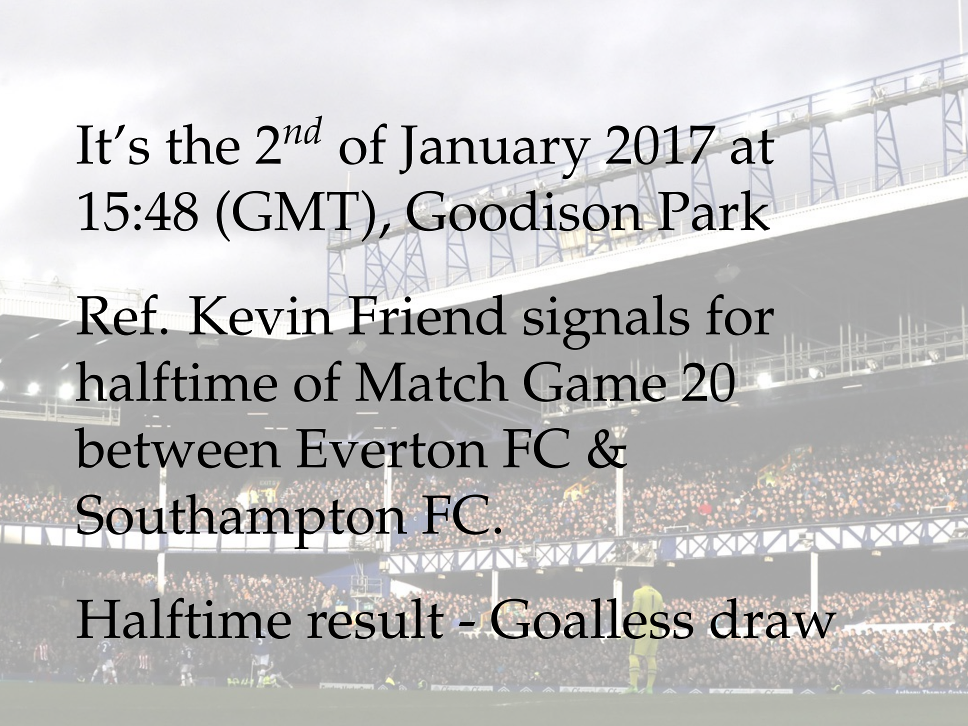
Department of AI, Faculty of ICT, University of Malta

Predictive Analysis of Football Matches using In-play Data

Matthew J. Zammit
matthew.zammit.09@um.edu.mt

September 11, 2018



It's the 2nd of January 2017 at 15:48 (GMT), Goodison Park

Ref. Kevin Friend signals for halftime of Match Game 20 between Everton FC & Southampton FC.

Halftime result - Goalless draw

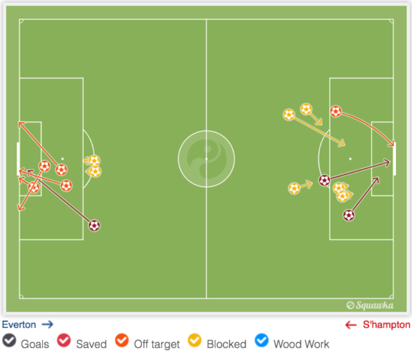


Figure: Shots per team

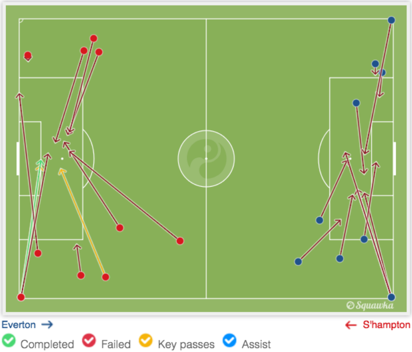


Figure: Crosses per team

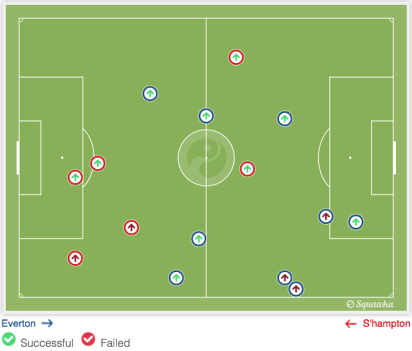


Figure: Dribbles per team

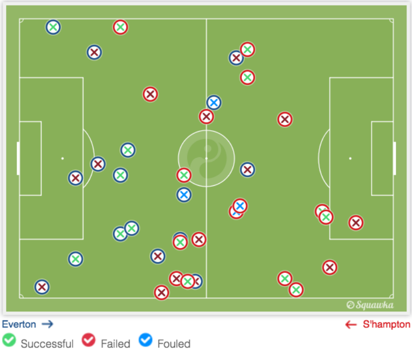


Figure: Tackles per team

Match day - Halftime Analysis (3)

