Predicting Soccer Match Results

Siyu Chen

Background

> Dataset of match statistics of *Bundesliga*, the top soccer league in Germany, in season 15/16.

> 18 teams, every team play against each other twice. (306 games per season)

Purposes

> To predict the match result based on home and away teams recent statistics.

> To help a team improve its performance.

Original Data (.csv)

Date	HomeTean	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
14/08/15	Bayern Mu	Hamburg	5	0	H	1	0	H	23	5	9	1	10	12	7	0	2	2	0	
15/08/15	Augsburg	Hertha	0	1	A	0	0	D	20	11	3	4	20	22	7	4	1	2	1	
15/08/15	Darmstadt	Hannover	2	2	D	1	0	H	11	14	4	5	21	22	5	9	1	2	0	
15/08/15	Dortmund	M'gladbacl	4	0	H	3	0	H	17	5	7	1	13	14	3	5	0	1	0	
15/08/15	Leverkuser	Hoffenhein	2	1	H	1	1	D	25	6	9	2	12	18	13	5	1	0	0	
15/08/15	Mainz	Ingolstadt	0	1	A	0	0	D	9	14	3	5	20	15	6	2	2	3	0	
15/08/15	Werder Bre	Schalke 04	0	3	A	0	1	A	15	16	2	5	7	16	5	6	0	2	0	
16/08/15	Stuttgart	FC Koln	1	3	A	0	0	D	28	9	8	5	13	8	13	3	3	2	0	
16/08/15	Wolfsburg	Ein Frankfu	2	1	H	2	1	H	5	13	3	6	14	19	7	3	1	1	0	
21/08/15	Hertha	Werder Bre	1	1	D	1	1	D	10	9	2	3	11	16	6	2	0	1	0	
22/08/15	Ein Frankfu	Augsburg	1	1	D	0	1	A	11	14	3	4	17	21	7	4	2	3	0	
22/08/15	FC Koln	Wolfsburg	1	1	D	1	0	H	9	15	6	7	15	11	2	4	2	1	0	
22/08/15	Hamburg	Stuttgart	3	2	H	1	2	A	14	12	5	4	18	22	0	5	3	3	0	
22/08/15	Hannover	Leverkuser	0	1	A	0	1	A	6	11	3	4	15	21	3	8	0	0	0	
22/08/15	Hoffenhein	Bayern Mu	1	2	A	1	1	D	2	5	1	2	7	2	1	2	0	0	0	
22/08/15	Schalke 04	Darmstadt	1	1	D	0	1	A	21	9	5	3	16	18	6	6	1	2	0	
23/08/15	Ingolstadt	Dortmund	0	4	A	0	0	D	3	18	2	8	18	16	1	4	2	3	0	
23/08/15	M'gladbacl	Mainz	1	2	A	0	1	A	15	13	6	4	12	12	3	0	2	1	0	

