IDS 572- Data Mining for Business
Game of Two Halves

Nikunj Vora[1] Vijendra Singh[2] Thoufeeq Ahamed Kajahussain[3]

December11, 2015

[1] Department of IDS. UIN: 662377890. Email: nvora5@uic.edu
[2] Department of IDS. UIN: 671831191. Email: vsingh31@uic.edu
[3] Department of IDS. UIN: 665731876. Email: tkajah2@uic.edu

Contents

Introduction

The English Premier League, also famously known as “EPL” or “Premiership” is the most followed professional football league in the world. Every year (season), 20 teams contest for winning the league and three teams are put into the relegation (moving down to lower leagues) zone. EPL’s high UEFA coefficient rank allows four teams every season to participate among elite teams from different leagues. With the increasing competition

Problem Statement

Objective:

We will build a model that predicts the outcome of a game based on the half time statistics of that game and historical full time statistics where statistics of a game will include variables like first half goals, shots taken, shots on target, fouls committed, etc.

Importance and Potential Implications:

Before the advent of the Data Mining techniques, sports organizations mostly depended on human experience that came from scouts, coaches, managers, players, for predicting the outcome of the game, as they converted the historical records into useful knowledge. Also, for example, pre-game analysis of a soccer game often include expert predictions but they are not always accurate as they are based on subjective claims and teams stature. Now, as the data size grew human experience was no longer reliable for prediction and organizations started looking for more methods. Data mining techniques can contribute for a better performance by leveraging historical game statistics.
Such Data Mining techniques are usually employed by,

- Betting companies
- Sports organizations
- Sport teams for evaluation of their game strategy

It is essential that we make progress in understanding the art of predicting soccer matches because it is a problem that many people really care about and this prediction can be used for the bookmakers making them rich!

Soccer Game Result prediction is very popular among fans around the world, which can be held responsible for making soccer betting famous. Since the stakes here are so high, prediction of game results had to be more than just a gut feeling of a soccer pundit, it had to be quantified which led to everyday development in soccer data mining techniques.

In this paper we take English Premier League for predictions because not only it is renowned worldwide for its fan base, but also for the top quality football played. In the past four seasons, three different teams have claimed the title with an almost negligible margin of difference. The purpose of our model building is to study the match data and look out for patterns which can be used in prediction of outcomes based on different match scenarios. With different teams having a different probability assigned to them after each match, it becomes interesting to understand what factors affect the probability of a team winning and predict the outcome.

Literature Review

There have been many papers dealing with the topic of predicting the final outcome in various sports. Sports like soccer, basketball have the majority of the work that has been done in the sports data mining field. Most of the papers use techniques varying from data mining to statistical models.

In a paper we reviewed as a part of our literature work had an implementation of naive Bayes and multinomial linear regression models in a combination, predicting the outcome of basketball matches. They used 141 variables divided into 2 groups for the prediction of the game results. They treated this as a classification problem with their system achieving the accuracy of 67% which means they achieved two-third correct predictions.[3]

Logistic Regression is one of the common techniques that is used in predictions of sports results having only 2 possible outcomes. But in one of the paper we studied a generalized logistic regression model was used to predict the World Cup 2014 group stage games. Historical World Cup data was used for the prediction.[6]

In an extensive study of predicting outcomes for premier league matches for 2011 season, a simulation model was built taking into consideration the betting odds and the team statistics. Based on the predicted outcomes, profit over time was calculated for the bets that were placed.[5]

In another paper that predicted basketball results, artificial neural networks were used along with regression analysis. Accuracy of the model was defined by taking the ratio of the correct predictions to total number of prediction made. It was new method of model evaluation we saw in out literature study.[4]

One of the problems we saw in all the papers we studied was the lack of historical data or the richness of the data available for the training of the models. This problem was shown in one of the papers where predictions were made for the college football games. Since no player plays more than 4 year of college football the features were limited and building statistical model was difficult.[2]

After reviewing the work done on the sports data mining we tried to overcome the limitations that were placed during their study and build a better and more robust model than the work which was already done.

Data Source

The dataset is collected from the English Premier League (EPL) matches from 2011-2014. The data set contains 1520 observations of 23 variables. The dataset was taken from http://www.football-data.co.uk/

Data Description

In football matches, the impact of being a home team vs being away team is significant and thus all the data points used for predicting are measured in terms of home team or away team.

The target variable is FTR = Full Time Result. It is a categorical variable with following three levels.