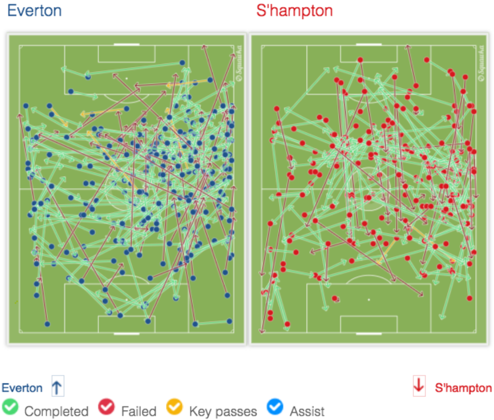


Figure: Passes per team

How do you think is the match going to end?

Match result market at halftime; Everton to win, odds at 2.86 (0.35). S'hampton to win, odds at 4.00 (0.25). Ends in a draw, odds at 2.50 (0.40).

Motivation

- Rise in popularity of betting exchanges through the Internet
- Prediction Markets, have been found to be accurate
- Sports data is recently being captured at precise and granular levels than ever before

Why Machine Learning?

- Multitude of complex variables associated with a football match
- Difficult for humans to think in terms of probability and to react to market changes
- Emotions might hinder the performance of humans to make rational decisions

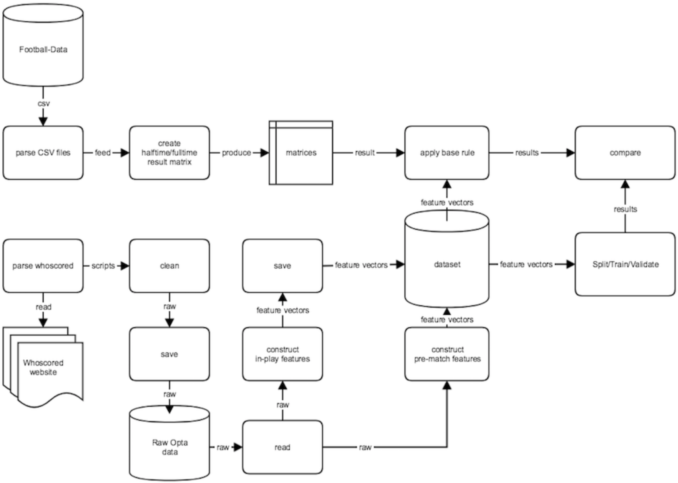
Aims and Objectives

- Predict the fulltime result (H/D/A) of matches drawn at halftime using in-play match data
- Investigate whether using Feature Selection (FS) by a Genetic Algorithm (GA) process would remove certain predictors and increase classification accuracy
- Test if the addition of pre-match data to the in-play game statistics would improve accuracy rate
- Compare the probabilistic classification of the classifier with that of the implied probability from the betting exchange market

Major Issues

- No publicly accessible datasets available
- No previously conducted studies using same data to compare with
- Most similar study was carried out using a Case Based Reasoning approach on the over/under 2.5 goals market

Dissertation Schematic



Halftime/Fulltime Result Base Rule

- Data was retrieved from *Fooball-Data*
- We investigated the transition probability of the halftime result to that of the fulltime result of each match
- Consists of 77,553 match instances spanning several major and minor leagues across Europe over multiple seasons
- We found that for a high percentage of matches the fulltime result remained the same as that of the halftime

Base rule

$$BR_{ftr}(R_{htr}) = R_{htr} \tag{1}$$

In-play Dataset

- *Football-Data* data was not granular enough for the study
- Data was parsed from *Whoscored* website and had to be engineered using conditional rules
- Main benefits why this site was chosen:
  1. Data recorded at a play-by-play rate
  2. Actions are labeled with a type, the x and y coordinates of the ball
  3. Opta as the source
  4. Continually being updated with data of major European competitions. Most importantly, English Premier League, Italian Serie A, Spanish La Liga, French Ligue and the German Bundesliga from 2009/10 to present

