PREDICTING THE OUTCOME OF 2020 ENGLISH PREMIER LEAGUE (EPL) FOOTBALL MATCHES

Group Name: `Group D`
Department of Computer Science
University College London
London, WC1E 6BT

January 6, 2020

1 Introduction [Terry]

2 Data Transformation & Exploration [Yun]

At first sight, we found that:

- The shape of the data frame is 4180 rows x 73 columns, but some columns are empty and unnamed.
- There are two different date formats, "%d%m%y" and "%d%m%Y".
- The involved data is from 2008-08-16 to 2019-05-12 (i.e. totally 11 seasons).

	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	...	Unnamed: 63	Unnamed: 64	U 6
0	16/08/08	Arsenal	West Brom	1	0	H	1	0	H	H Webb	...	NaN	NaN	N
1	16/08/08	Bolton	Stoke	3	1	H	3	0	H	C Foy	...	NaN	NaN	N
2	16/08/08	Everton	Blackburn	2	3	A	1	1	D	A Marriner	...	NaN	NaN	N
3	16/08/08	Hull	Fulham	2	1	H	1	1	D	P Walton	...	NaN	NaN	N
4	16/08/08	Middlesbrough	Tottenham	2	1	H	0	0	D	M Atkinson	...	NaN	NaN	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4175	12/05/2019	Liverpool	Wolves	2	0	H	1	0	H	M Atkinson	...	NaN	NaN	N
4176	12/05/2019	Man United	Cardiff	0	2	A	0	1	A	J Moss	...	NaN	NaN	N
4177	12/05/2019	Southampton	Huddersfield	1	1	D	1	0	H	L Probert	...	NaN	NaN	N
4178	12/05/2019	Tottenham	Everton	2	2	D	1	0	H	A Marriner	...	NaN	NaN	N
4179	12/05/2019	Watford	West Ham	1	4	A	0	2	A	C Kavanagh	...	NaN	NaN	N

4180 rows × 73 columns

Fig. 1

2.1 Data Cleaning

After we dropped the unnamed columns, the number reduced to 22.

We verified that there is no row containing invalid values (i.e., None, NaN, infinite or overflowed number), so we don't need to drop any rows. The size remains 4180.

We then unified the date formats, converting into “%Y-%m-%d” for later exploration and transformation.

2.2 Initial Data Exploration

2.2.1 Number of matches per season

The full set is of huge amount. To help learn the data, we separated rows by date from August to May (i.e., one season) to check how many matches there are per season.

```
2008 [380 rows x 22 columns]
2009 [380 rows x 22 columns]
2010 [380 rows x 22 columns]
2011 [380 rows x 22 columns]
2012 [380 rows x 22 columns]
2013 [380 rows x 22 columns]
2014 [380 rows x 22 columns]
2015 [380 rows x 22 columns]
2016 [380 rows x 22 columns]
2017 [380 rows x 22 columns]
2018 [380 rows x 22 columns]
```

Fig. 2

We found that the number of matches each season stays constant (380).

2.2.2 Percentage of match result

We also computed the average percentage of each match result per season and that over the 11 years. See Fig. 3

```
========================================
2008 [380]
----------------------------------------
home team wins: 45.526%
away team wins: 28.947%
draw: 25.526%

========================================
2009 [380]
----------------------------------------
home team wins: 50.789%
away team wins: 23.947%
draw: 25.263%

========================================
2010 [380]
----------------------------------------
home team wins: 47.105%
away team wins: 23.684%
draw: 29.211%

========================================
2011 [380]
----------------------------------------
home team wins: 45.000%
away team wins: 30.526%
draw: 24.474%

========================================
2012 [380]
----------------------------------------
home team wins: 43.684%
away team wins: 27.895%
draw: 28.421%

========================================
2013 [380]
----------------------------------------
home team wins: 47.105%
away team wins: 32.368%
draw: 20.526%
```

(a)

```
========================================
2014 [380]
----------------------------------------
home team wins: 45.263%
away team wins: 30.263%
draw: 24.474%

========================================
2015 [380]
----------------------------------------
home team wins: 41.316%
away team wins: 30.526%
draw: 28.158%

========================================
2016 [380]
----------------------------------------
home team wins: 49.211%
away team wins: 28.684%
draw: 22.105%

========================================
2017 [380]
----------------------------------------
home team wins: 45.526%
away team wins: 28.421%
draw: 26.053%

========================================
2018 [380]
----------------------------------------
home team wins: 47.632%
away team wins: 33.684%
draw: 18.684%

========================================
Overall [4180]
----------------------------------------
home team wins: 46.196%
away team wins: 28.995%
draw: 24.809%
```

(b)

Fig. 3

From the result we noticed that in all cases the result 'home team wins' ('H') is of the highest probability, and 'H':'A':'D' $\approx$ 5:3:2 in general.

