FOOTBALL (SOCCER) MATCH RESULT PROJECTIONS

Nikolai Stambler-Tennant

Brown University – DSI

October 19th, 2022

https://github.com/Niktennant/Premier-League-Result-Projections.git

INTRODUCTION

The problem?

- Predicting football matches outcomes - specifically in the Premier League.

Importance?

- Billion $ industry, sports betting, personal interest

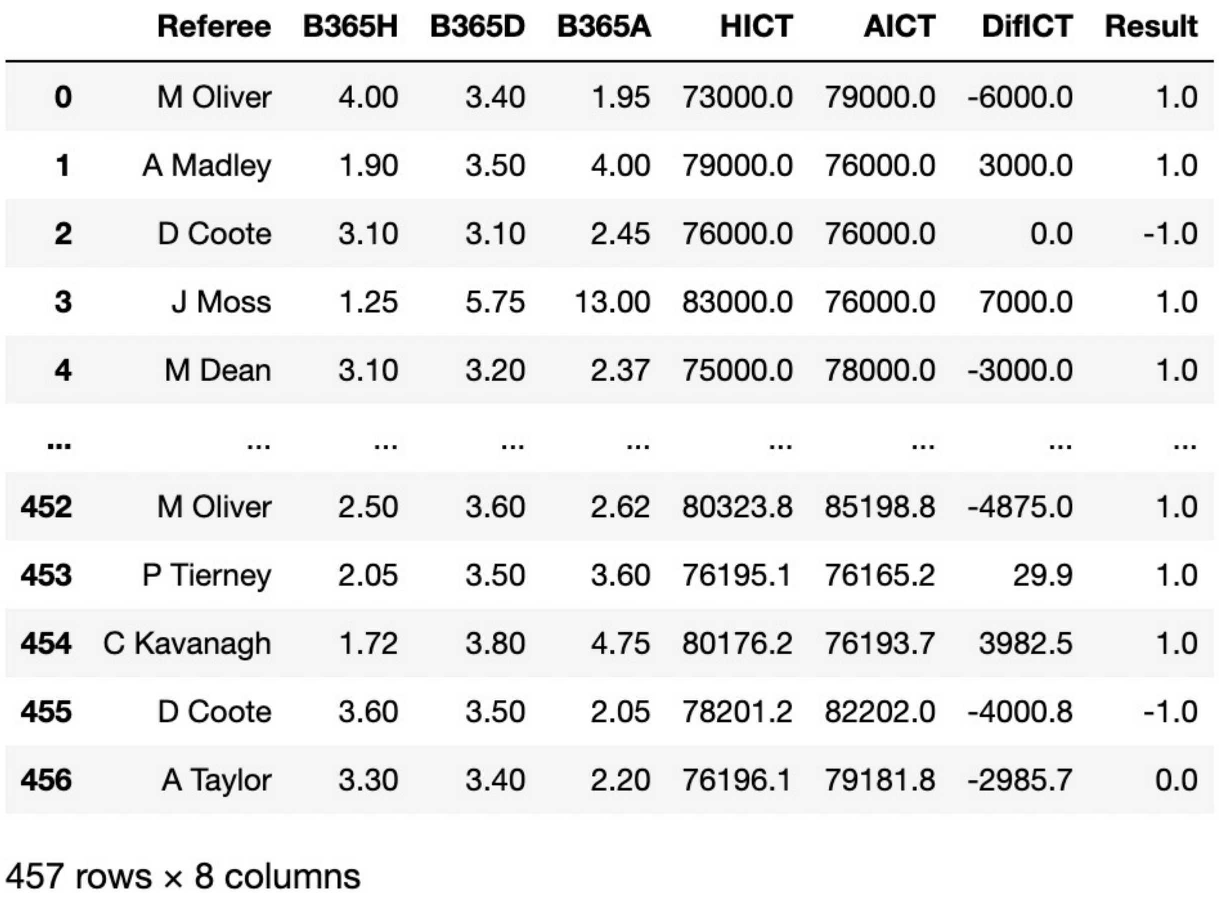
Type?