Bb1X2	BbMxH	BbAvH	BbMxD	BbAvD	BbMxA	BbAvA	BbOU	BbMx>2.5	BbAv>2.5	BbMx<2.5	BbAv<2.5	BbAH	BbAHh	BbMxAHH	BbAvAHH	BbMxAHA	BbAvAHA	PSCH	PSCD	PSCA
43	1.15	1.08	13	10.6	40	29.37	40	1.4	1.35	3.5	3.1	30	-2.5	1.91	1.84	2.06	1.99	1.1	12.57	30.92
43	2.1	2	3.5	3.34	4.2	3.9	41	2.33	2.23	1.68	1.63	27	-0.5	2.04	1.99	1.91	1.86	2.13	3.43	3.87
43	2.65	2.49	3.4	3.25	3.02	2.87	41	2.22	2.13	1.76	1.7	27	0	1.88	1.8	2.14	2.05	2.7	3.28	2.9
42	1.8	1.71	4	3.78	5.21	4.83	41	1.92	1.82	2.05	1.96	25	-0.5	1.75	1.71	2.27	2.19	1.63	4.24	5.85
43	1.57	1.49	4.8	4.42	7	6.24	36	1.61	1.55	2.56	2.38	27	-1	1.88	1.82	2.09	2.03	1.48	4.84	7.28
43	2.1	2	3.6	3.41	4.08	3.8	42	2.06	1.98	1.89	1.81	27	-0.5	2.05	1.99	1.92	1.86	1.97	3.63	4.2
43	3.3	3.01	3.62	3.42	2.4	2.31	41	1.78	1.72	2.2	2.09	26	0.25	1.9	1.86	2.06	2	3.22	3.74	2.26
43	2.15	2.03	3.6	3.38	3.96	3.7	41	2.12	2.04	1.85	1.76	27	-0.5	2.09	2.03	1.88	1.83	2.07	3.67	3.77
43	1.5	1.46	5	4.62	7	6.54	39	1.5	1.46	2.8	2.63	26	-1.25	2.06	1.98	1.93	1.88	1.44	5.27	7.29
38	2.15	2.07	3.6	3.41	3.86	3.6	37	2.13	2.03	1.82	1.77	25	-0.5	2.14	2.06	1.85	1.8	2.1	3.54	3.82
40	2.25	2.18	3.65	3.43	3.7	3.26	38	1.88	1.78	2.1	2	26	-0.25	1.93	1.88	2.03	1.96	2.3	3.52	3.33
40	4.4	4.13	3.9	3.63	1.91	1.86	38	1.8	1.72	2.2	2.09	27	0.5	2.06	1.99	1.9	1.86	3.45	3.67	2.19
40	3.01	2.85	3.6	3.4	2.6	2.43	38	1.82	1.74	2.17	2.07	26	0	2.17	2.07	1.82	1.76	2.87	3.58	2.55
40	4.65	4.35	4	3.68	1.9	1.8	38	1.84	1.76	2.11	2.03	26	0.5	2.14	2.05	1.85	1.79	4.65	3.85	1.83
40	11	9.77	6.5	5.73	1.33	1.29	38	1.45	1.41	3	2.82	26	1.5	2.15	2.06	1.84	1.79	14.58	7.11	1.23
40	1.45	1.42	5	4.59	9	7.54	37	1.69	1.64	2.34	2.22	26	-1	1.74	1.69	2.28	2.2	1.5	4.59	7.5
40	8	7.08	5	4.38	1.5	1.46	38	1.77	1.7	2.2	2.12	26	1.25	1.86	1.82	2.11	2.03	7.15	4.54	1.52

> Eliminated the betting features

Dictionary for Remaining features

~~Date = Match Date (dd/mm/yy)~~

~~HomeTeam = Home Team~~

~~AwayTeam = Away Team~~

~~FTHG = Full Time Home Team Goals~~

~~FTAG = Full Time Away Team Goals~~

Target

FTR = Full Time Result

~~HTHG = Half Time Home Team Goals~~

~~HTAG = Half Time Away Team Goals~~

~~HTR = Half Time Result~~

HS = Home Team Shots

AS = Away Team Shots

HST = Home Team Shots on Target

AST = Away Team Shots on Target

HC = Home Team Corners

AC = Away Team Corners

HF = Home Team Fouls Committed

AF = Away Team Fouls Committed

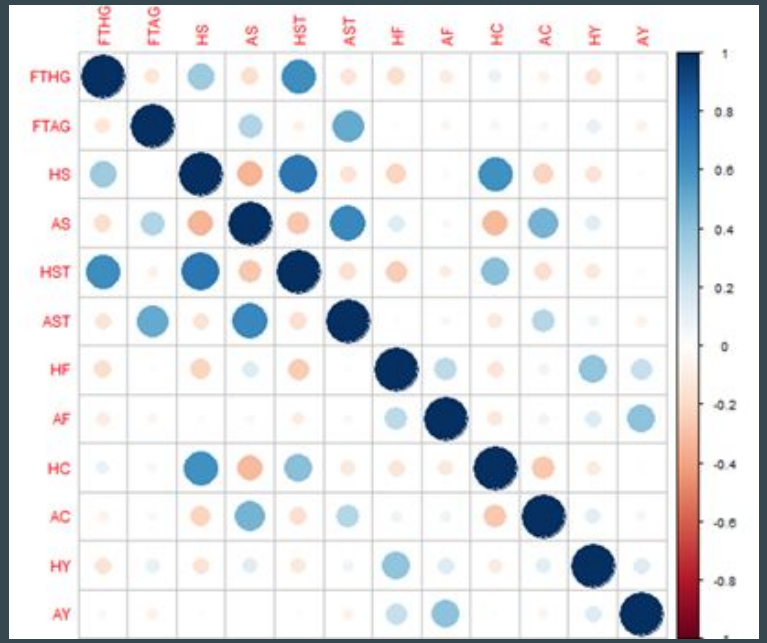
HY = Home Team Yellow Cards

AY = Away Team Yellow Cards

HR = Home Team Red Cards

AR = Away Team Red Cards

Correlation Test



More Feature Modification

> Home/Away team shots accuracy

HSTR(ASTR) = HS(AS) / HST(AST)