Feature Engineering of In-play Predictors

Dataset URL: https://bit.ly/2QdlCs6

- SHOT_TOTAL
- SHOT_ON_GOAL
- ASSIST_SHOT
- ASSIST_INTENT'L
- ASSIST_INTENT'L_GOAL
- PASS_TOTAL/SUCCESS
- PASS_LONG
- PASS_FORW/BACK
- PASS_TRGT_FINAL_TRD
- PASS_TRGT_MID_TRD
- PASS_TRGT_DEF_TRD
- CORNER_FAVOUR
- FOUL_RECEIVED
- CROSS_FAV_TOTAL
- CROSS_FAV_SUCCESS
- OFFSIDE_COMMITTED
- POSSESSION_TOTAL
- POSSESSION_ATT
- POSSESSION_DEF
- INTERCEPTION
- CARD_YELLOW/RED
- TACKLE_TOT/SUCCESS
- DRIBBLE_TOT/SUCCESS

Feature and Target Vector

- The feature vector constitutes of the difference in the halftime statistics between the home and away team
- A positive value for a particular feature means that the home team had accumulated more of that statistic till the halftime than the away team
- The target vector consists of only one element for each feature vector. The value could be from the set {0,1,2}, where the elements represents home, draw and away win respectively

Feature and Target Vector Example

match	target (FTR)	shotTotalDiff	shotOnGoalDiff	passTotalDiff	passLongDiff	passSuccessDiff	passBackwardDiff	passTargetFinalThirdDiff	…	tackleTotalDiff	tackleSuccessDiff	dribbleTotalDiff
Arsenal v Aston Villa	2	2	2	115	-10	123	44	22	…	-9	-3	8
Arsenal v Cardiff	0	5	1	132	-4	138	49	64	…	-9	-5	8
Arsenal v Chelsea	1	-4	-3	89	4	86	46	12	…	-2	-5	1
Arsenal v Crystal Palace	0	4	2	297	-2	302	120	139	…	0	-3	-6
Arsenal v Everton	1	-2	1	-123	-20	-131	-69	-2	…	8	5	-4
Arsenal v Fulham	0	7	3	42	-19	52	12	95	…	-5	-1	3
Arsenal v Hull	0	11	5	232	-20	243	86	123	…	-7	-4	8
Arsenal v Liverpool	0	2	0	118	1	102	58	76	…	12	10	-7
Arsenal v Man City	1	-6	-3	-80	0	-75	-37	-25	…	-3	-1	7
Arsenal v Man Utd	1	6	4	33	1	27	1	52	…	-6	-4	8
Arsenal v Newcastle	0	8	5	130	-1	133	48	50	…	-1	-3	-1
Arsenal v Norwich	0	3	0	22	-8	27	16	23	…	5	4	-3
Arsenal v Southampton	0	-1	-2	7	-3	8	-4	19	…	-3	-4	6
Arsenal v Stoke	0	7	4	179	-9	184	94	45	…	-7	-2	13
Arsenal v Sunderland	0	9	5	280	0	279	131	131	…	-8	-4	9
Arsenal v Swansea	1	7	4	105	-4	110	52	143	…	-8	-7	7
Arsenal v Tottenham	0	3	-2	-27	-14	-40	-18	-1	…	12	5	-8
Arsenal v West Brom	0	4	4	182	15	176	87	55	…	0	-1	-3
Arsenal v West Ham	0	1	2	147	2	146	55	50	…	-6	-4	2

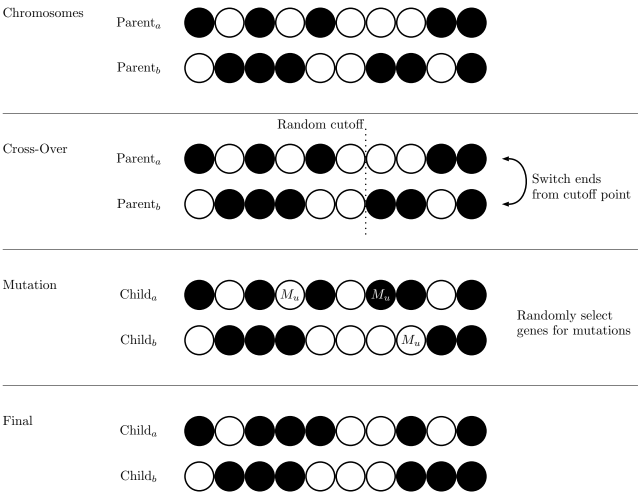
Manual Feature Selection Experiments

- Instances from the English 2015/16 season were used as a sample for initial experimentation
- Features were iteratively being added to the feature space depending on the accuracy and based on our football intuition
- Machine Learning Algorithms used; Neural Nets (NN), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF)
- Dataset was normalized for algorithms which trained faster and perform better with scaled data
- Random Forest was found to be consistently accurate across all the tests

