Predicting Football Match Results for the 2018 FIFA World Cup Could we have “Beaten the Bookies” ?

Gordon Hunter, Benjamin Jauvion*, Ishan Rashid and Mohammed Sharif-Ali

School of Computer Science & Mathematics,

Kingston University, London, U.K.

* Visiting Student from University of Clermond-Ferrand, France, Summer 2018

Kingston University London

Motivation

- Huge interest in professional Association Football (Soccer) around the World,
- Particularly in English Premier League, other European Leagues,
- And, in particular, major international tournaments such as FIFA World Cup
- Interest also in Football-related Gambling and prediction of match/league results & tables :
  - Fans, Gambling Industry, serious gamblers, sport/legal authorities.
  - Gaining competitive advantage OR Detecting match/spot fixing

Kingston University London

Previous Predictions - Some verging on the Bizarre

- Many people have attempted to predict FIFA World Cup match outcomes.
- For reasons listed previously, plus raising National enthusiasm, etc.
- Some of these have had little or no Scientific basis :
  - Astrologers' predictions
  - Animal behaviour, such as "Paul the Octopus" (2013 FIFA World Cup)
- But these probably rely more on luck than real knowledge of factors genuinely influencing the outcomes.
- How many cases of "Not-Paul the Not-Octopus" did we not hear about because their predictions went wrong from the start ?

Previous Work -
Based on Serious Mathematical/Statistical Modelling

- Maher (1982) model – treats “Home Team” score and “Away Team” score as Poisson variables, coupled only by “Attack Strength” & “Defence Strength”, distribution parameters computed iteratively using Maximum Likelihood.
- This type of model was further developed by Dixon & Coles (1997).

- ELO Models & Logistic Regression Models (e.g. Reade & Akie, 2012) : mainly based on teams’ previous form

Kingston University London

In-Play Odds Models

- Bedford & Bagley (2008) modelled in-play odds for USA/Canada professional Ice Hockey results using “phases of play” :
- Modelling the influence of the most recent events (goals, shots, passes, targets, fouls, etc.) on the odds of a given team winning the match :
- Ordinal logistic regression model (win/draw/lose) for each “phase” of the game
- Similar models have also been used for in-play models of tennis match odds (see, e.g. Knottenbelt et al), but the focus of these is often to predict or model very short-term (e.g. next point or next game).

Our In-Play Odds Model for Football (1)

- Can we produce an analogous model for football to those of Bedford & Bagley (2008) ?
- Divide a football match into “chunks” or “phases”
  - For convenience, 9 minutes each, (5+1) phases for each half, 12 for each match
- Record each noteworthy “event” in each phase
  - Goals, Red/Yellow Cards, Penalties, Corners, Shots, Passes, Fouls
  - Note FRACTION of total of each in each phase for each team
- Initial approach : Record matches, then “mark-up” manually
  - Proved far too labour intensive & totally impractical
- Instead, use detailed minute-by-minute updates on www.whoscored.com
- STILL very labour intensive, and difficult to obtain sufficient detailed data for International Matches

Kingston University London

Our In-Play Odds Model (2)

- Log total events of each type & fraction of total for each team for each “phase” of match.

	A	B	C	D	E	F	G
1	Crystal Palace v Chelsea 03.01.2016						
2	10-18mins	C'Palace	Chelsea	Ratio	CP Decimal	Che Decimal	
3	Goals	0	0	1:1	0.5000	0.5000	
4	Total Shots	1	1	1:1	0.5000	0.5000	
5	Shots On Target	1	0	1:0	1.0000	0.0000	
6	Shots Off Target	0	0	1:1	0.5000	0.5000	
7	Blocked Shots	0	1	0:1	0.0000	1.0000	
8	Possession	47	53	47:53	0.4700	0.5300	
9	Touches	44	71	44:71	0.3826	0.6174	
10	Total Passes	29	57	29:57	0.3372	0.6628	
11	Accurate Passes	23	47	23:47	0.3286	0.6714	
12	Key Passes	1	1	1:1	0.5000	0.5000	
13	Dribbles Attempted	3	0	1:0	1.0000	0.0000	
14	Dribbles Won	1	0	1:0	1.0000	0.0000	
15	Aerials Won	1	1	1:1	0.5000	0.5000	
16	Tackles Attempted	0	5	0:1	0.0000	1.0000	
17	Successful Tackles	0	4	0:1	0.0000	1.0000	
18	Clearances	0	1	0:1	0.0000	1.0000	
19	Interceptions	0	1	0:1	0.0000	1.0000	
20	Corners	1	0	1:0	1.0000	0.0000	
21	Fouls	1	2	1:2	0.3333	0.6667	
22	Offsides	0	0	1:1	0.5000	0.5000	
23	Yellow Cards	1	0	1:0	1.0000	0.0000	
24	Red Cards	0	0	1:1	0.5000	0.5000	
25	Substitutions	0	1	0:1	0.0000	1.0000	
26							

0-9 | 10-18 | 19-27 | 28-36 | 37-45 | 45+ | 46-54 | 55-63 | 64-72 | 73-81 | 82-90 | 90+ | ...

