Barclays Premier League winning team prediction

Course : Artificial Intelligence Lab
Group - 09

Team Members

1. Afrin Akther - 14101046
2. Israth Chowdhury - 14101064
3. Monirul Islam Pavel - 14301141
4. Shahriar Hasan Parash - 14101233
5. Akash Biswas - 14101206

Introduction

For some countries Football or Soccer is more than just a game. It is played by **250 million** players in over **200 countries** making it the most popular sport. Each of these countries has a domestic league of their own in which teams compete for being labelled the best football team of that country.

The aim of this project is to evaluate the performance and prediction accuracy by using different regression models.

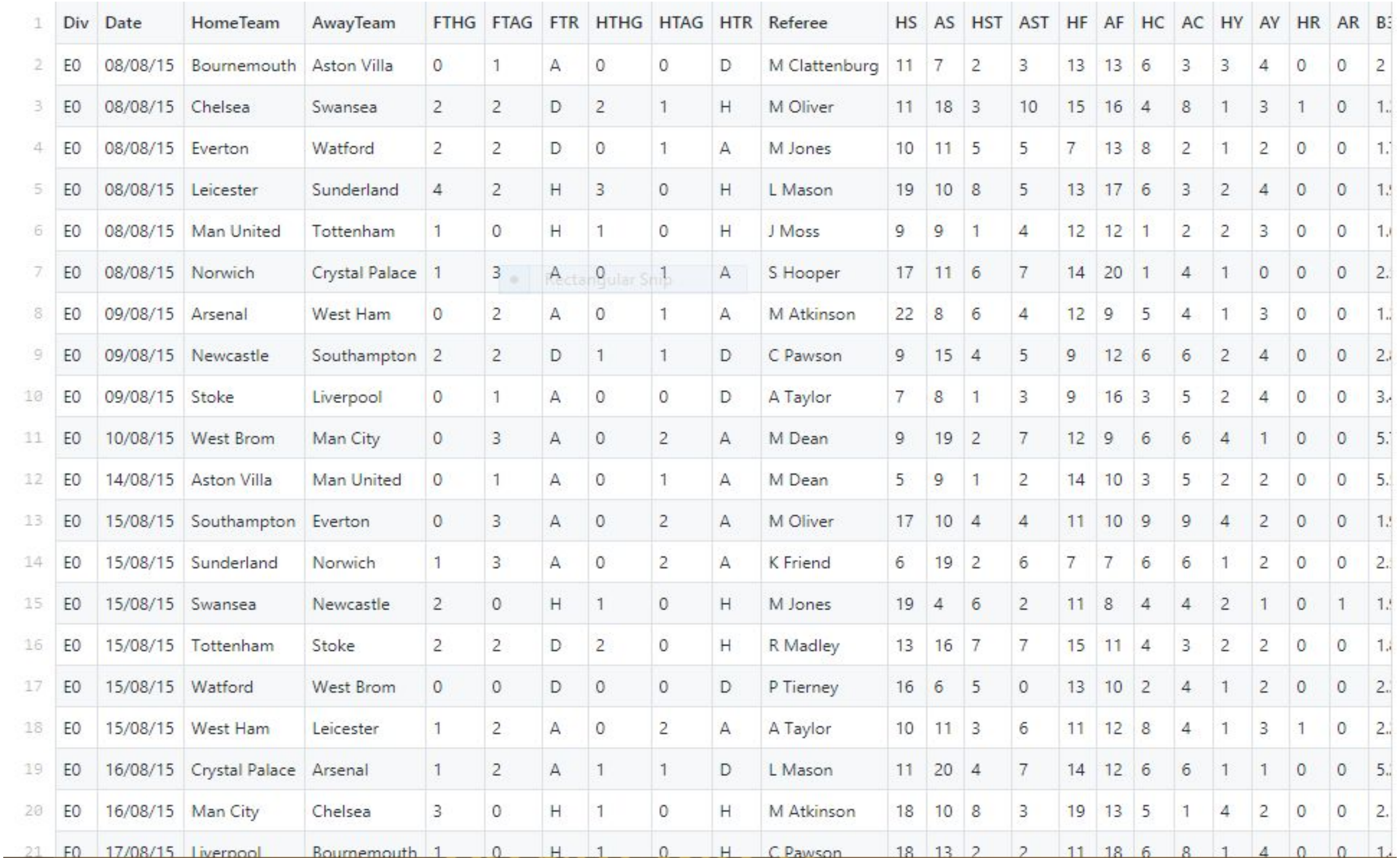
After completion, the purpose of this project is to achieve:

- Evaluate different regression predictive models.
- Perform data transform to improve model performance.
- Implement algorithm tuning to improve model performance.
- Employ ensemble methods and tuning of ensemble methods to improve model performance.

