Premier League

Premier League Prediction Model

By:

Snehal Shinde (142103013)

Prashanth Sonawane (112003140)

Shyam Aradhye (112003158)

Introduction

This project aims to predict the outcome of Premier League matches using *data science* techniques. We will use historical data to create a model that will help us understand which *variables* are most important in determining match outcomes.

This project will require us to use a range of data science tools and techniques, including data cleaning and preprocessing, feature engineering, and machine learning algorithms. By the end of the project, we hope to have developed a powerful predictive model that can help football fans and bettors make more informed decisions about upcoming Premier League matches.

Objectives

The main objectives are :

- Predict the match results
- Predict the Table at the end of season

Pipeline of our project

1. Data Collection
2. Data Preprocessing
3. Feature Engineering
4. Exploratory Data Analysis
5. Model Selection
6. Model Building
7. Model Evaluation
8. Prediction and Results

Data Collection

Identify the data sources needed for the project, such as team and player statistics, historical match data, and betting odds. Collect the data from reliable and comprehensive sources, such as the Premier League’s official website and sports betting websites. Store the collected data in a structured format, such as a relational database or a spreadsheet.

RAW DATA

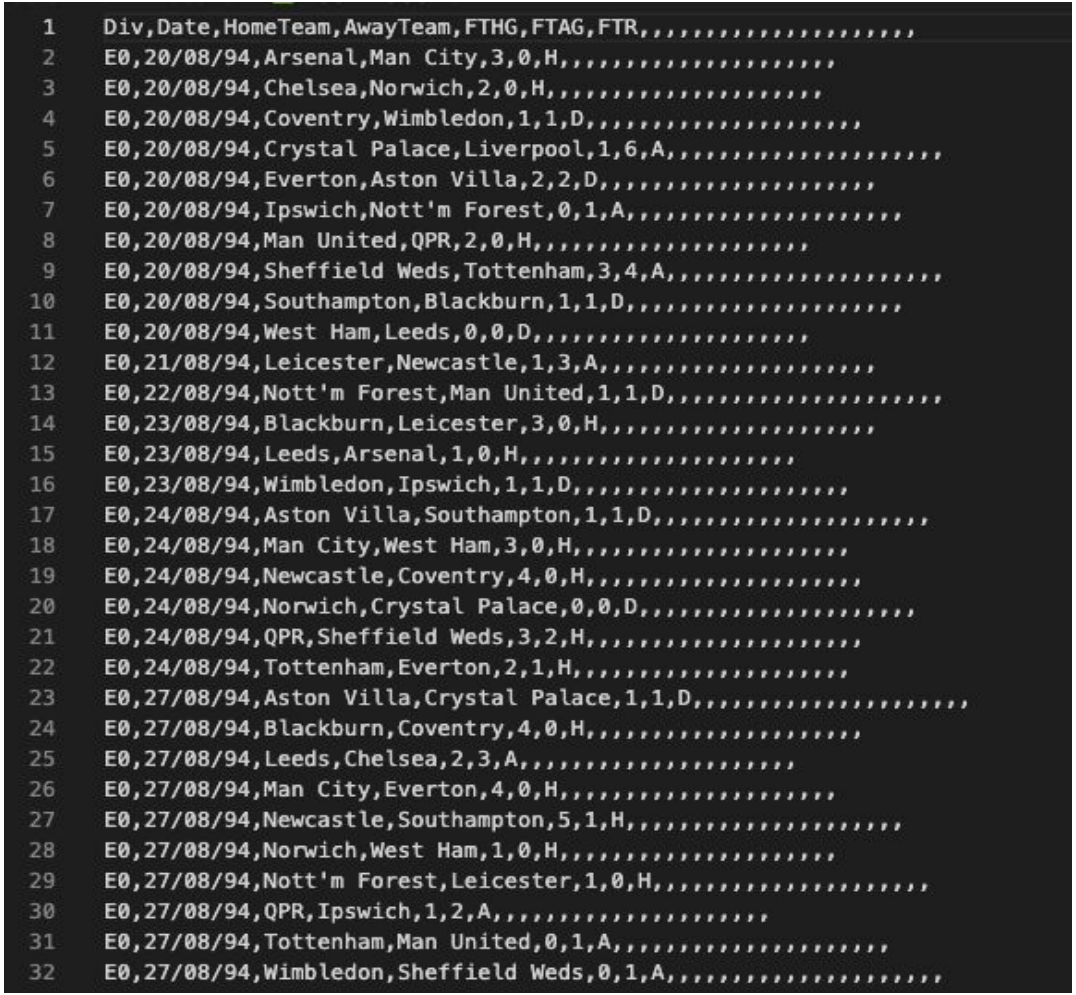
Match-wise data

Season-wise table

Official websites

Sofifa.com

Various stat websites



```
Div,Date,HomeTeam,AwayTeam,FTHG,FTAG,FTR,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Arsenal,Man City,3,0,H,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Chelsea,Norwich,2,0,H,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Coventry,Wimbledon,1,1,D,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Crystal Palace,Liverpool,1,6,A,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Everton,Aston Villa,2,2,D,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Ipswich,Nott'm Forest,0,1,A,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Man United,QPR,2,0,H,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Sheffield Weds,Tottenham,3,4,A,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,Southampton,Blackburn,1,1,D,,,,,,,,,,,,,,,,,,,,,
E0,20/08/94,West Ham,Leeds,0,0,D,,,,,,,,,,,,,,,,,,,,,
E0,21/08/94,Leicester,Newcastle,1,3,A,,,,,,,,,,,,,,,,,,,,,
E0,22/08/94,Nott'm Forest,Man United,1,1,D,,,,,,,,,,,,,,,,,,,,,
E0,23/08/94,Blackburn,Leicester,3,0,H,,,,,,,,,,,,,,,,,,,,,
E0,23/08/94,Leeds,Arsenal,1,0,H,,,,,,,,,,,,,,,,,,,,,
E0,23/08/94,Wimbledon,Ipswich,1,1,D,,,,,,,,,,,,,,,,,,,,,
E0,24/08/94,Aston Villa,Southampton,1,1,D,,,,,,,,,,,,,,,,,,,,,
E0,24/08/94,Man City,West Ham,3,0,H,,,,,,,,,,,,,,,,,,,,,
E0,24/08/94,Newcastle,Coventry,4,0,H,,,,,,,,,,,,,,,,,,,,,
E0,24/08/94,Norwich,Crystal Palace,0,0,D,,,,,,,,,,,,,,,,,,,,,
E0,24/08/94,QPR,Sheffield Weds,3,2,H,,,,,,,,,,,,,,,,,,,,,