Kingston University London

Our In-Play Odds Model (3)

- Produce an ordinal logistic regression model (home win/draw/away win) for each time period ("phase") :
  - Using event statistics for that "phase" alone, and
  - Using "cumulative" event statistics for the "whole match" so far
  - Need to get the correct balance between the individual and cumulative contributions.
  - This will change throughout the match,
    - E.g. effect of a "red card" in first 5 minutes, compared with a "red card" in last 5 minutes
    - E.g. effect of a goal scored in 23rd minute when score was 0-0, compared with a goal scored in 85th minute when score was 5-0.

Kingston University London

Results of Our In-Play Odds Model (for English Premier League)

- Compare our “in-play” odds with those from market (via Odds Portal website)
- All converted to “fair probabilities” (scaled to sum to exactly 1 over the three outcomes).

Crystal Palace - Chelsea

Today, 03 Jan 2016, 13:30

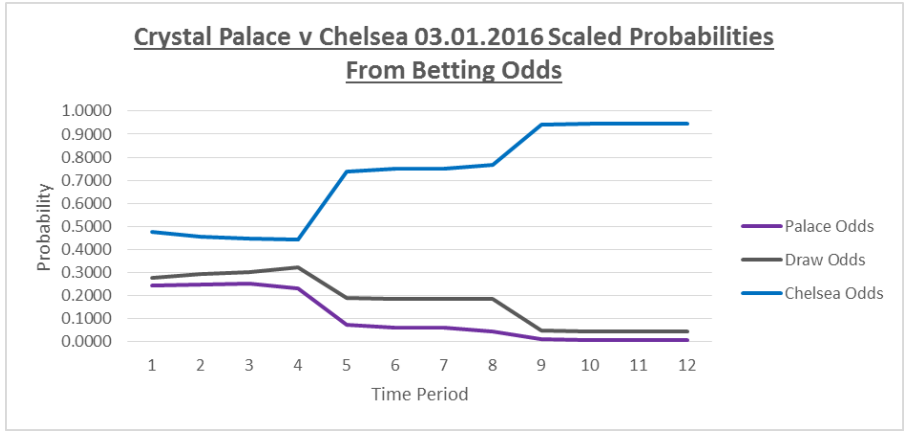
36' 0:1 (0:1)

PRE-MATCH ODDS | IN-PLAY ODDS

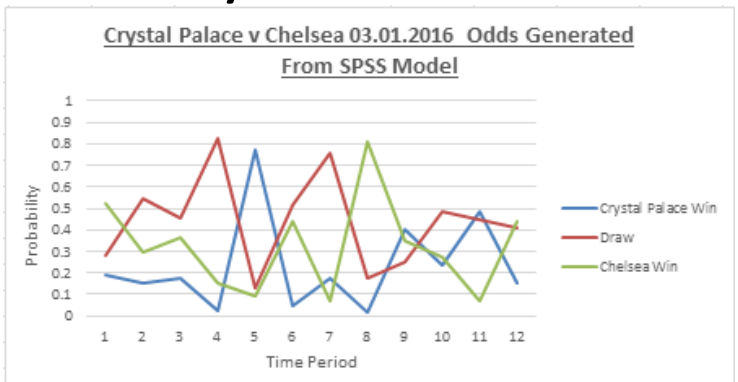
1X2 | AH | O/U | DNB | EH | DC | O/E | BTS

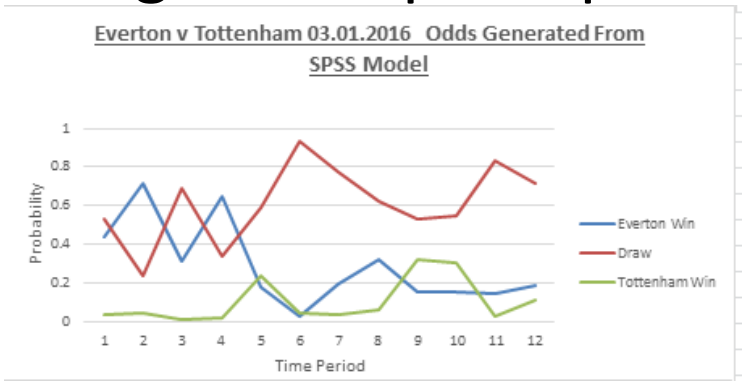
Full Time | 1st Half | 2nd Half

Bookmakers	1	X	2	Payout
10Bet	12/1	15/4	29/100	94.1%
bet-at-home	11/1	18/5	3/10	93.5%
bet365	12/1	4/1	29/100	95.0%
Marathonbet	11/1	77/20	8/25	95.5%
Pinnacle Sports	111/10	203/50	29/100	94.7%
RealDealBet	46/4	39/10	7/25	93.7%
Unibet	10/1	18/5	8/25	93.8%
= Average	1119/100	191/50	3/10	94.4%
↑ Highest	12/1	203/50	8/25	96.9%



- Our predicted odds vary rather erratically over games – perhaps need to be smoothed ?





