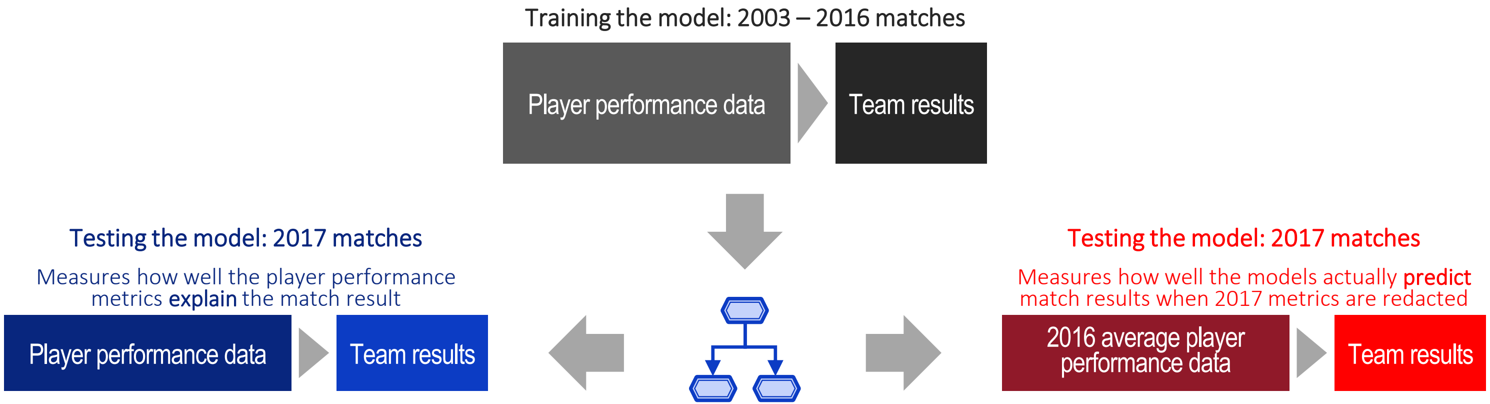
Before detailing the models, a quick point about how the models will be measured:

Data up to 2016 used to train the model and 2017 matches used to test it

- This is the usual set up for testing predictive accuracy of machine learning models
- The models will aim to classify whether a win was achieved or not, therefore accuracy is measured by how many wins and non-wins were correctly predicted

Realistically, 2017 player performance data would not be available so true predictive accuracy will be tested

- 2017 player performance data will itself be predicted by taking the players' 2016 match averages
- These averages will be used to aggregate team totals and indices, then the models applied to predict the match result
- The models will again be measured on classification accuracy



Training the model: 2003 – 2016 matches

Testing the model: 2017 matches

Measures how well the player performance metrics **explain** the match result

Testing the model: 2017 matches

Measures how well the models actually **predict** match results when 2017 metrics are redacted

MACHINE LEARNING – 3 OF 5

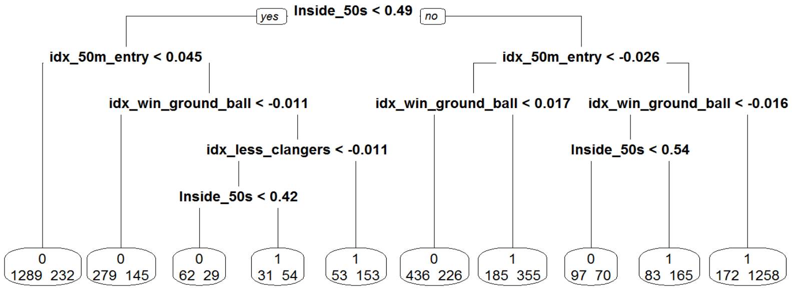
A Decision Tree model iterates through each index to "split" by the most deterministic variable at each point. The tree below is "tuned" to balance over-fitting the model to the training data, predictive accuracy, and interpretability.