H	Home Win
D	Draw
A	Away Win

The predictor variables are as follows:

Variable	Description
FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HTR	Half Time Result (Home Win,Draw, Away Win)
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HHW	Home Team Hit Woodwork
AHW	Away Team Hit Woodwork
HC	Home Team Corners
AC	Away Team Corners
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HO	Home Team Offsides
AO	Away Team Offsides
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards
HBP	Home Team Bookings Points (10 Yellow, 25 Red)
ABP	Away Team Bookings Points (10 Yellow, 25 Red)

Data Harmonization

Derived Variables: All the below mentioned variables were calculated using Excel macros.

Variable	Description
HomeTeamMatchesPlayed	Home Team Matches played during the season
HomeTeamMatchesPlayed	Away Team Matches played during the season
HomeTeamPointsScored	Home Team Points Scored in this season
HomeTeamPointsScored	Away Team Points Scored in this season
HomeTeamHomeWin%	Percentage of home team winning home matches
AwayTeamAwayWin%	Percentage of away team winning away matches

The following two variables were calculated in the SPSS

PointsofDifference: Using HomeTeamPointsScored and AwayTeamPointsScored we calculated the difference of the points which indicated how the home team rated as compared to away team.

SeasonForm: It is the ratio of HomeTeamHomeWin% and AwayTeamAwayWin%. SeasonForm gives a understanding of the ratings of the two team for a particular match. A Home team with greater % Home win will have a large season form for the match and if away team has large %Away win then the Season form will be a small number indicating that away team is a stronger one.

Descriptive Statistics

Variable	Count	Mean	Min	Max	Range	Variance	Standard Deviation	Standard Error of Mean
HomeTeamHomeWin%	1518	0.381	0	1	1	0.073	0.27	0.007
Away TeamAwayWin%	1520	0.264	0	1	1	0.049	0.222	0.006
HTHG	1520	0.695	`	5	5	0.728	0.853	0.022
HTAG	1520	0.524	0	4	4	0.529	0.727	0.019
HS	1520	14.55	2	43	41	29.773	5.456	0.14
AS	1520	11.41	0	30	30	22.654	4.76	0.122
HST	1520	6.478	0	24	24	12.861	3.586	0.092
AST	1520	5.082	0	20	20	9.41	3.068	0.079
HF	1520	10.51	2	23	21	10.866	3.296	0.085
AF	1520	10.86	1	24	23	12.409	3.523	0.09
HC	1520	6.237	0	19	19	10.11	3.18	0.082
AC	1520	4.763	0	19	19	7.607	2.758	0.071
HY	1520	1.447	0	7	7	1.402	1.184	0.03
AY	1520	1.803	0	8	8	1.657	1.287	0.033
HR	1520	0.061	0	2	2	0.06	0.245	0.006
AR	1520	0.097	0	2	2	0.097	0.311	0.008
SeasonForm	1518	1.452	0	13.286	13.286	2.651	1.628	0.042
PointDifference	1520	-0.014	-3	3	6	0.592	0.769	0.02

Algorithms, Evaluation and Measurements

1. <u>LOGISTIC REGRESSION</u>

We have a classification problem and we need to predict a categorical variable (Full Time Result FTR) with 3 possible outcomes (H, A and D). Next we have 17 continuous variables and 3 categorical variables which serve as input. Since, we had vast numeric data we started with Logistic Regression. We used the following variables for logistic regression- HomeTeam, AwayTeam, HTHG, HTAG, HTR, HS, AS, HST, AST, HF, AF, HC, AC, HY, AY, HR, AR, PointDifference, HomeTeamHomeWin%, Away Team AwayWin%. For the model options we used multinomial procedure with stepwise regression and set base category for target as ‘D’. We got the following parameter estimates for the target FTR

Evaluation:

Results for output field FTR

Comparing $L-FTR with FTR

'Partition'	1_Training		2_Testing	
Correct	713	70.38%	355	70.02%
Wrong	300	29.62%	152	29.98%
Total	1,013		507	

Coincidence Matrix for $L-FTR (rows show actuals)

'Partition' = 1_Training	A	D	H
A	235	43	38
D	71	96	78
H	30	40	382
'Partition' = 2_Testing	**A**	**D**	**H**
A	112	14	18
D	31	45	51
H	13	25	198

And we were able to get the above accuracy for the logistic regression model

2. NEURAL NETWORK

Since our data exhibited seasonal variability and derivation of new fields such as PointDifference which gave the difference between points scored by home team and away team during the current season. So we used neural networks to gather more inferences on the data. We also used AutoDataPrep to transform continuous variables based on a Max/Min Transformation method to values in the range of {0, 1}. But the results were far disappointing since neural networks didn't do well with larger number of numerical inputs.