Kingston University London

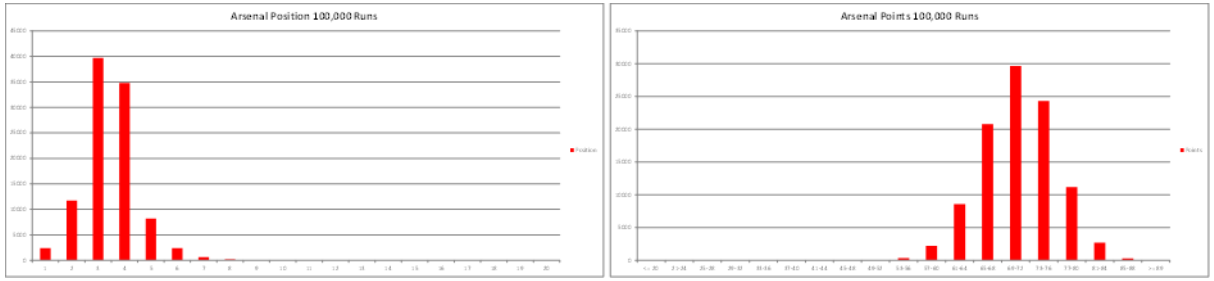
Modelling Matches & League Seasons

- Model English Premier League games
  - Categorise each team (A to E) according to recent previous form
  - Produce ordinal (home win/draw/away win) logistic regression model for Category i (home) versus Category j (away) game, for each i, j ϵ {A, B, C, D, E}
  - Use last season, average of last 4 seasons, or exponentially weighted average of last 4 seasons' data.
  - Compute probabilities of home win, draw, away win for each game of the season.
- Use Monte Carlo simulation approach to simulate whole Premier League Season
- Calculate total points & final position for each team
- Use many (100 000) repetitions to compute distributions for each team's final points & league position.

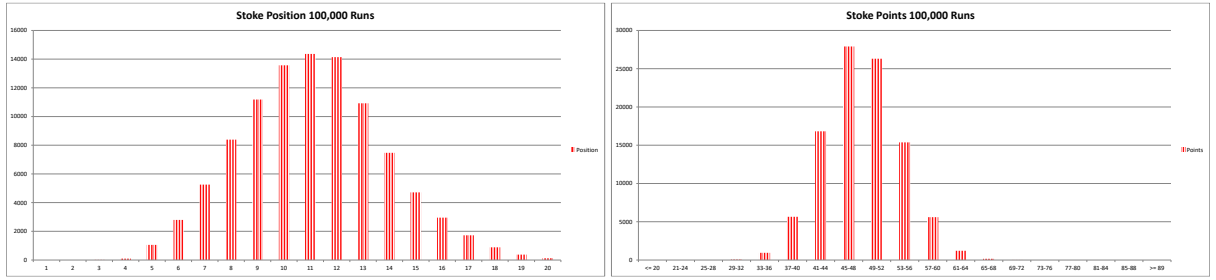
Kingston University London

Results of Modelling Matches & League Seasons (1)

- Distributions of “Big Five” and typical weaker teams were realistic

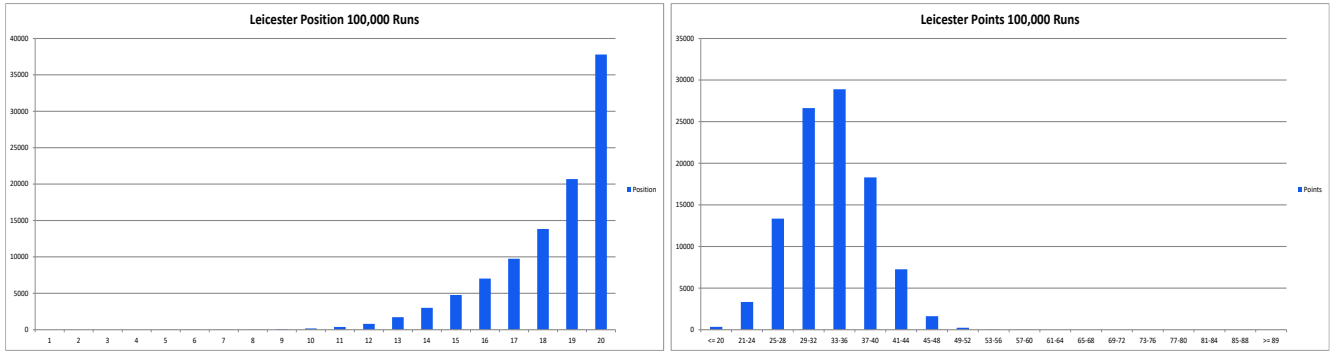


Arsenal Distribution of Position & Points



Stoke City Distribution of Position & Points

- But we couldn’t have predicted that Leicester City FC would win the Premier League !