2.2.3 Relationship between attributes

We plotted a Pearson Correlation Heatmap (Fig. 4) to see the top 10 features related to the match result (FTR).

As shown in the graph, the top 10 features are:

HTR, FTHG, HTHG, HST, HS, HR, AS, AST, HTAG, FTAG,

ordered from the greatest to least.

It is notable that the goal scored at full time (FTHG, FTAG) & goal scored at half time (HTHG, HTAG) and the total number of shots on goal (HS, AS) & that on target (HST, AST) are the two pairs of data which are highly correlated (> 0.65).

2.3 Feature Construction

So, within the top 10 we picked FTHG, FTAG , HS, AS, HR, AR to create features:

- FTHG, FTAG $\Rightarrow$ the cumulative full-time goal difference by home team and away team [HCGD, ACGD]

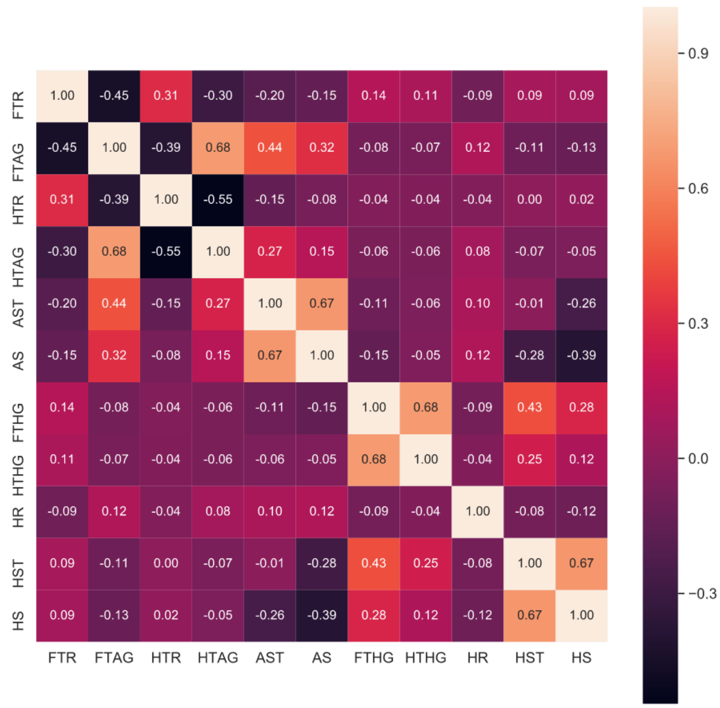

Fig. 4

- HS, AS ⇒ the average number of shots on goal in the past 3 matches by home team and away team [HAHS, AAHS]

- HR, AR (as features directly)

Apart from that, we also derived features from the following attributes:

- Date ⇒ the delta time from last match of home team and away team [HDT, ADT]

- HomeTeam, AwayTeam ⇒ the distance needed to travel for the away team (with the help of extra data source) [DIS]

- FTR ⇒ the performance of past 3 matches of the home team and away team [HM1,AM1, HM2,AM2. HM3,AM3]

Due to the lack of data in the beginning of each year, there are a few rows containing empty values. After removing these rows and also the intermediate data (which we used to create features), the feature set is shown in Fig. 5.

2.4 Second Data Exploration - Analyse Numerical Features

To learn the characteristics of each feature, we derive the minimum, maximum, median, mean, variance and standard deviation:

	HR	AR	HCGD	ACGD	HAHS	AAHS	HDT	ADT	DIS	HM1	AM1	HM2	AM2	HM3	AM3
29	0	0	-2	3	13.000000	19.666667	14.0	14.0	290.604156	L	W	D	L	W	W
32	0	1	4	5	12.000000	12.333333	13.0	13.0	261.179108	W	D	W	W	L	W
33	1	0	-2	-4	8.333333	9.333333	14.0	14.0	159.281448	L	L	W	D	D	W
34	0	0	-2	1	10.333333	14.666667	14.0	14.0	420.982727	W	W	L	L	L	W
35	0	0	-2	1	10.333333	9.666667	14.0	14.0	175.303436	D	W	L	L	L	W
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4175	0	0	65	3	16.333333	13.666667	8.0	8.0	108.891106	W	W	W	W	W	W
4176	0	0	13	-37	14.000000	12.666667	7.0	8.0	229.968140	D	L	D	L	L	L
4177	0	0	-20	-54	13.666667	8.333333	8.0	7.0	306.418793	L	D	D	L	D	L
4178	0	0	28	8	18.000000	19.000000	8.0	9.0	283.650818	L	W	L	D	W	W
4179	1	0	-4	-6	13.333333	14.666667	7.0	8.0	33.253616	L	W	L	W	D	D