Dataset

- Using features to predict
- Clearing the dataset
- Targeting Full Time Result(FTR)
- Multiclass classification problem

Link: http://football-data.co.uk/englandm.php



Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR	B3
E0	08/08/15	Bournemouth	Aston Villa	0	1	A	0	0	D	M Clattenburg	11	7	2	3	13	13	6	3	3	4	0	0	2
E0	08/08/15	Chelsea	Swansea	2	2	D	2	1	H	M Oliver	11	18	3	10	15	16	4	8	1	3	1	0	1.
E0	08/08/15	Everton	Watford	2	2	D	0	1	A	M Jones	10	11	5	5	7	13	8	2	1	2	0	0	1.
E0	08/08/15	Leicester	Sunderland	4	2	H	3	0	H	L Mason	19	10	8	5	13	17	6	3	2	4	0	0	1.
E0	08/08/15	Man United	Tottenham	1	0	H	1	0	H	J Moss	9	9	1	4	12	12	1	2	2	3	0	0	1.
E0	08/08/15	Norwich	Crystal Palace	1	3	A	0	1	A	S Hooper	17	11	6	7	14	20	1	4	1	0	0	0	2.
E0	09/08/15	Arsenal	West Ham	0	2	A	0	1	A	M Atkinson	22	8	6	4	12	9	5	4	1	3	0	0	1.
E0	09/08/15	Newcastle	Southampton	2	2	D	1	1	D	C Pawson	9	15	4	5	9	12	6	6	2	4	0	0	2.
E0	09/08/15	Stoke	Liverpool	0	1	A	0	0	D	A Taylor	7	8	1	3	9	16	3	5	2	4	0	0	3.
E0	10/08/15	West Brom	Man City	0	3	A	0	2	A	M Dean	9	19	2	7	12	9	6	6	4	1	0	0	5.
E0	14/08/15	Aston Villa	Man United	0	1	A	0	1	A	M Dean	5	9	1	2	14	10	3	5	2	2	0	0	5.
E0	15/08/15	Southampton	Everton	0	3	A	0	2	A	M Oliver	17	10	4	4	11	10	9	9	4	2	0	0	1.
E0	15/08/15	Sunderland	Norwich	1	3	A	0	2	A	K Friend	6	19	2	6	7	7	6	6	1	2	0	0	2.
E0	15/08/15	Swansea	Newcastle	2	0	H	1	0	H	M Jones	19	4	6	2	11	8	4	4	2	1	0	1	1.
E0	15/08/15	Tottenham	Stoke	2	2	D	2	0	H	R Madley	13	16	7	7	15	11	4	3	2	2	0	0	1.
E0	15/08/15	Watford	West Brom	0	0	D	0	0	D	P Tierney	16	6	5	0	13	10	2	4	1	2	0	0	2.
E0	15/08/15	West Ham	Leicester	1	2	A	0	2	A	A Taylor	10	11	3	6	11	12	8	4	1	3	1	0	2.
E0	16/08/15	Crystal Palace	Arsenal	1	2	A	1	1	D	L Mason	11	20	4	7	14	12	6	6	1	1	0	0	5.
E0	16/08/15	Man City	Chelsea	3	0	H	1	0	H	M Atkinson	18	10	8	3	19	13	5	1	4	2	0	0	2.
E0	17/08/15	Liverpool	Bournemouth	1	0	H	1	0	H	C Pawson	18	13	2	2	11	18	6	8	1	4	0	0	1.

Libraries that are used

- **Pandas :** Loading the data, data wrangling and manipulation, reshaping and pivoting of data sets ,data set merging and joining.

- **Scikit-learn :** Libraries for classifiers, model evaluation, metrics, cross-validation

- **Matplotlib and Seaborn :** Data visualization

Algorithms/Methods

- Logistic Regression
- SVM
- XGBoost

Logistic regression

- **Logistic regression** is the appropriate regression analysis to conduct when the dependent variable is dichotomous (multiclass).

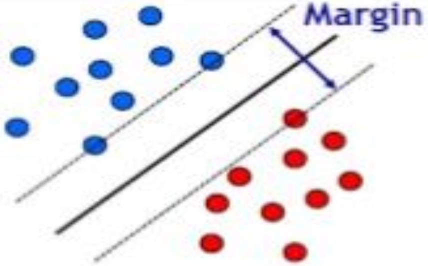
Support Vector Machine (SVM):

- **Basic idea:**

- The SVM tries to find a classifier which maximizes the margin between pos. and neg. data points.