Evaluation:

Results for output field FTR

Comparing $N-FTR with FTR

'Partition'	1_Training		2_Testing	
Correct	683	67.42%	320	63.12%
Wrong	330	32.58%	187	36.88%
Total	1,013		507	

3. K Means Clustering

We used K means clustering to build clusters for three output we were predicting. We set the cluster size to 3 and used selected variables to generate the below cluster.

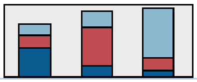
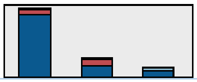
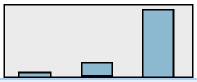
Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	cluster-1	cluster-3	cluster-2
Label			
Description			
Size	41.8% (635)	34.3% (521)	23.9% (364)
Inputs	HTR D (100.0%)	HTR H (100.0%)	HTR A (100.0%)
	HTHG 0.31	HTHG 1.56	HTHG 0.14
	HTAG 0.31	HTAG 0.14	HTAG 1.45
	PointDifference -0.06	PointDifference 0.22	PointDifference -0.27
	AST 4.87	AST 4.54	AST 6.23
	HST 6.03	HST 7.36	HST 5.99
	SeasonForm 1.32	SeasonForm 1.82	SeasonForm 1.14
	HomeTeamHome Win%	HomeTeamHome Win%	HomeTeamHome Win%
	HY 1.47	HY 1.29	HY 1.63
	AS 11.26	AS 11.00	AS 12.27
	Away Team AwayWin%	Away Team AwayWin%	Away Team AwayWin%
	HR 0.05	HR 0.05	HR 0.10
	AR 0.09	AR 0.13	AR 0.07
	HS 14.18	HS 15.14	HS 14.39
	HC 6.37	HC 5.93	HC 6.43
	AF 11.09	AF 10.67	AF 10.73
	AY 1.80	AY 1.74	AY 1.90
	AC 4.87	AC 4.77	AC 4.57
	HF 10.51	HF 10.55	HF 10.49

After segmenting into three clusters we connected it to a type node and set the partition variable. Then we connected it to a logistic regression model and our prediction got increased by 1% on testing set.

Graph	Model	$KM-K-Means	No. Records in Split	No. Fields Used	Overall Accuracy (%)
		cluster-1	635	11	63.150
		cluster-2	364	7	74.725
		cluster-3	521	5	81.574

Evaluation:

Results for output field FTR

Comparing $L-FTR with FTR

'Partition'	1_Training		2_Testing	
Correct	738	72.85%	360	71.01%
Wrong	275	27.15%	147	28.99%
Total	1,013		507	

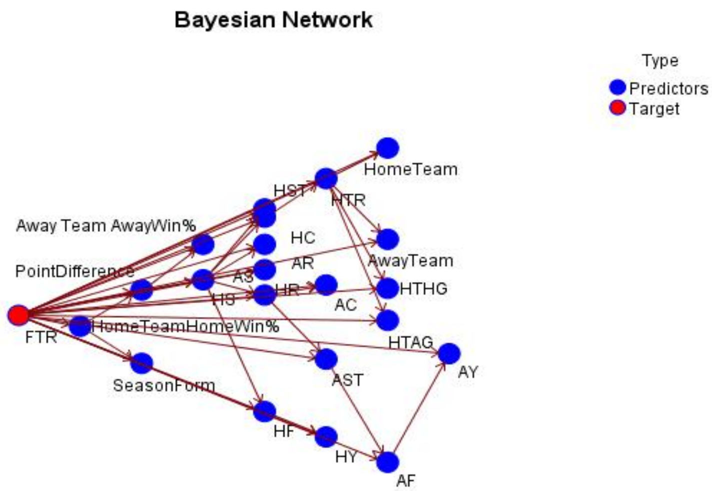
Coincidence Matrix for $L-FTR (rows show actuals)

'Partition' = 1_Training	A	D	H
A	239	42	35
D	58	112	75
H	33	32	387

'Partition' = 2_Testing	A	D	H
A	105	23	16
D	25	57	45
H	17	21	198

4. **<u>Bayesian Network Model</u>**

We used Bayesian network to find the better accuracy by using the following network.