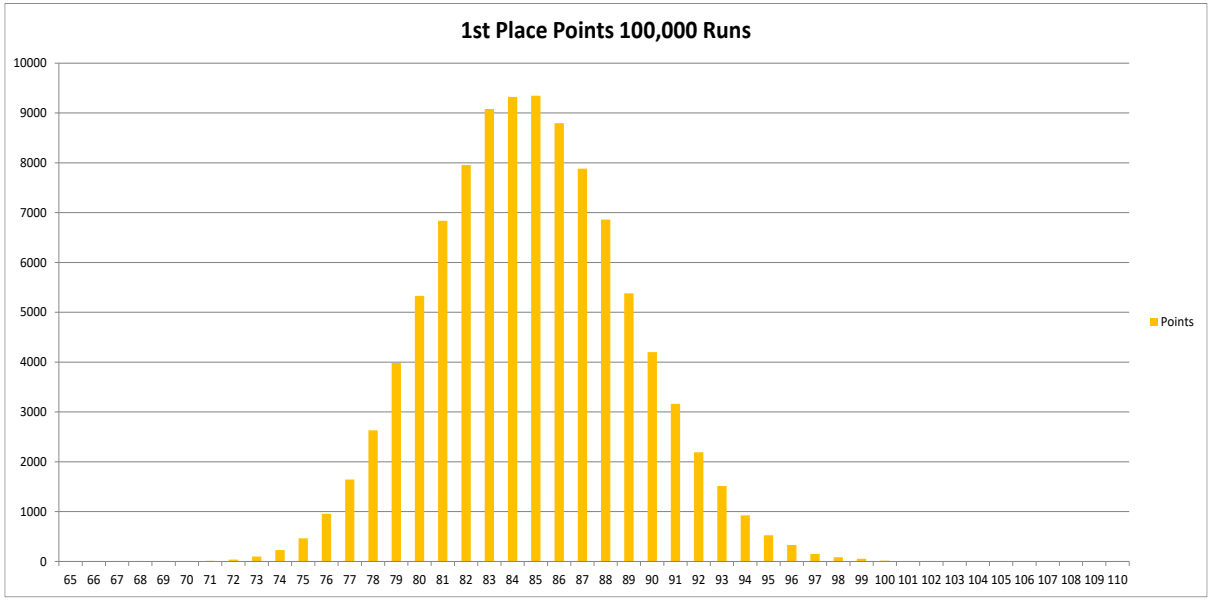


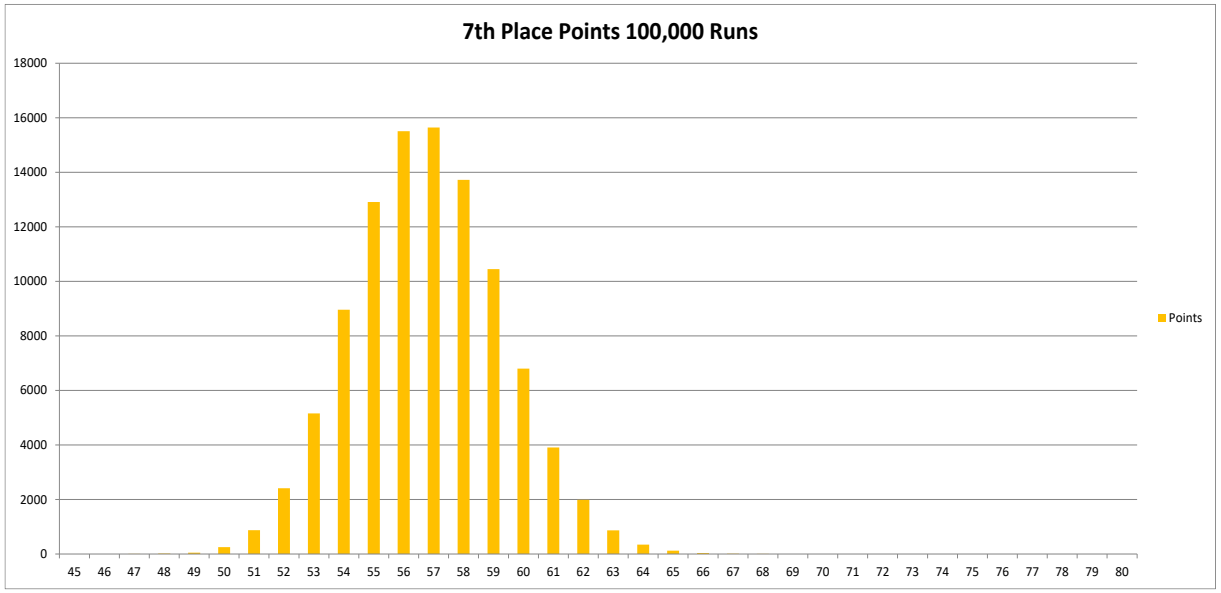
Leicester City Distribution of Position & Points