- Up to now: consider linear classifiers

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$$



• Formulation as a convex optimization problem
Find the hyperplane satisfying

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

under the constraints

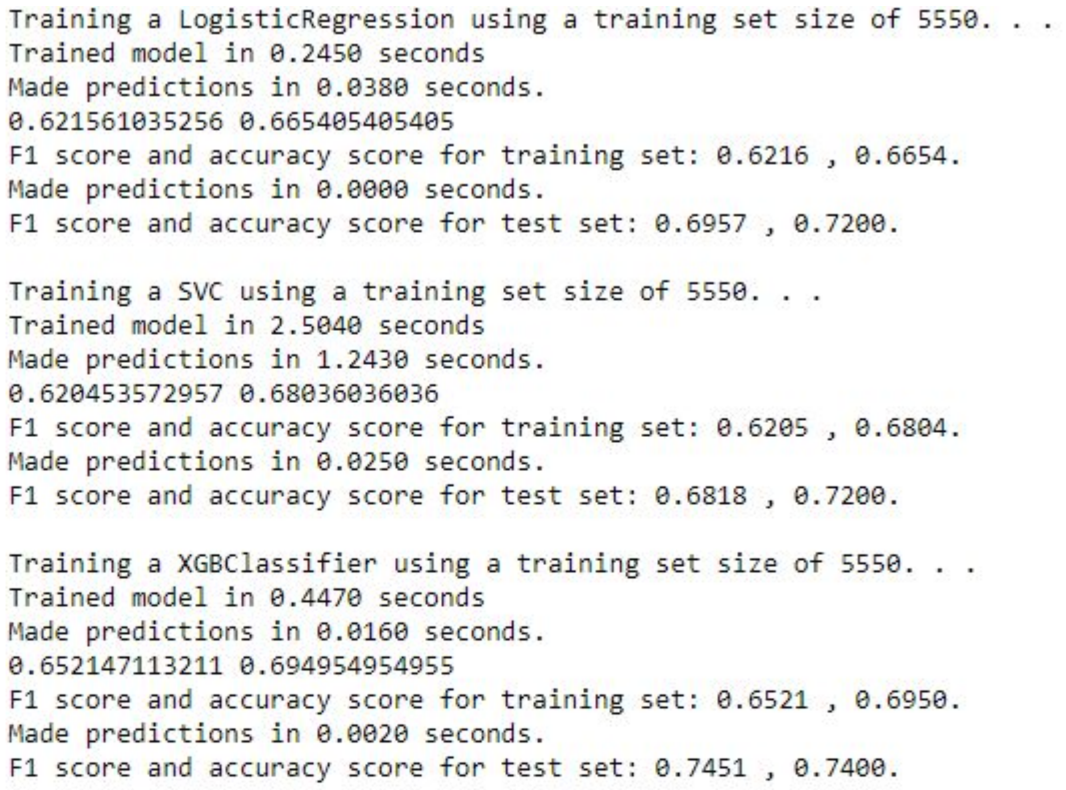
$$t_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) \geq 1 \quad \forall n$$

based on training data points $\mathbf{x}_n$ and target values $t_n \in \{-1, 1\}$.

XGBoost

- XGBoost is used for supervised learning problems, where we use the training data (with multiple features) *xi* to predict a target variable *yi*. Before we dive into trees.
- XGBoost scales beyond billions of examples using far fewer resources than existing systems.

Accuracy Test



```
Training a LogisticRegression using a training set size of 5550. . .
Trained model in 0.2450 seconds
Made predictions in 0.0380 seconds.
0.621561035256 0.665405405405
F1 score and accuracy score for training set: 0.6216 , 0.6654.
Made predictions in 0.0000 seconds.
F1 score and accuracy score for test set: 0.6957 , 0.7200.

Training a SVC using a training set size of 5550. . .
Trained model in 2.5040 seconds
Made predictions in 1.2430 seconds.
0.620453572957 0.68036036036
F1 score and accuracy score for training set: 0.6205 , 0.6804.
Made predictions in 0.0250 seconds.
F1 score and accuracy score for test set: 0.6818 , 0.7200.

Training a XGBClassifier using a training set size of 5550. . .
Trained model in 0.4470 seconds
Made predictions in 0.0160 seconds.
0.652147113211 0.694954954955
F1 score and accuracy score for training set: 0.6521 , 0.6950.
Made predictions in 0.0020 seconds.
F1 score and accuracy score for test set: 0.7451 , 0.7400.
```

Conclusion

Based on the above we can see that the past head to head record for 5 seasons is not a very reliable metric. This can be because of the following reasons:

- Every EPL season is different. Some teams might perform better in a particular season and then be average for the next few years.
- Most teams are generally inconsistent throughout the season, i.e., they may perform better in the first quarter and then slack off in the next quarter. Because of this, there is no certain way to predict the outcome purely based on past head to head record.

- Past head to head record can be really useful for predicting the performance of top teams against the mid table teams. Since the top teams tend to be consistent and the mid table teams inconsistent, past head to head record can be a good metric.
- Similarly, past head to head record alone is unreliable but if combines with another metric like past form, it can give pretty accurate results.

References

- https://github.com/RudrakshTuwani/Football-Data-Analysis-and-Prediction/blob/master/Prediction/Scraping%20and%20Cleaning.ipynb
- https://github.com/llSourcell/Predicting_Winning_Teams/blob/master/Scraping%20and%20Cleaning.ipynb
- https://www.sciencedirect.com/science/article/pii/S0950705106000724

Thank You

Any Question?