> Shots on target accuracy difference

STRdf = HSTR - ASTR

> Fouls committed difference

Fdf = HF - AF

> Yellow Cards difference

Ydf = HY - AY

> Red Cards difference

Rdf = HR - AR

> Corners difference

Cdf = HC - AC

> Added new feature HT (Home Stadium Attendance).

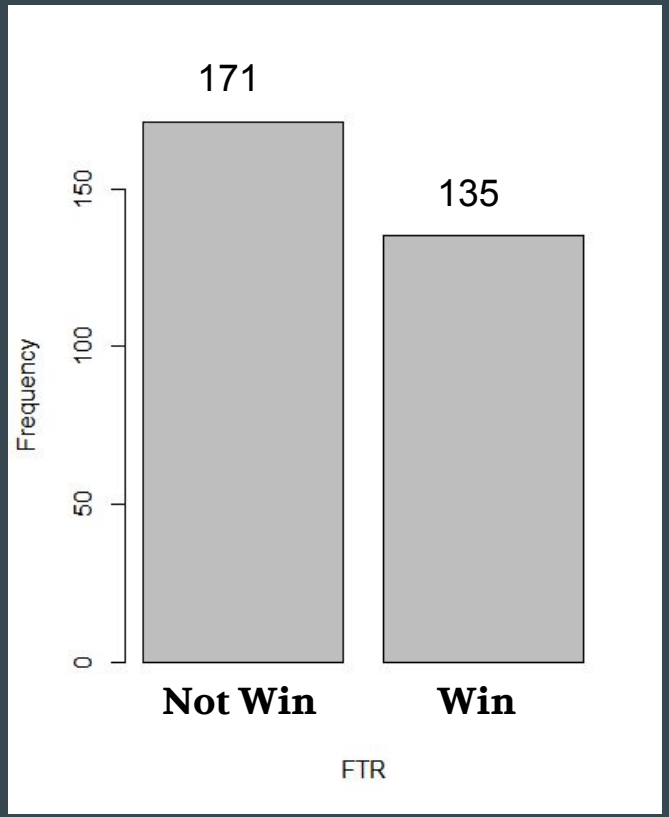
> Changed target feature (FTR) from 3 classes to 2 classes: Win or Not Win

Final Data Set (before selection)

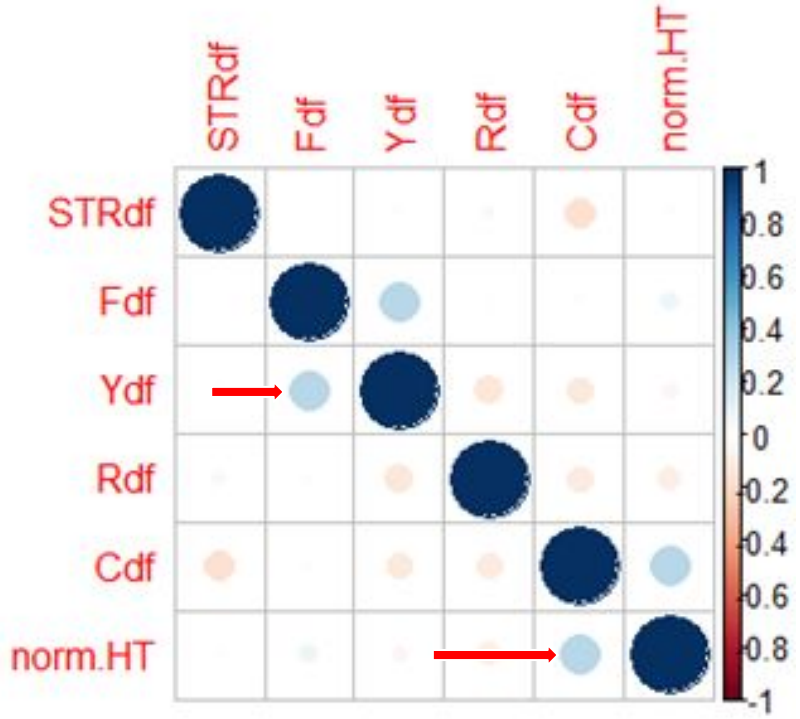
	A	B	C	D	E	F	G
1	FTR	HT	STRdf	Fdf	Ydf	Rdf	Cdf
2	Win	75000	0.191304	-2	0	0	7
3	NotWin	29017	-0.21364	-2	-1	0	3
4	NotWin	16647	0.006494	-1	-1	0	-4
5	Win	81178	0.211765	-1	-1	0	-2
6	Win	29085	0.026667	-6	1	0	8
7	NotWin	31053	-0.02381	5	-1	0	4
8	NotWin	40935	-0.17917	-9	-2	0	-1
9	NotWin	51983	-0.26984	5	1	0	10
10	Win	28945	0.138462	-5	0	0	4
11	NotWin	49704	-0.13333	-5	-1	0	4
12	NotWin	46676	-0.01299	-4	-1	0	3
13	NotWin	48676	0.2	4	1	0	-2
14	Win	53700	0.02381	-4	0	-1	-5
15	NotWin	41246	0.136364	-6	0	0	-5
16	NotWin	27615	0.1	5	0	0	-1