Kingston University London

Results of Modelling Matches & League Seasons (2)

Distribution of total points over simulated seasons of top placed & 7th placed sides





Kingston University London

Modelling & Predicting International Matches

- Data more “sparse” – fewer matches, particularly previous “head to head” results between specific teams.
- Use factors such as FIFA World Rankings, Recent Goal Difference, Recent Win Fraction, Previous “Head to Head” Win Fraction (where available – otherwise use “default value”).

- What sort of model should be used ?

Regression Modelling for Predicting International Matches

- Use an Ordinal Logistic Regression Model (Win/Draw/Lose).
- Use predictor variable such as FIFA World Rankings, Recent Goal Difference, Recent Win Fraction, Previous “Head to Head” Win Fraction (where available – otherwise use “default value”).
- Sharif-Ali used this to (retrospectively) model the 2014 FIFA World Cup
- Results were not particularly good – but the 2014 World Cup DID contain a high proportion of very surprising results (e.g. the strong performances of Costa Rica).

Modelling the 2018 FIFA World Cup (1)

- This project ran from May to July 2018, so the actual results were coming-in as we were running the models;
- Thus, we were able to model and predict the knock-out stages once the actual Group Stage results were available.
- Again, the actual results included several surprises (e.g. early exit of Germany), so we might have expected the predictions of our models to be rather poor.

- Use several models to forecast probabilities of the outcome of each match.
- Feed these individual match probabilities to simulate the entire tournament 100 000 times via a set of Monte Carlo simulations.
- Data sets used for training models :

  (a) Results from all previous World Cups (since 1930).
  Problems : Not all 2018 teams have been in so many previous tournaments (e.g. Croatia),
  How much do results from long ago (e.g. 1930s) influence outcomes in 2018 ?
  Relatively few "Head to Heads" for most pairs of countries.

  (b) Last 100 International Matches played by each country.
  Problems : Should we re-weight results according to quality of opposition
  (e.g. Team X beat San Marino 1-0, whereas team Y beat Brazil 3-2) ?

Modelling the 2018 FIFA World Cup (2)

Three basic types of models were used for each match :

(i) A Maher-type model to predict the goals scored by Team X and Team Y in the game between X and Y.

(ii) Ordinal Logistic Regression for Win/Draw/Lose
NOTE : This won't allow estimation of Goal Difference or Goals Scored for Group Stage, which could affect final table positions & progression to next stage.

(iii) Ordinal Logistic Regression for number of goals (assumed to be in range $0 \leq$ team_goals ≤ 8) scored by team X when playing "at home" against team Y, then use corresponding Y values for overall match result probabilities.

- Use predictor variable such as FIFA World Rankings, Past Results at World Cups (since 1930, weighted by number of games played) Recent Goal Difference, Recent Win Fraction, Previous "Head to Head" Win Fraction (where available – otherwise use "default value").

Modelling the 2018 FIFA World Cup (3)

- Normalised Maher model : The 32 teams in the 32 FIFA World Cup had previously played each other 970 times in WC tournaments, but not equally distributed between team pairs (e.g. Spain v Germany more common than Iran v Croatia), or even by individual teams.
  - Normalise goals scored relative to a notional 100 matches played, equivalent to average goals scored by team per WC game.
- Adjusting probabilities for 0-0 draws : Maher's "Product of Poissons" model tends to overestimate probability of a 0-0 draw.
- Note : Draws not allowed in "knock out" stages.
- Allow the possibility of introducing a "Confederation Coefficient" to allow adjustment of different difficulties of local qualifying tournaments.
- Or use of a FIFA_ranking coefficient to adjust actual goals scored by Team X against Team Y when there was a big difference between their FIFA rankings.