Each split checks if a condition is met

- LHS for Yes
- RHS for No

Each "leaf" at the bottom shows the predicted match result and how many losses vs wins were in the training data

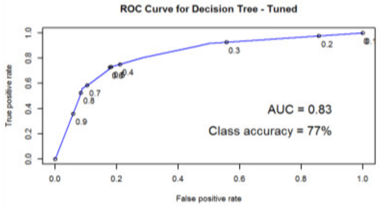
- 0 for predicted loss, 1 for predicted win
- LHS: number of actual losses in leaf
  RHS: number of actual wins in leaf



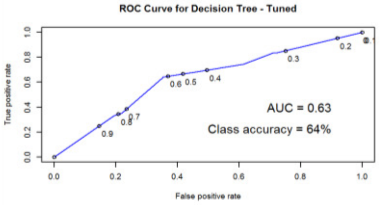
__Classification accuracy__ = proportion of correct predictions against total matches. The benchmark to beat is flipping a coin (50% accuracy).

__AUC__ = Area Under the Curve, a measure of how often a win is predicted for an actual win versus a win is predicted for an actual loss. The benchmark to beat is random (AUC = 0.50).

The Decision Tree explained results well in testing (**316** of **414** matches classified correctly, 2017 player performance available)



However, prediction was less accurate (**267** of **414** matches classified correctly, 2017 player performance estimated)



Using one model per cluster, prediction marginally improved

Cluster	Games	Classification accuracy	AUC
1	11	**100%**	**1.00**
2	17	**65%**	**0.50**
3	43	**67%**	**0.65**
4	133	**65%**	**0.65**
5	146	**64%**	**0.63**
6	64	**66%**	**0.66**

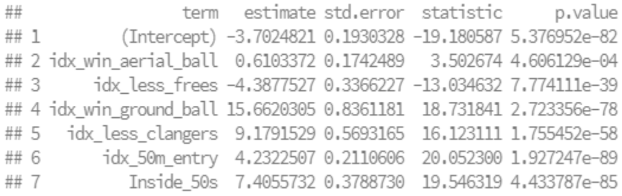
Weighted avg accuracy	65%
Weighted avg AUC	0.65

MACHINE LEARNING – 4 OF 5

A Logistic Regression model seeks a linear relationship between indices and winning. The model below ignores non-significant indices to improve predictive accuracy, and checks for multicollinearity to avoid over-fitting.

Each variable has a coefficient ("estimate" below) which is a measure of influence on winning

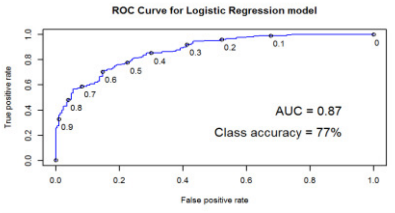
- Note: estimate units are non-standardised so cannot be interpreted verbatim (see next page for variable significance)



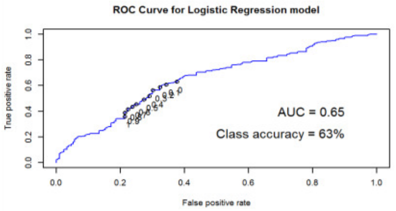
```
##                term   estimate std.error   statistic      p.value
## 1       (Intercept) -3.7024821 0.1930328 -19.180587 5.376952e-82
## 2 idx_win_aerial_ball  0.6103372 0.1742489   3.502674 4.606129e-04
## 3       idx_less_frees -4.3877527 0.3366227 -13.034632 7.774111e-39
## 4 idx_win_ground_ball 15.6620305 0.8361181  18.731841 2.723356e-78
## 5    idx_less_clangers  9.1791529 0.5693165  16.123111 1.755452e-58
## 6        idx_50m_entry  4.2322507 0.2110606  20.052300 1.927247e-89
## 7          Inside_50s  7.4055732 0.3788730  19.546319 4.433787e-85
```

Very small p values indicate the variable is significant to prediction of win/loss result

The Logistic Regression model explained results well in testing (**316** of **414** matches classified correctly, 2017 player performance available) – same accuracy as Decision Tree but better AUC



Prediction was less accurate (**259** of **414** matches classified correctly, 2017 player performance estimated) – less accurate than Decision Tree but better AUC