Automated Feature Selection using GA

- Using only the Random Forest algorithm with custom GA separately for each league
- Investigate classification performance and predictors chosen with default parameter settings and with model tuning
  - Nested Cross Validation (CV)
  - Grid search used for parameter tuning
  - Fitness function promotes fewer predictors
  - Growth function for mutation rate
  - Stopping criteria: score of latest generation subtracted by the mean of the scores from the previous ten generations greater than a threshold (0.1)

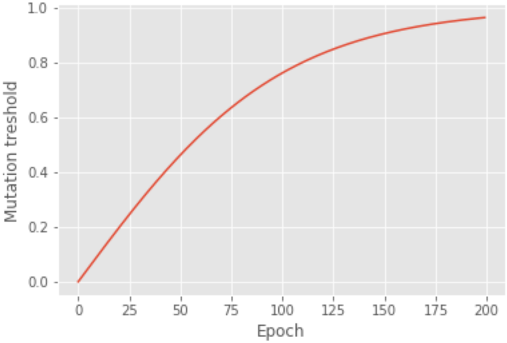
Genetic Algorithm - Feature selection



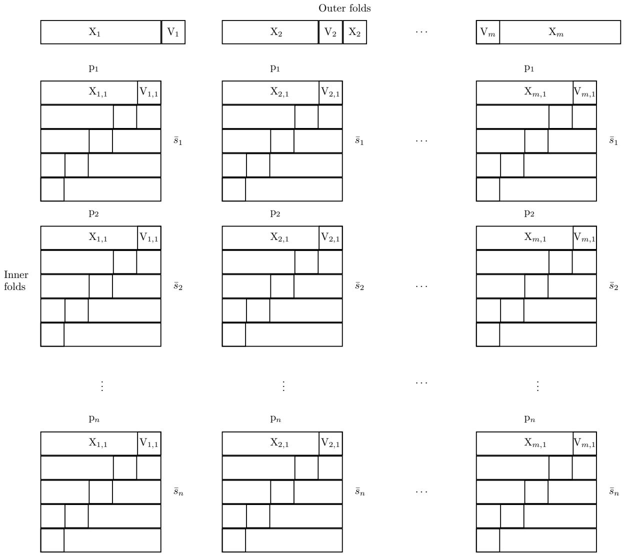
Mutation Threshold

$$mt = \tanh\left(\frac{2i}{n}\right) \qquad (2)$$

Where i is the current epoch and n is the maximum number of epochs.



Nested CV with Parameter Tuning



Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

- TP - True Positive
- TN - True Negative
- FP - False Positive
- FN - False Negative

Pre-match data

- Prematch data added to contextualise the instances
- Prematch feature set includes simple attributes:
  - Goals Scored
  - Goals Conceded
  - Points
- and computed ones
  - Team Form based on the teams' latest performances
  - Attacking Strength
  - Defensive Strength
- Inner partitioning loop customised to train/test on a seasonal basis because of temporal data
- Same as with in-play data, the feature vector consisted of the difference between the home and away team pre-match statistics

Pre-match Features - Att & Def Strength

where, t denotes the team for which the strength is being calculated, n represents the match game and m describes the total number of teams.

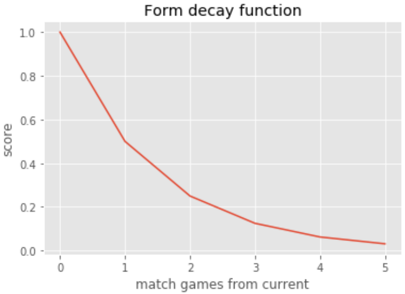
S and C represent the goals scored and goals conceded matrices, respectively.

Attack Strength

$$AttackStr(t, n, m) = \frac{\frac{1}{n}\sum_{i=1}^{n} S_{it}}{\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} S_{ij}} \quad (4)$$

Defence Strength

$$DefenceStr(t, n, m) = \frac{\frac{1}{n}\sum_{i=1}^{n} C_{it}}{\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} C_{ij}} \quad (5)$$