Probabilistic Models and MC Simulations

- Models for Probabilities for Team A v Team B probabilities were developed using Maximum Likelihood Estimation in SAS, and the resulting probabilities used in a Monte Carlo simulation model developed in C++.
- The whole tournament was simulated 100 000 times using the MC model.
- Progression of teams from group stages to knockout phase, and between knockout stages, followed the FIFA WC rules, neglecting the “Fairer playing side progresses if two teams otherwise tied” final rule.

Results (1) – Using unmodified Maher model

Teams	Round of 1	Quarter	Fourth	Third	Second	Winner
Germany	0,26523	0,14343	0,05415	0,07012	0,07135	0,07355
England	0,24588	0,25489	0,05062	0,06426	0,08318	0,08526
Saudia Ara	0,18268	0,04287	0,01091	0,00416	0,00378	0,00113
Argentina	0,32207	0,19355	0,04123	0,06074	0,05225	0,06813
Australia	0,1839	0,06295	0,01513	0,00772	0,00724	0,00372
Belgium	0,21998	0,13111	0,02874	0,01846	0,01719	0,00899
Brazil	0,28409	0,17148	0,04499	0,07552	0,11345	0,15904
Colombia	0,30316	0,16473	0,04037	0,03533	0,03257	0,02404
Costa Rica	0,28239	0,08959	0,02745	0,01792	0,01677	0,00847
Croatia	0,33428	0,15515	0,03893	0,03896	0,03308	0,02703
Denmark	0,30225	0,15472	0,03988	0,04135	0,03488	0,03066
Egypt	0,32685	0,12252	0,02897	0,02048	0,01934	0,01153
Spain	0,23146	0,17452	0,04268	0,06717	0,06165	0,08742
France	0,29158	0,20886	0,04002	0,07883	0,07838	0,12738
Iceland	0,0622	0,00935	0,00092	0,00022	0,00019	4E-05
Japan	0,2764	0,12693	0,02589	0,01401	0,01354	0,00624
Morocco	0,20515	0,15241	0,03722	0,04241	0,03739	0,03424
Mexico	0,2263	0,10643	0,03653	0,03315	0,03272	0,02471
Nigeria	0,31795	0,13631	0,03244	0,02832	0,02599	0,02067
Panama	0,15736	0,0694	0,01357	0,00637	0,00566	0,00223
Peru	0,18577	0,0681	0,01579	0,00907	0,00757	0,00425
Poland	0,2792	0,12163	0,02696	0,01587	0,01646	0,00876
Portugal	0,21662	0,13507	0,0362	0,03657	0,03274	0,03178
South Kore	0,20807	0,08834	0,03003	0,02399	0,0235	0,01511
Iran	0,16229	0,10374	0,0238	0,01927	0,01713	0,01107
Russia	0,32841	0,11011	0,02698	0,01677	0,01419	0,00817
Senegal	0,26415	0,12967	0,02796	0,01803	0,01793	0,01017
Serbia	0,25101	0,06533	0,02071	0,01014	0,00956	0,00359
Sweden	0,23371	0,10622	0,03878	0,03437	0,03449	0,02572
Switzerlan	0,2492	0,06004	0,01946	0,00878	0,00798	0,00304
Tunisia	0,25387	0,17078	0,03863	0,02884	0,02768	0,01705
Uruguay	0,34654	0,16977	0,04406	0,0528	0,05017	0,05681

- We see that this model suggests that Brazil had the highest *a priori* chance of winning the World Cup (15.9%), but with France (12.7%), Spain (8.7%) and Germany (7.3%) serious contenders.
- However, this simple model also predicts a probability of 32% that Germany would NOT progress beyond the group stage – so perhaps we shouldn't have been so surprised when that did indeed occur !
- Similarly, Spain had a 33% chance, and Argentina a 25% chance, of being eliminated at the group stage according to this model.

Results (2) – Using various other models

Models\Periods	Probability score matches and coefficients	Group results	Final phase results
MaherHTHC	-Scores closer to reality -Differences between teams as the previous model -Probabilities of draws closer to 25%	-Belgium is likely to be eliminated -Germany still has high probabilities of being eliminated -The rest is the same as before	-More luck to win for Brazil (18%) - Less probability for England -The rest is the same as before
Maher100	-Coefficients corresponding better to the form of the teams (better coefficient for Belgium and worse for Uruguay) -Probabilities between tighter teams -High scores	-Favorite teams have trouble leaving groups (example: France 31%) -Belgium has higher probabilities	-Brazil always raw for the win -10% chance for Iran to win the world cup -No teams finishing in the last four of the World Cup have a probability greater than 4% to win in this model
MaherConf	-More differences between "small" and "big" teams.	-Good end results on the groups	-The teams have similar probabilities -Iran still has a good chance of winning
MaherRank	-Large differences between "small" and "big" teams.	-The best teams have the best coefficients -Brazil has a 81% chance of finishing first in its group -Belgium is the team that has the best chance of leaving their group	-Brazil has a 42% chance of winning the World Cup -Argentina and Spain have about 10% to win the world cup, France 6% and Germany 4%.
MaherABK	-Iceland has a worse defense yet	-Sensibly the same results as MaherHTHC	-Sensibly the same results as MaherHTHC
MaherIndep	-Probabilities of fairly realistic scores -Probabilities of draws close to 25%	-Big teams are struggling in the group stage (France 34% chance of being eliminated, Germany 41%, Brazil 26%)	-Brazil with the highest probability of winning (10%) -Other large teams have probabilities between 5 and 8%
LogHTH	-Difference of probabilities between "small" and "large" teams more marked	-Sensibly the same results as MaherHTHC	-Sensibly the same results as MaherHTHC
Log100	-Only model where Portugal has a better probability of victory than Spain -Less difference between the teams	-Uruguay has a 71% chance of finishing first. -Switzerland is as likely to leave the group as Brazil	-Germany is the team most likely to win the World Cup with 10.7% followed by Portugal with 10.4% -The other big teams have between 5 and 8% chance of winning