- Multi-class Classification (H : 1.0, D : 0.0, A : -1.0)

Data Origin?

- Multiple sources
  - Fantasy Premier League API
  - FIFA 22 & FIFA 23
  - PL Match Records

MY DATA

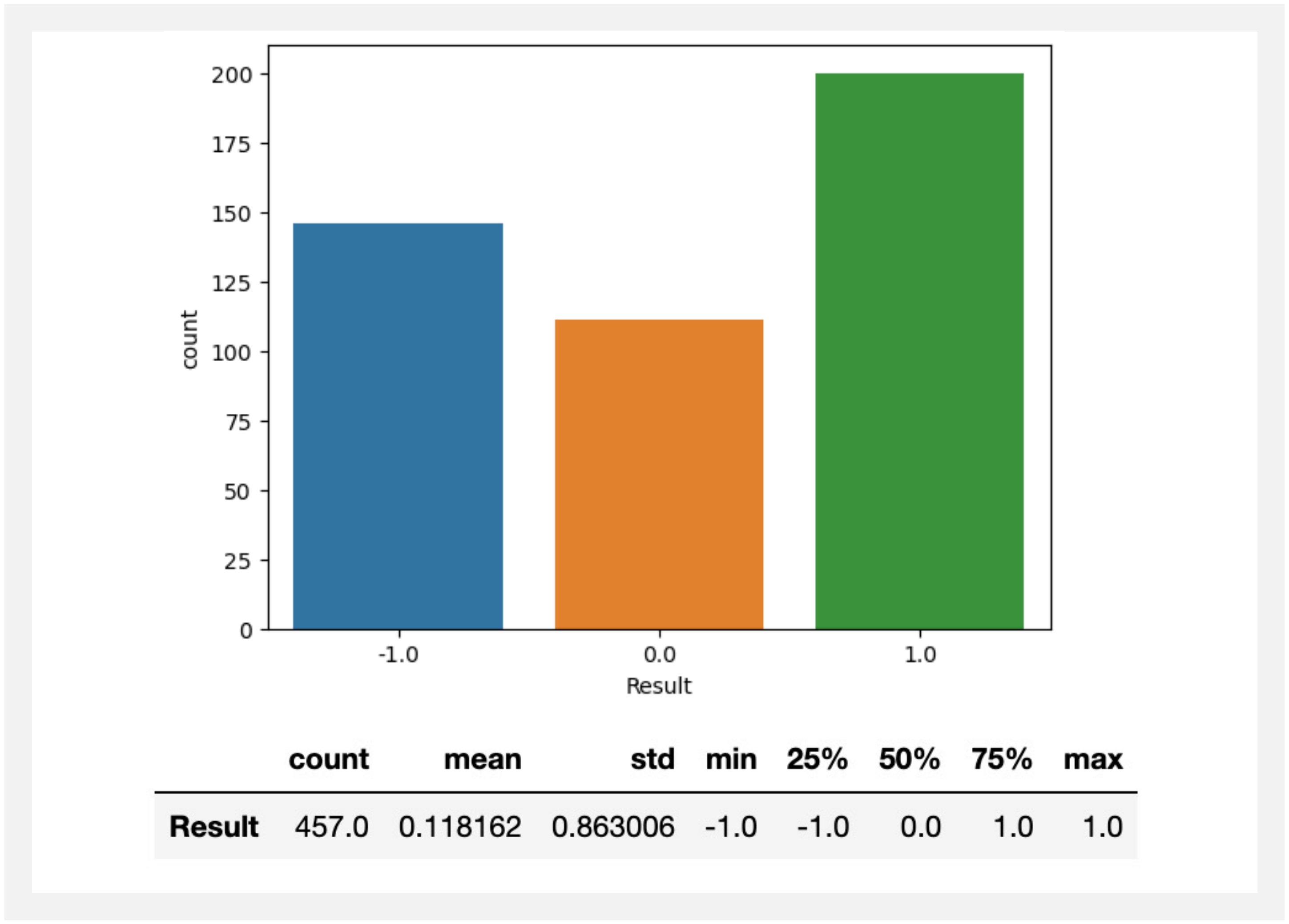


	Referee	B365H	B365D	B365A	HICT	AICT	DifICT	Result
0	M Oliver	4.00	3.40	1.95	73000.0	79000.0	-6000.0	1.0
1	A Madley	1.90	3.50	4.00	79000.0	76000.0	3000.0	1.0
2	D Coote	3.10	3.10	2.45	76000.0	76000.0	0.0	-1.0
3	J Moss	1.25	5.75	13.00	83000.0	76000.0	7000.0	1.0