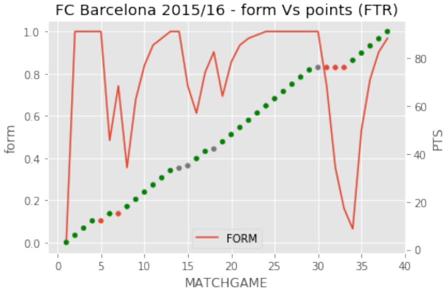


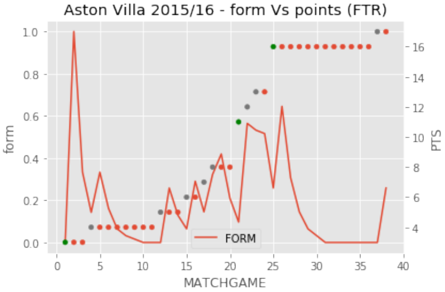
Form

$$Form(t, j) = \frac{\sum_{i=0}^{5} (\frac{1}{2})^i (y_{j-i,t})}{\sum_{i=0}^{5} (\frac{1}{2})^i} \quad (6)$$

Importance is given to the result of the previous games depending on how recent they have been played by assigning them different weights.

Pre-match Form - Examples





Custom Inner Loop for Prematch data

<table>
<tr><th></th><th colspan="5">Split</th></tr>
<tr><th>Data Set</th><th>1</th><th>2</th><th>3</th><th>4</th><th></th></tr>
<tr><td>2011/12</td><td>t_1</td><td rowspan="2">t_2</td><td rowspan="3">t_3</td><td rowspan="4">t_4</td><td></td></tr>
<tr><td>2012/13</td><td>v_1</td><td></td></tr>
<tr><td>2013/14</td><td></td><td>v_2</td><td></td></tr>
<tr><td>2014/15</td><td colspan="2"></td><td>v_3</td><td></td></tr>
<tr><td>2015/16</td><td colspan="3"></td><td>v_4</td><td></td></tr>
<tr><td>Score</td><td>S_1</td><td>S_2</td><td>S_3</td><td>S_4</td><td>$\bar{S}$</td></tr>
</table>

Evalutation of Probability Outcomes

Brier Score Function

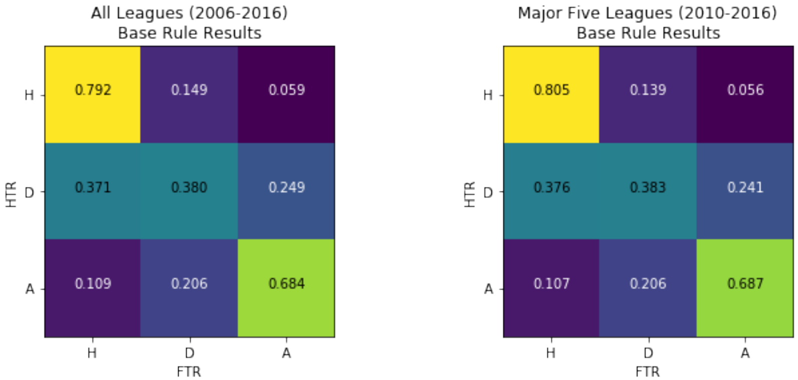
$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}(p_{ij}-o_{ij})^2 \tag{7}$$

Where, n is the total number of instances and r is the number of possible outcomes (three is our case).

$p_{i,j}$ is the probability of the j^{th} outcome for the i^{th} instance from the model. For example, when $i = 1$ the probability vector is [0.7, 0.2, 0.1].

$o_{i,j}$ is the actual probability of the j^{th} outcome for the i^{th} instance after its occurrence. For example, for the same instance $i = 1$, the actual result was [1.0, 0.0, 0.0].

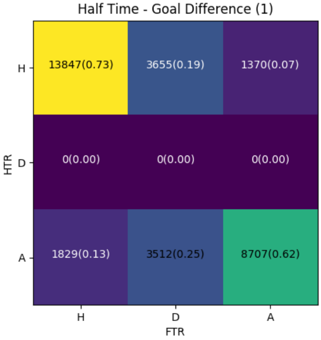
Base Rule Matrices Results

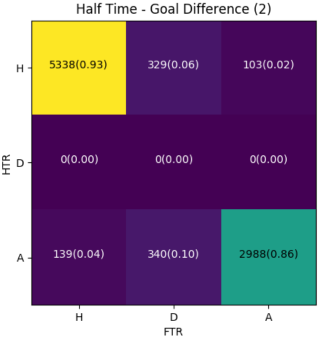


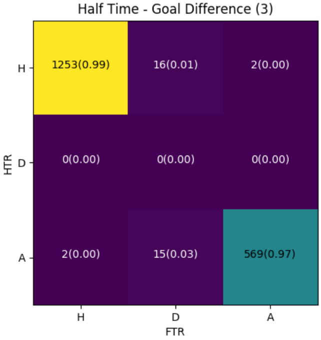
HTR - Halftime Result, FTR - Fulltime Result.