- Key : “HTH” means “trained on Head to Head” match results only, “100” means “trained on results of team’s last 100 International matches”. “Maher” means based on Maher model, “Log” based on logistic regression, ABK – three independent coefficients in model, ABCD – four independent coefficients in model, C – includes confederation difficulty coefficient. Logistic models build by stepwise inclusion of terms.
- Each model has some good points, but also some quirks or drawbacks.

Results – Betting without Kelly

Models	Threshold	Average gain (or loss) by match (using threshold)	Average gain (or loss) by match (without threshold)
Simple Probability	0,38	0*	-82,5406
MaherConf	0,32	37,49253	-37,6474
MaherConfABK	0,39	31,53048	-35,9751
MaherConfNonIndep	0,38	27,9164	-56,7737
MaherFIFARank	0,48	22,61239	-59,9167
MaherFIFAABK	0,46	14,92991	-59,4549
MaherFIFANonIndep	0,45	8,359727	-69,5249
Maher100	0.31	32,61901	-43,271
Maher100ABK	0,33	25,97849	-40,3388
Maher100NonIndep	0,38	24,37394	-57,9367
MaherHTH	0,43	26,14366	-63,295
MaherHTHABK	0,44	17,0969	-69,5585
MaherHTHNonIndep	0,39	15,13019	-73,5203
Log100	0,42	10,29746	-191,223
Log100Maher	0,38	41,70668	-58,2374
LogHTH	0,53	0,557063	-196,679
LogWDL**	0,44	39,92152	-54,3656

- Virtual bets of £ 1000 x (Decimal Odds) placed on all 738 theoretically possible games.
- “Threshold used” means bet only placed if model calculated probability exceeded empirically determined threshold.
- Thresholded strategy yielded positive net gain for almost every model.
- Unthresholded models always gave a net loss.

- But how can we determine the threshold if we don’t know the actual results in advance ?