E0,24/08/94,Tottenham,Everton,2,1,H,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Aston Villa,Crystal Palace,1,1,D,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Blackburn,Coventry,4,0,H,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Leeds,Chelsea,2,3,A,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Man City,Everton,4,0,H,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Newcastle,Southampton,5,1,H,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Norwich,West Ham,1,0,H,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Nott'm Forest,Leicester,1,0,H,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,QPR,Ipswich,1,2,A,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Tottenham,Man United,0,1,A,,,,,,,,,,,,,,,,,,,,,
E0,27/08/94,Wimbledon,Sheffield Weds,0,1,A,,,,,,,,,,,,,,,,,,,,,
```

DATA PREPROCESSING

- Handled Null/NaN values, Redundant Columns.
- Changed the format of ‘DATE’ column to follow one format.(YYYY-MM-DD)
- Separated one huge file of raw data into multiple files having season-wise data for better analysis and model implementation.
- Getting Rid of teams which didn’t play enough seasons to contribute to proper data analysis and evaluation.

CLEAN DATA

```
Team,Points,Goal_Diff,Win_Rate
Man United,82,38,0.6578947368421053
Newcastle,78,29,0.631578947368421
Liverpool,71,36,0.5263157894736842
Aston Villa,63,17,0.47368421052631576
Arsenal,63,17,0.4473684210526316
Blackburn,61,14,0.47368421052631576
Everton,61,20,0.4473684210526316
Tottenham,61,12,0.42105263157894735
Nott'm Forest,58,-4,0.39473684210526316
West Ham,51,-9,0.3684210526315789
Chelsea,50,2,0.3157894736842105
Leeds,43,-17,0.3157894736842105
Middlesbrough,43,-15,0.2894736842105263
Wimbledon,41,-15,0.2631578947368421
Sheffield Weds,40,-13,0.2631578947368421
Man City,38,-25,0.23684210526315788
Coventry,38,-18,0.21052631578947367
Southampton,38,-18,0.23684210526315788
QPR,33,-19,0.23684210526315788
Bolton,29,-32,0.21052631578947367
```

```
Date,HomeTeam,AwayTeam,FTHG,FTAG,FTR
1995-08-19,Aston Villa,Man United,3,1,H
1995-08-19,Blackburn,QPR,1,0,H
1995-08-19,Chelsea,Everton,0,0,D
1995-08-19,Liverpool,Sheffield Weds,1,0,H
1995-08-19,Man City,Tottenham,1,1,D
1995-08-19,Newcastle,Coventry,3,0,H
1995-08-19,Southampton,Nott'm Forest,3,4,A
1995-08-19,West Ham,Leeds,1,2,A
1995-08-19,Wimbledon,Bolton,3,2,H
1995-08-20,Arsenal,Middlesbrough,1,1,D
1995-08-21,Leeds,Liverpool,1,0,H
1995-08-22,Bolton,Newcastle,1,3,A
1995-08-23,Coventry,Man City,2,1,H
1995-08-23,Everton,Arsenal,0,2,A
1995-08-23,Man United,West Ham,2,1,H
1995-08-23,Nott'm Forest,Chelsea,0,0,D
1995-08-23,QPR,Wimbledon,0,3,A
1995-08-23,Sheffield Weds,Blackburn,2,1,H
1995-08-23,Tottenham,Aston Villa,0,1,A
1995-08-26,Bolton,Blackburn,2,1,H
1995-08-26,Coventry,Arsenal,0,0,D
1995-08-26,Everton,Southampton,2,0,H
1995-08-26,Leeds,Aston Villa,2,0,H
1995-08-26,Man United,Wimbledon,3,1,H
1995-08-26,Middlesbrough,Chelsea,2,0,H
1995-08-26,Nott'm Forest,West Ham,1,1,D
1995-08-26,QPR,Man City,1,0,H
```