Using one model per cluster, prediction was unchanged*

Cluster	Games	Classification accuracy	AUC
1	11	**91%**	**0.94**
2	17	**65%**	**0.76**
3	43	**70%**	**0.60**
4	133	**60%**	**0.62**
5	146	**60%**	**0.57**
6	64	**69%**	**0.75**

Weighted avg accuracy	63%
Weighted avg AUC	0.63

***Classification accuracy** = proportion of correct predictions against total matches. The benchmark to beat is flipping a coin (50% accuracy).*

***AUC** = Area Under the Curve, a measure of how often a win is predicted for an actual win versus a win is predicted for an actual loss. The benchmark to beat is random (AUC = 0.50).*

***Note:** * Further tuning is available by choosing threshold value (currently probability > 50%)*

MACHINE LEARNING – 5 OF 5

The relative significance of variables in the Logistic Regression model can be tested by leaving the variable out of the model and measuring the amount of AUC "lost"; i.e. the AUC lift the variable provided.

idx_50m_entry and idx_win_ground_ball both lift AUC strongly

idx_less_clangers and Inside_50s show some AUC lift

idx_less_frees had a negligible impact to AUC, while idx_win_aerial_ball actually dropped AUC, probably due to multicollinearity with inside 50 measures

Variable	AUC lift	Interpretation
idx_50m_entry	**+0.04**	Marking more 50m entries than opponents is very important
idx_win_ground_ball	**+0.03**	Winning more contested ground ball than opponents is very important
idx_less_clangers	**+0.01**	Making less unforced errors than opponents is somewhat important
Inside_50s	**+0.01**	Having a high proportion of top ten players who get the ball inside 50m is somewhat important
idx_less_frees	**±0.00**	Giving away less free kicks than opponents is not really important at all
idx_win_aerial_ball	**–0.01**	*We ignore this finding due to moderate correlation between idx_win_aerial_ball and idx_50m_entry*

The most significant variables in the Logistic Regression model align with those in the Decision Tree – this is expected and confirms their relative importance in this analysis

CONCLUSIONS – 1 OF 2

This analysis has been an insightful proof-of-concept to the value of data to AFL. Beyond the usual sports focus areas of fan engagement, direct marketing and ticket sales, it shows that player performance data can be used as a predictor to match results and more importantly to define what factors are vital to winning.

The analysis confirmed conventional wisdom about how to win AFL matches

- Coaches and talent managers should focus on finding players who are excellent at winning contested ground balls
- They should also seek accuracy in entering the forward 50m arc
- It helps to have more inside 50m contributors generally

Player performance data is a surprisingly strong predictor of match outcomes

- This is surprising due to the adversarial nature of sport where one would expect there are many ways to win
- Decision Trees and Logistic Regression models were both reasonable prediction methods, including when 2017 player data was redacted

Further predictive accuracy could be found… but probably isn't valuable

- Clustering showed promise as a way of boosting accuracy
- Trying different numbers of clusters might find improvement, as could setting individual threshold probability values for each Logistic Regression model
- A more robust method of predicting player metrics for the predicted season (than prior season averages) is probably needed
- However, finding the nth degree of accuracy in prediction may not be valuable – the results must be interpretable to be useful to the football department (coach or talent manager)

CONCLUSIONS – 2 OF 2

Where to from here?

There are almost unlimited opportunities to expand this analysis to keep seeking deeper insights and greater competitive advantage.

Combining multiple, proprietary data sources could lead to huge tactical gains

- This analysis only used free data – much more detail is in Champion Data
- Combining this with player GPS data could lead to invaluable insights into play-by-play outcomes
- What if we could go from understanding we need to win the contested ball, to knowing how to win it with the right running patterns and team plans?

Going back to sports analytics roots: Moneyball version 2

- An understanding of how players can successfully combine to create winning plays could lead to the next revolution in Moneyball
- The relative value of players in the draft and trade period probably has more to do with a team's specific needs than taking the best midfielders and big forwards on offer
- Are we sure we've taken the lessons of Moneyball in AFL?

This approach can be applied to any sport with rich data

- The use of data analysis in short-form cricket (T20) is already defining competitive advantage
- It is evident that teams have specific plans by game phase, by match-up and by player