4	M Dean	3.10	3.20	2.37	75000.0	78000.0	-3000.0	1.0
...	...	...	...	...	...	...	...	...
452	M Oliver	2.50	3.60	2.62	80323.8	85198.8	-4875.0	1.0
453	P Tierney	2.05	3.50	3.60	76195.1	76165.2	29.9	1.0
454	C Kavanagh	1.72	3.80	4.75	80176.2	76193.7	3982.5	1.0
455	D Coote	3.60	3.50	2.05	78201.2	82202.0	-4000.8	-1.0
456	A Taylor	3.30	3.40	2.20	76196.1	79181.8	-2985.7	0.0

457 rows × 8 columns

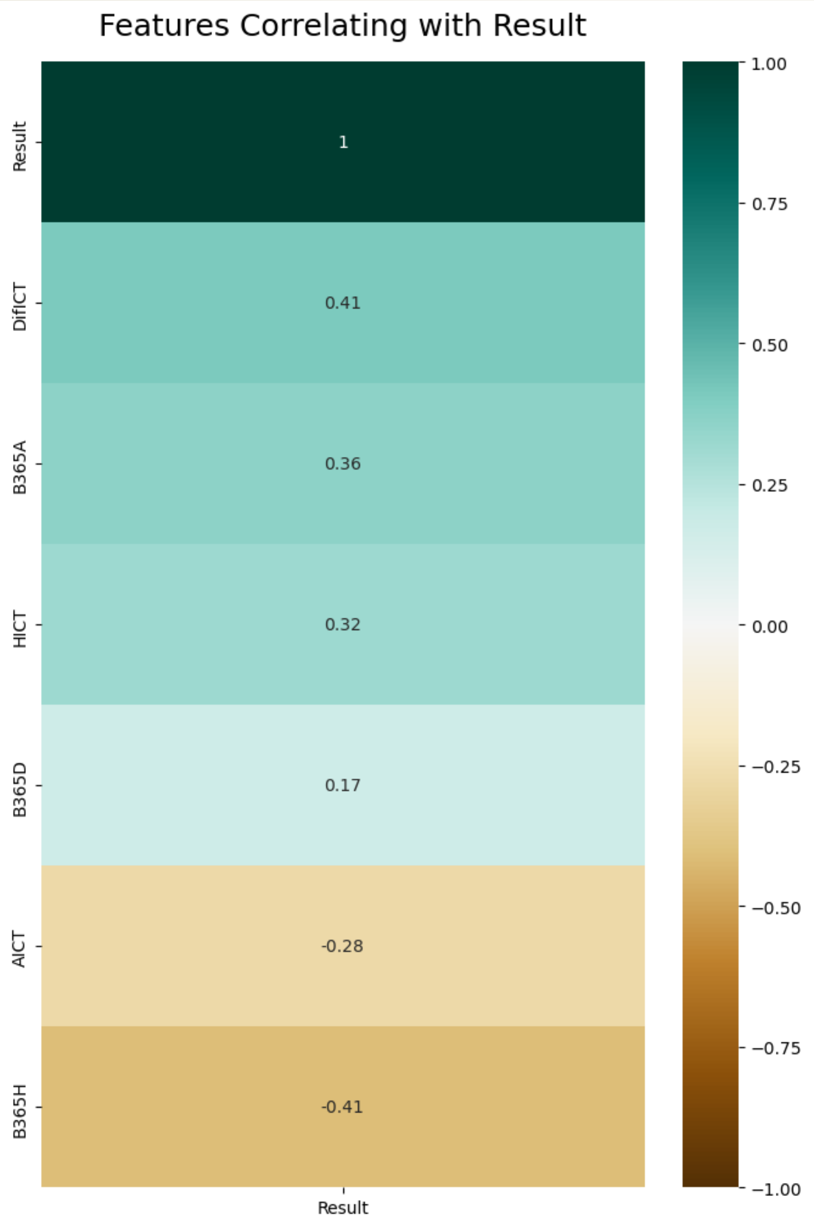
EDA - TARGET VARIABLE (RESULT)