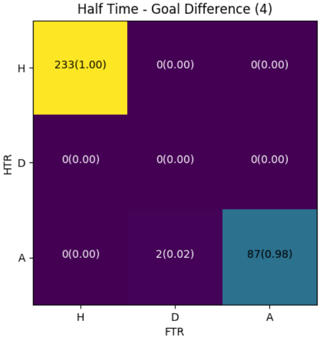
A darker color represents a lower value.

BR Matrices Results by Goal Difference (1)

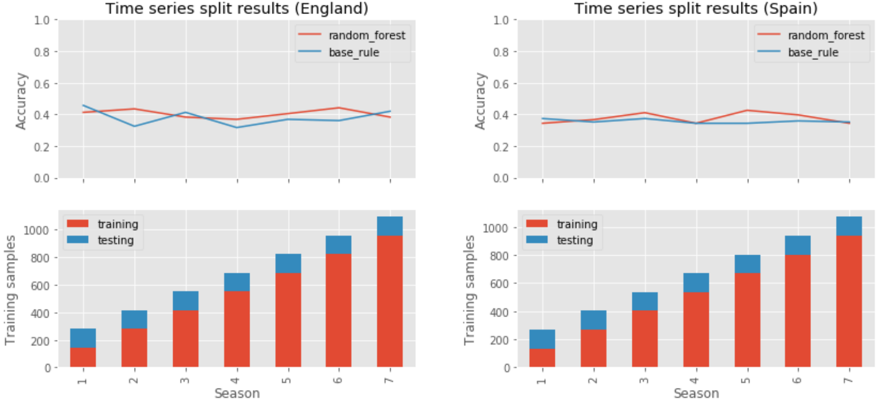




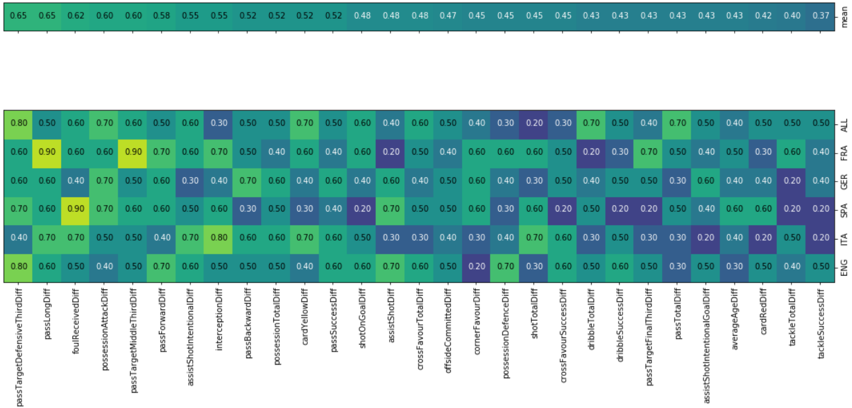

Half Time - Goal Difference (3)
H
D
A
HTR
1253(0.99)
16(0.01)
2(0.00)
0(0.00)
0(0.00)
0(0.00)
2(0.00)
15(0.03)
569(0.97)
H
D
A
FTR


Half Time - Goal Difference (4)
H
D
A
HTR
233(1.00)
0(0.00)
0(0.00)
0(0.00)
0(0.00)
0(0.00)
0(0.00)
2(0.02)
87(0.98)
H
D
A
FTR

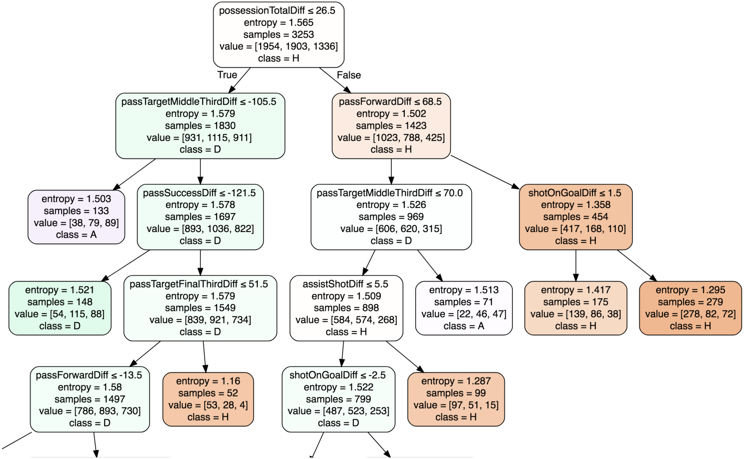
Time Series Results



One tailed paired t-test showed that the accuracy of the time series random forest was not significantly different from that of the base-rule with t-statistic of 1.33 and p-value of 0.19.