- **306 instances**
- **No missing value**

FTR(Full Time Result)



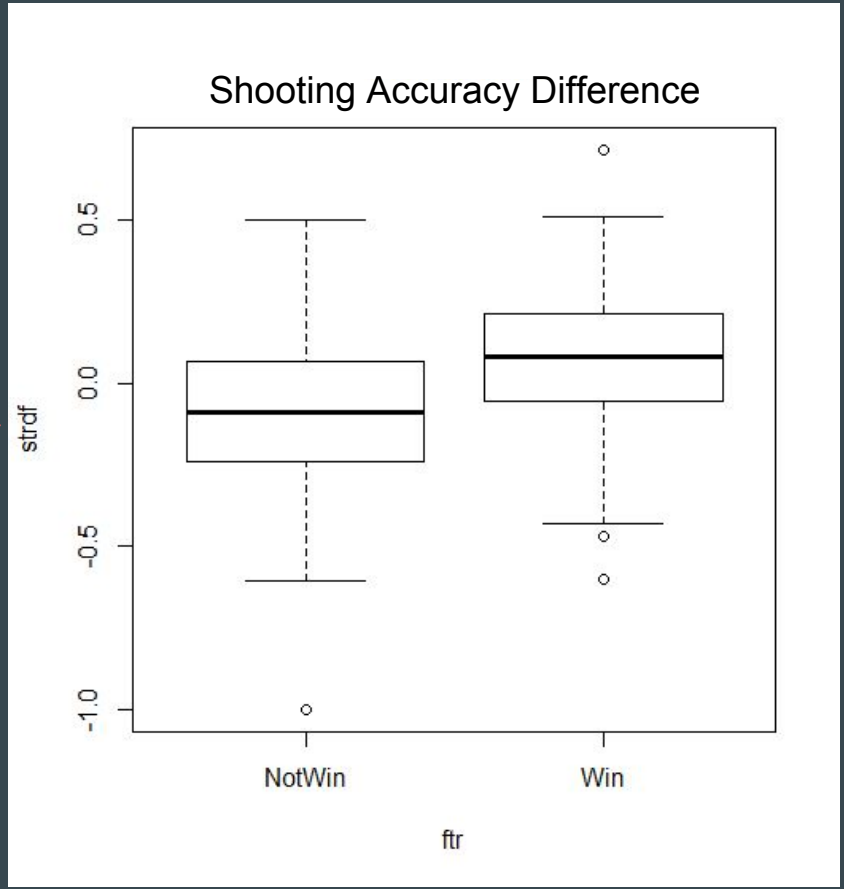
New Correlation Plot



Feature Selection

Attribute Importance Ranking

Feature	Importance
STRdf	36.3399895
HT	25.8640686
Fdf	-0.9601667
Ydf	8.6205440
Cdf	7.3069773
Rdf	6.3807209



Backward Greedy Search

Results (1~10):