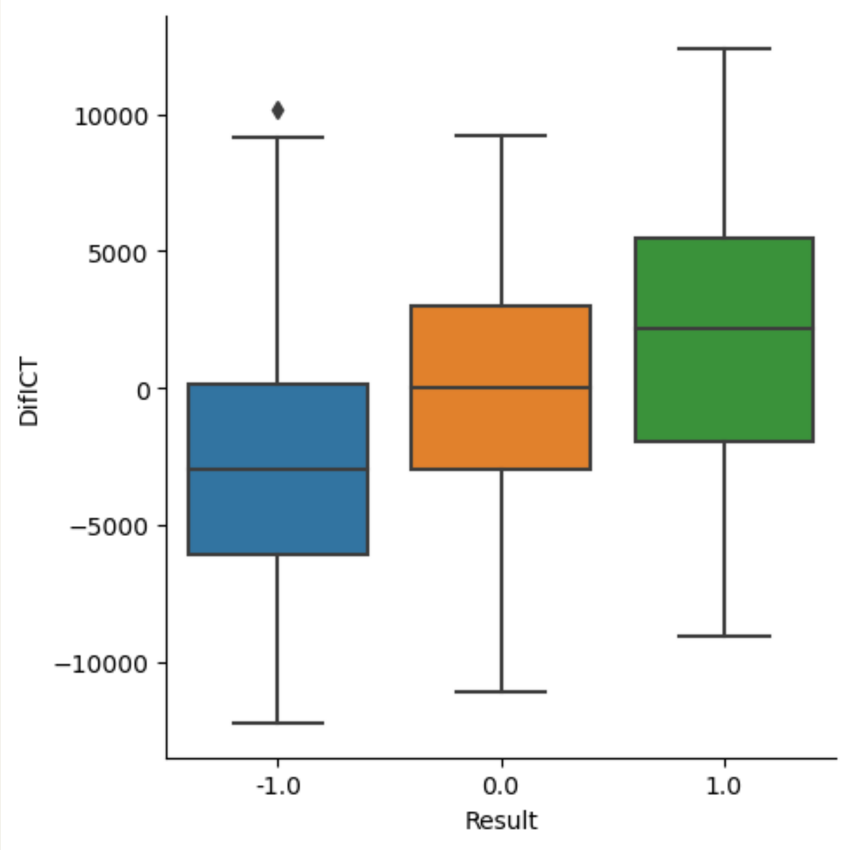
- Home Wins – 200 (43.8%)
- Away Wins – 146 (31.9&)
- Draws – 111 (24.3%)



	count	mean	std	min	25%	50%	75%	max
Result	457.0	0.118162	0.863006	-1.0	-1.0	0.0	1.0	1.0

EDA- PT. 2


Features Correlating with Result
Result
DifICT
B365A
HICT
B365D
AICT
B365H
1
0.41
0.36
0.32
0.17
-0.28
-0.41
Result
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


10000
5000
0
−5000
−10000
DifICT
-1.0
0.0
1.0
Result

SPLITTING/PREPROCESSING

SPLIT

- Time Series (no group structure)
- Lagged data (5 matches)
- 60/20/20 split with df sorted by date (ascending order)
  - 271/91/90

PREPROCESSES

- StandardScaler: continuous features w/ normal/tailed distribution
- One-hot: 'Referee' - unordered categorical data
- Min/MaxScaler: HICT/AICT (73000- 86309)

(FEATURES, DATA POINTS)

- Before preprocess (457, 8)
- After lag (452, 38)
- After preprocessing (452, 58)

MISSING VALUES?

- Nan