0.65 0.65 0.62 0.60 0.60 0.58 0.55 0.55 0.52 0.52 0.52 0.52 0.48 0.48 0.48 0.47 0.45 0.45 0.45 0.45 0.43 0.43 0.43 0.43 0.43 0.43 0.42 0.40 0.37
mean
0.80 0.50 0.60 0.70 0.60 0.50 0.60 0.30 0.50 0.50 0.70 0.50 0.60 0.40 0.60 0.50 0.40 0.30 0.20 0.30 0.70 0.50 0.40 0.70 0.50 0.40 0.50 0.50 0.50
ALL
0.60 0.90 0.60 0.60 0.90 0.70 0.60 0.70 0.50 0.40 0.60 0.40 0.60 0.20 0.50 0.40 0.60 0.60 0.60 0.50 0.20 0.30 0.70 0.50 0.40 0.50 0.30 0.60 0.40
FRA
0.60 0.60 0.40 0.70 0.50 0.60 0.30 0.40 0.70 0.60 0.40 0.60 0.40 0.60 0.40 0.50 0.60 0.40 0.30 0.50 0.40 0.50 0.50 0.30 0.60 0.40 0.40 0.20 0.40
GER
0.70 0.60 0.90 0.70 0.60 0.60 0.50 0.60 0.30 0.50 0.30 0.40 0.20 0.70 0.50 0.50 0.60 0.30 0.60 0.20 0.50 0.20 0.20 0.50 0.40 0.60 0.60 0.20 0.20
SPA
0.40 0.70 0.70 0.50 0.50 0.40 0.70 0.80 0.60 0.60 0.70 0.60 0.50 0.30 0.30 0.40 0.30 0.40 0.70 0.60 0.30 0.50 0.30 0.30 0.20 0.40 0.20 0.50 0.20
ITA
0.80 0.60 0.50 0.40 0.50 0.70 0.60 0.50 0.50 0.50 0.40 0.60 0.60 0.70 0.60 0.50 0.20 0.70 0.30 0.60 0.50 0.60 0.50 0.30 0.50 0.30 0.50 0.40 0.50
ENG
passTargetDefensiveThirdDiff
passLongDiff
foulReceivedDiff
possessionAttackDiff
passTargetMiddleThirdDiff
passForwardDiff
assistShotIntentionalDiff
interceptionDiff
passBackwardDiff
possessionTotalDiff
cardYellowDiff
passSuccessDiff
shotOnGoalDiff
assistShotDiff
crossFavourTotalDiff
offsideCommittedDiff
cornerFavourDiff
possessionDefenceDiff
shotTotalDiff
crossFavourSuccessDiff
dribbleTotalDiff
dribbleSuccessDiff
passTargetFinalThirdDiff
passTotalDiff
assistShotIntentionalGoalDiff
averageAgeDiff
cardRedDiff
tackleTotalDiff
tackleSuccessDiff

Decision Tree Example



Results Summary

Leagues	RF_{HT+PM}	$RF_{GA(T)}$	RF_{GA}	RF_{TS}	RF_{10CV}	BR
English Premier League	**0.482**	0.471 (±.041)	0.440 (±.034)	0.403 (±.028)	0.390	0.379 (±.052)
Italian Serie A	**0.485**	0.442 (±.051)	0.426 (±.039)	0.404 (±.035)	0.372	0.392 (±.018)
Spanish La Liga	0.438	**0.462** (±.038)	0.455 (±.031)	0.375 (±.035)	0.418	0.356 (±.013)
German Bundesliga	**0.449**	0.415 (±.037)	0.433 (±.047)	0.357 (±.043)	0.372	0.346 (±.053)
French Ligue 1	**0.458**	0.438 (±.036)	0.435 (±.040)	0.384 (±.046)	0.392	0.388 (±.037)
Mean	**0.461** (±0.020)	0.450 (±.016)	0.438 (±.011)	0.384 (±0.040)	0.389	0.371 (±0.041)
All leagues	-	**0.434** (±.027)	0.407 (±.015)	-	-	-

- BR - Base Rule
- RF_{10CV} - Random Forest 10-fold Cross Validation
- RF_{TS} - Random Forest Time Series
- RF_{GA} - Default Random Forest with Genetic Algorithm
- $RF_{GA(T)}$ - Random Forest with Genetic Algorithm & Tuned
- RF_{HT+PM} - Random Forest & Genetic Algorithm with In-play & Pre-match Data

Brier Score for the PM was 0.544 and for the RF was 0.623.