<u>Evaluation:</u>

Results for output field FTR

Comparing $B-FTR with FTR

'Partition'	1_Training		2_Testing	
Correct	771	76.11%	371	73.18%
Wrong	242	23.89%	136	26.82%
Total	1,013		507	

Coincidence Matrix for $B-FTR (rows show actuals)

'Partition' = 1_Training	A	D	H	$null$
A	248	40	27	1
D	48	144	52	1
H	24	49	379	0

'Partition' = 2_Testing	A	D	H
A	112	19	13
D	26	65	36
H	16	26	194

5. **<u>Decision Tree Models</u>**

We also used different tree models like CHAID, CRT and C5 to predict the outcome. But decisions tree were not performing well on our dataset due to wide seasonality of data. For decision trees we got the best prediction accuracy for CRT tree with the following result.

Evaluation:

Results for output field FTR

Comparing $R-FTR with FTR

'Partition'	1_Training		2_Testing	
Correct	669	66.04%	313	61.74%
Wrong	344	33.96%	194	38.26%
Total	1,013		507	

Coincidence Matrix for $R-FTR (rows show actuals)

'Partition' = 1_Training	A	D	H
A	183	80	53
D	51	104	90
H	16	54	382

'Partition' = 2_Testing	A	D	H
A	77	41	26
D	21	44	62
H	16	28	192

Difficulties & Limitations

The difficult part was to makes meaningful interpretation of the data. The original data had limited capabilities and thus harmonization or generating variables which would represent the target variable in a better way was important.

Limitations:

- Team Statistics:
  The data does not include the variables stating the team's historical performances. This is important since the historical trends of the team can help us better understand the team's position as compared to other team.
- Player Statistics:
  With every new season, players are transferred from one team to other and thus the overall team performance is dependent on the players. So, having details about the team's players and their injuries, form, availability can be of utmost importance in deducing the team's performance and thus the result.
- Result Uncertainty:
  Football is a game and thus there is always uncertainty of the final results. Thus, achieving a very high accuracy is difficult task. But we can try to predict the final results and the variables used for predicting can help in building strategy for the team.

Observations & Results

The algorithm used for predicting the full time result based on half time statistics and other criteria can be summarized as follows:

Model	Training Accuracy	Testing Accuracy
Logistic Regression	70.38%	70.02%
Neural Network	67.42%	63.12%
K-Means Clustering	72.85%	71.01%
Bayesian Network Model	76.11%	73.18%
Decision Tree Model	66.04%	61.74%

From the table, we can conclude that Naïve Bayes model gives the best accuracy. But as we know that Naïve Bayes model assumes that the variables used for predicting are independent of each other which is not the case in our dataset. Thus, we can conclude that Logistic Regression is the best model for predicting the output. Also, the accuracy for logistic model increases with the increase in training data and so we can expect accuracy increase when the historical data increases and thus can serve as a best model for predicting the output.

Using the best model, we predicted the outcomes of the matches played in a Game week. Following table describes the prediction,

Match	Predicted Outcome	Probability	Actual Outcome
1	H	0.839	H
2	H	0.93	H
3	H	0.999	H
4	H	0.864	H
5	A	0.417	D
6	A	0.731	A
7	D	0.358	H
8	H	0.897	H
9	A	0.493	A
10	D	0.536	H

As we can see, the model predicts 7 out of 10 observations accurately.

Conclusion

This paper presents models for predicting the outcomes of English Premier League matches. The model predicts the outcome of the match based on home team. We used 5 different models for the problem and compared them based on accuracy in predicting the outcome. During the process of building the model we identified few features that affect the outcome of the game and that analysis can be used by different soccer managing authorities for building their strategy for the game. The results that were achieved were satisfactory and according to the final predictions.

This technique can be used for future implementation in other domains of sports and few extra features can be added like player transfers, manager quotient, etc. in order to make the predictions more accurate and realistic.

References

[1] Data Source: http://www.football-data.co.uk
[2] Carson K. Leung*, Kyle W. Joseph, "Sports data mining: predicting results for the college football games", 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES 2014
[3] Dragan Miljkovic, Ljubiša Gajic, Aleksandar Kovacevic, Zora Konjovic, "The Use of Data Mining for Basketball Matches Outcomes Prediction", 2010 IEEE 8th International Symposium on Intelligent Systems and Informatics, Sept 10-11,2010.
[4] Chenjie Cao, "Sports Data Mining Technology Used in Basketball Outcome Prediction", Dublin Institute of Technology , 2012-01-01
[5] Je_rey Alan Logan Snyder, "What Actually Wins Soccer Matches: Prediction of the 2011-2012 Premier League for Fun and Profit", May 6, 2013
[6] Virgile Galle and Ludovica Rizzo, "Analytics Edge Project: Predicting the outcome of the 2014 World Cup", June 27, 2014