3845 rows × 15 columns

Fig. 5

```
========================================
HR [size: 3845]
----------------------------------------
min: 0.0000
max: 2.0000
median:0.0000
mean: 0.0583
variance: 0.0585
standard deviation: 0.2419

========================================
AR [size: 3845]
----------------------------------------
min: 0.0000
max: 2.0000
median:0.0000
mean: 0.0887
variance: 0.0891
standard deviation: 0.2986

========================================
HCGD [size: 3845]
----------------------------------------
min: -54.0000
max: 76.0000
median:-2.0000
mean: -0.1545
variance: 268.4770
standard deviation: 16.3874
```

(a)

```
========================================
ACGD [size: 3845]
----------------------------------------
min: -54.0000
max: 78.0000
median:-2.0000
mean: 0.2195
variance: 266.7305
standard deviation: 16.3340

========================================
HAHS [size: 3845]
----------------------------------------
min: 3.3333
max: 27.0000
median:12.0000
mean: 12.4158
variance: 12.0246
standard deviation: 3.4681

========================================
AAHS [size: 3845]
----------------------------------------
min: 3.6667
max: 28.6667
median:12.3333
mean: 12.8305
variance: 12.3910
standard deviation: 3.5205
```

(b)

```
========================================
HDT [size: 3845]
----------------------------------------
min: 2.0000
max: 27.0000
median:7.0000
mean: 7.4637
variance: 11.7795
standard deviation: 3.4326

========================================
ADT [size: 3845]
----------------------------------------
min: 2.0000
max: 22.0000
median:7.0000
mean: 7.4780
variance: 11.9130
standard deviation: 3.4520

========================================
DIS [size: 3845]
----------------------------------------
min: 0.9710
max: 473.8653
median:179.0834
mean: 187.5142
variance: 12289.0815
standard deviation: 110.8705
```

(c)

Fig. 6

From the figure, we can draw such conclusions:

- HR & AR: The range is very small (2). From the median, the mean and also the small variance we can know that most values are 0 (as these two features are discrete) while value=2 is of low occurrence.
- HCGD & ACGD: Large range (> 130) with negative values involved. The median and the mean demonstrates that there is a relatively greater number of negative values within the data set.
- HAHS & AAHS: Moderate range (around 25) with all positive values. The median and the mean is at the half of the range while the variance is reasonable.
- HDT & ADT: Similar moderate range (around 25) and variance with the above pair of data. But the median and the mean is at the one third of the range. Outliers may exist.
- DIS: Large range (> 450) with all positive values. Reasonable median and mean. But from the variance we can know that the value fluctuates significantly.

- Comparing to the other features, the values of HR & AR are too small while that of DIS too large.

2.5 Data Transformation

2.5.1 Label mapping

We mapped the label (i.e., FTR) into numbers for later model training by the rule:

- 'H' $\rightarrow$ 1
- 'A $\rightarrow$ 0
- 'D' $\rightarrow$ 2

2.5.2 Rescale and standardize numerical features

With the conclusions from 3.3, we applied the z-score standardization and min-max rescaling to the numerical features.

2.5.3 Transform categorical features

The categorical data within the feature set is:

HM1,AM1, HM2,AM2, HM3,AM3,

which only take the values 'W', 'L', 'D'.

So we introduced the binary features

HM1_W, HM1_L, HM1_D
AM1_W, AM1_L, AM1_D
......
AM3_W, AM3_L, AM3_D

such that if, for example, HM1 takes the value of 'W', then HM1_W = 1, HM1_L= 0, HM1_D = 0.

3 Methodology Overview [Yanke]

4 Model Training & Validation [Yanke]

5 Results [Yi]

6 Final Predictions on Test Set [Yusi]

7 Conclusion [Terry]

References

[1] Sharma, Mohit. What Steps should one take while doing Data Preprocessing?. June 20th 2018. June 1st 2019. <*https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa*>.

[2] J. González. Scaling/ normalisation/ standardisation: a pervasive question. Oct 18th 2018. June 3st 2019. <*https://quantdare.com/scaling-normalisation-standardisation-a-pervasive-question/*>.