Match	Prediction	Home	Draw	Away
AC Milan vs Cagliari	Actual	**1**	0	0
	Random Forest	**0.45**	0.38	0.17
	BetFair	**0.61**	0.29	0.09
Crotone vs Empoli	Actual	**1**	0	0
	Random Forest	0.39	**0.42**	0.19
	BetFair	0.34	**0.41**	0.24
Empoli vs Udinese	Actual	**1**	0	0
	Random Forest	0.26	**0.43**	0.30
	BetFair	0.28	**0.43**	0.29
Lazio vs Chievo	Actual	0	0	**1**
	Random Forest	**0.54**	0.30	0.16
	BetFair	**0.62**	0.30	0.08
Lazio vs Crotone	Actual	**1**	0	0
	Random Forest	**0.60**	0.24	0.16
	BetFair	**0.68**	0.25	0.06
Napoli vs Pescara	Actual	**1**	0	0
	Random Forest	**0.56**	0.28	0.16
	BetFair	**0.75**	0.21	0.04

Match	Prediction	Home	Draw	Away
Napoli vs Pescara	Actual	**1**	0	0
	Random Forest	**0.56**	0.28	0.16
	BetFair	**0.75**	0.21	0.04
Palermo vs Inter	Actual	0	0	**1**
	Random Forest	0.19	**0.41**	0.40
	BetFair	0.12	0.33	**0.56**
Roma vs Cagliari	Actual	**1**	0	0
	Random Forest	**0.48**	0.36	0.16
	BetFair	**0.72**	0.21	0.06
Sampdoria vs Empoli	Actual	0	**1**	0
	Random Forest	0.38	**0.40**	0.22
	BetFair	**0.46**	0.36	0.18
Sassuolo vs Torino	Actual	0	**1**	0
	Random Forest	0.26	**0.39**	0.34
	BetFair	0.23	0.36	**0.40**
Udinese vs AC Milan	Actual	**1**	0	0
	Random Forest	0.32	**0.40**	0.27
	BetFair	0.27	**0.40**	0.33

Comparision with Betting Exchange (Eng)

Brier Score for the PM was 0.622 and for the RF, 0.655.

Match	Prediction	Home	Draw	Away
Arsenal vs Burnley	Actual	**1**	0	0
	Random Forest	**0.70**	0.17	0.13
	BetFair	**0.70**	0.22	0.07
Burnley vs Southampton	Actual	**1**	0	0
	Random Forest	0.14	0.38	**0.48**
	BetFair	0.16	0.41	**0.44**
Hull vs Bournemouth	Actual	**1**	0	0
	Random Forest	0.35	0.26	**0.39**
	BetFair	0.29	**0.38**	0.32
Liverpool vs Swansea	Actual	0	0	**1**
	Random Forest	**0.71**	0.14	0.15
	BetFair	**0.66**	0.28	0.07
Man City vs Burnley	Actual	**1**	0	0
	Random Forest	**0.55**	0.31	0.14
	BetFair	**0.53**	0.33	0.14
Man City vs Tottenham	Actual	0	**1**	0
	Random Forest	**0.52**	0.22	0.26
	BetFair	**0.46**	0.34	0.20
Watford vs Middlesbrough	Actual	0	**1**	0
	Random Forest	**0.40**	0.32	0.27
	BetFair	0.31	**0.46**	0.23
West Ham vs Man Utd	Actual	0	0	**1**
	Random Forest	0.29	**0.36**	0.35
	BetFair	0.06	0.23	**0.71**

Everton won the match by three goals to nil.

Everton vs Southampton	Actual	**1**	0	0
	Random Forest	**0.59**	0.28	0.13
	BetFair	0.35	**0.40**	0.25

Conclusion

- We have derived a base rule for predicting fulltime results at the halftime interval of football matches
- Parsed in-play data from an Opta source and developed a dataset consisting of the differences between team statistics till the halftime for both pre-match and match day data
- Shown that random forest using both types of available data produced the best results
- Similar accuracy as the betting market when considering probabilities for predictions, in some cases out-performing the market and thus giving an edge to the user

Future Work

- Addition of other predictors such as inclusion of key players in team, player individual form and their scoring and defensive abilities
- Rate or number of entrances into opposition penalty box and dangerous areas
- Split match data into several minute time-frames and investigate predictions along the time of play
- Investigate predictions on other markets such as over/under goals and next team to score
- Use predictions as part of a betting strategy

Thank you for your attention