FTR ~ STRdf + Fdf + Ydf + Cdf + HT
FTR ~ STRdf + Fdf + Ydf + HT
FTR ~ STRdf + Fdf + Ydf + Rdf + Cdf + HT
FTR ~ STRdf + Fdf + Ydf + HT
FTR ~ STRdf + Fdf + Rdf + Cdf + HT
FTR ~ STRdf + Fdf + Ydf + Rdf + HT
FTR ~ STRdf + Fdf + Ydf + Rdf + HT
FTR ~ STRdf + Fdf + Ydf + Rdf + Cdf
FTR ~ STRdf + Fdf + Ydf + Rdf + Cdf + HT
FTR ~ STRdf + Fdf + Rdf + Cdf + HT

Times each feature was selected:

STRdf = 10
Fdf = 10
HT = 9
Ydf = 8
Rdf = 7
Cdf = 6

NO feature was eliminated yet.

Modeling

< 6 potential classifiers were tested: One Rule, Naïve Bayes, Random Tree, PART, Random Forest and RIPPER.

< linear and logistic regression, and SVM were also tested.

< The best classifiers: **1.Naïve Bayes:** 70.19%
2.RIPPER: 65.38%

Model Evaluation: RIPPER (65.38%)

! Features: Cdf (corner) and Fdf (fouls) were eliminated.

Rules:

```
(STRdf >= -0.119048) and (HT >= 49704) => FTR=Win (53.0/11.0)
 => FTR=NotWin (149.0/48.0)

Number of Rules : 2
```

=== Confusion Matrix ===

Win	NotWin	<--classified as
18	**27**	Win
9	50	NotWin

	TP Rate	FP Rate	Precision	Recall
win	0.400	0.153	0.667	0.400
Not Win	0.847	0.600	0.649	0.847

======== Evaluation result =======

Kappa statistic	0.26
Mean absolute error	0.44
Root mean squared error	0.48
Relative absolute error	88.35%
Root relative squared error	96.97%
Total Number of Instances	104

Model Evaluation: Naïve Bayes (70.19%)

! Features: Cdf (corner) and Fdf (fouls) were eliminated.

Attribute	Class Win (0.45)	NotWin (0.55)
HT Attendance		
mean	49734.8196	38815.645
std. dev.	20687.2752	15264.9473
weight sum	90	112
precision	3902.4706	3902.4706
STRdf Shooting accuracy		
mean	0.0773	-0.0555
std. dev.	0.1979	0.23
weight sum	90	112
precision	0.0064	0.0064
Ydf Yellow card difference		
mean	-0.4889	-0.0179
std. dev.	1.6882	1.6636
weight sum	90	112
precision	1	1
Rdf Red card difference		
mean	-0.0444	0
std. dev.	0.3624	0.3273
weight sum	90	112
precision	1	1

======== Evaluation result =======

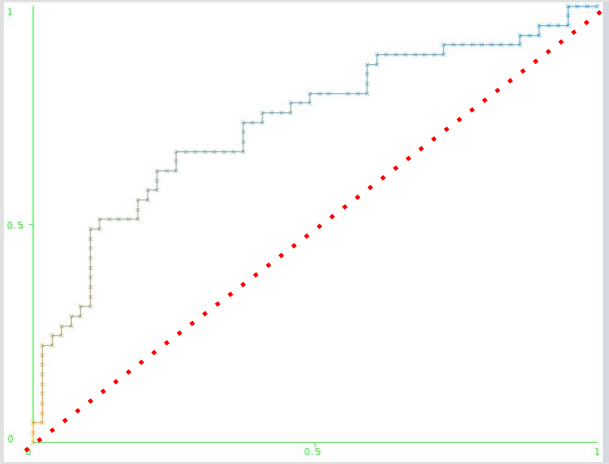
Kappa statistic	0.36
Mean absolute error	0.40
Root mean squared error	0.46
Relative absolute error	81.52%
Root relative squared error	92.29%
Total Number of Instances	104

=== Confusion Matrix ===

Win	NotWin	<--classified as
20	**25**	Win
6	53	NotWin

	TP Rate	FP Rate	Precision	Recall
win	0.444	0.102	0.769	0.444
Not Win	0.898	0.556	0.679	0.898

ROC Curve



Conclusion

Possible deployments:

- *Training:* 1. To increase shooting accuracy. 2. To play “cleaner” (less cards gained)
- *Future plan:* To build larger stadium (more fans)
- *Betting:* To predict the result based on past data.

Concerns:

- Other factors to a match’s result: Average age; judges; injuries; possession rate, passing accuracy….
- Dataset size: more instances