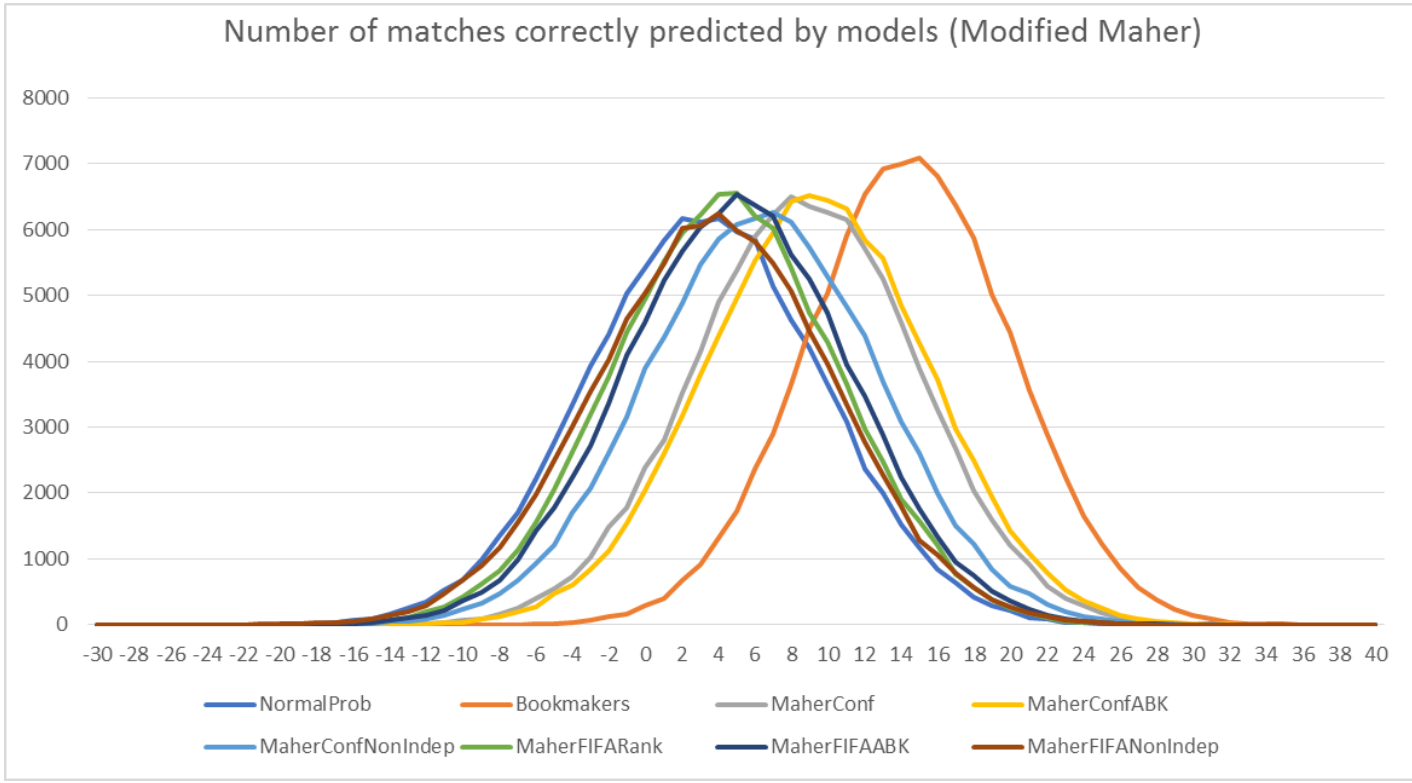
Results – Betting using Kelly

Models	Normal Kelly	Kelly with threshold		Fraction of Kelly	
	Final money	Best threshold	Final money	Best threshold	Final money
NormalProb	1,86E-25	0,36	1000*	0*	1000
MaherConf	1,07E-11	0,6	1986,888	0,1	2069,49
MaherConfABK	7,16E-12	0,6	1916,832	0,1	2092,775
MaherConfNonIndep	3,93E-15	0,35	18037,11	0,01	1002,127
MaherFIFARank	3,89E-26	0,64	1000*	0*	1000
MaherFIFAABK	3,03E-30	0,76	1000*	0*	1000
MaherFIFANonIndep	6,81E-25	0,5	1000*	0*	1000
Maher100	1,29E-12	0,47	7332,05	0,08	1551,107
Maher100ABK	2,45E-13	0,49	2378,202	0,07	1533,498
Maher100NonIndep	1,12E-16	0,35	48288,48	0*	1000
MaherHTH	2,36E-25	0,49	1000*	0*	1000
MaherHTHABK	2,53E-25	0,47	1412,282	0*	1000
MaherHTHNonIndep	7,05E-24	0,39	1000*	0*	1000
Log100	4,13E-17	0,29	537665,7	0*	1000
Log100Maher	1,07E-16	0,32	172383	0*	1000
LogHTH	4,08E-16	0,31	16723,75	0*	1000
LogWDL	1,23E-10	0,39	1542012	0,04	1105,753

- Now try the same virtual betting experiment using various strategies based on Kelly's approach to betting :
- (i) "Standard" Kelly,
- (ii) Kelly, but only betting if prob from model exceeds a threshold,
- (iii) Kelly with threshold, but only an empirically-determined fraction of the Kelly amount is staked.
- A "Best threshold" of 0 tells us that we shold never bet using this money, so we are left with our original £ 1000.
- However, several models give substantial positive returns

Results – Overall Summary

- Overall, the bookmaker's favorite predicted the winners of each game better than any of our models – in terms of the number of results predicted correctly. Thus, we couldn't have made a profit by placing a fixed stake on the favourite to win.
- However, using a carefully-chosen threshold on "when to bet", and a Kelly approach, did lead us to strategy which, for this data, could have given us a net profit.



Discussion & Conclusions

- These models yielded interesting results, but require considerable further work – and would have benefitted from more extensive data !
- Always backing bookie's favourites would have given us the highest number of correctly-predicted match results, **BUT**
- We couldn't have "beaten the bookies" using fixed stakes if we had always backed their favourites.
- However, **ON THIS DATA**, with careful choice of thresholds when deciding to bet, we **COULD** have made a net profit.
- But were we just lucky ? And how could we have known the best thresholds ?
- In-play odds modelling is interesting and shows some promise, but again requires plenty of suitable data.

Kingston University London

Thank You ! Ευχαριστω σας !

Any Questions ?

G.Hunter@Kingston.ac.uk

Kingston University London