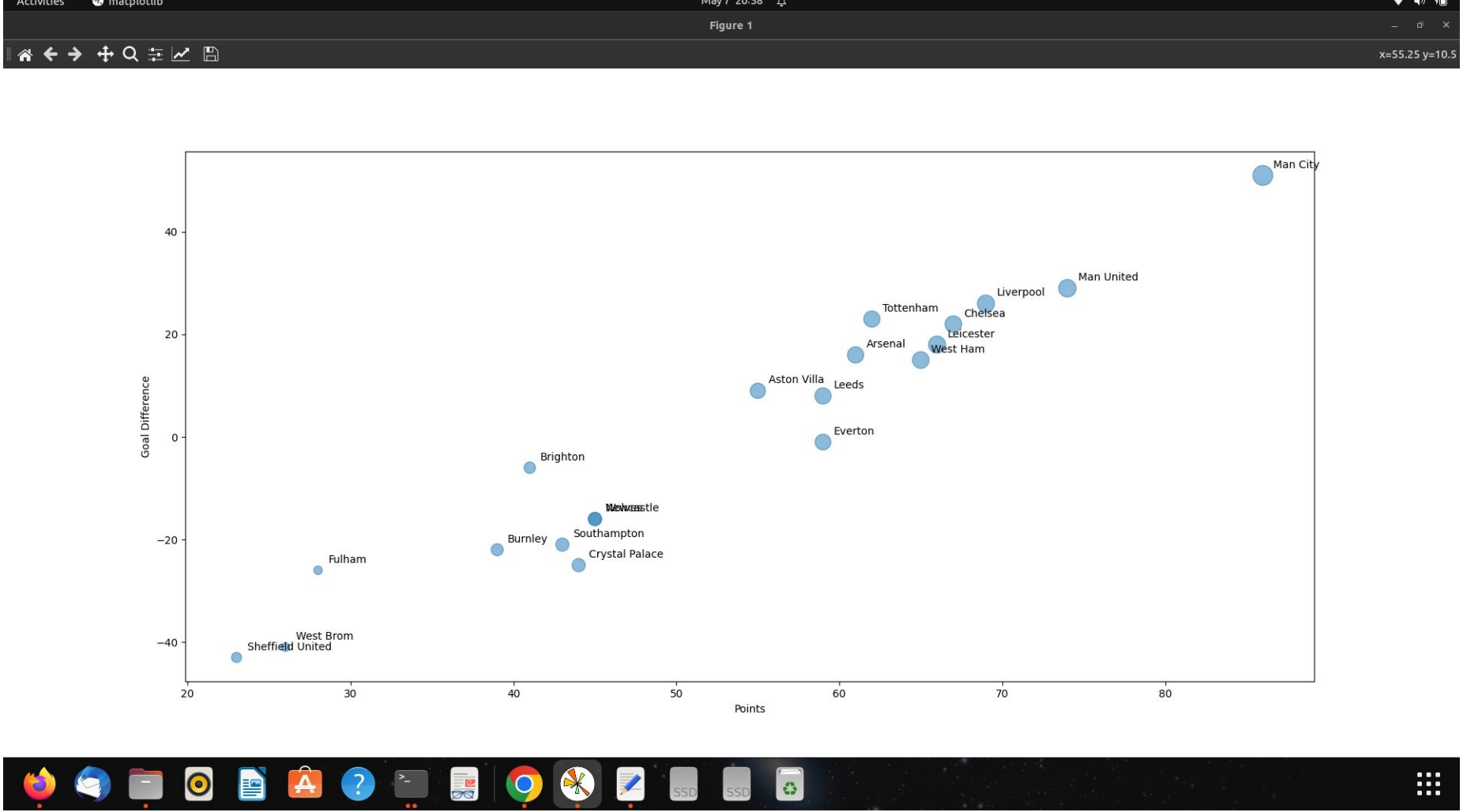
FEATURE ENGINEERING

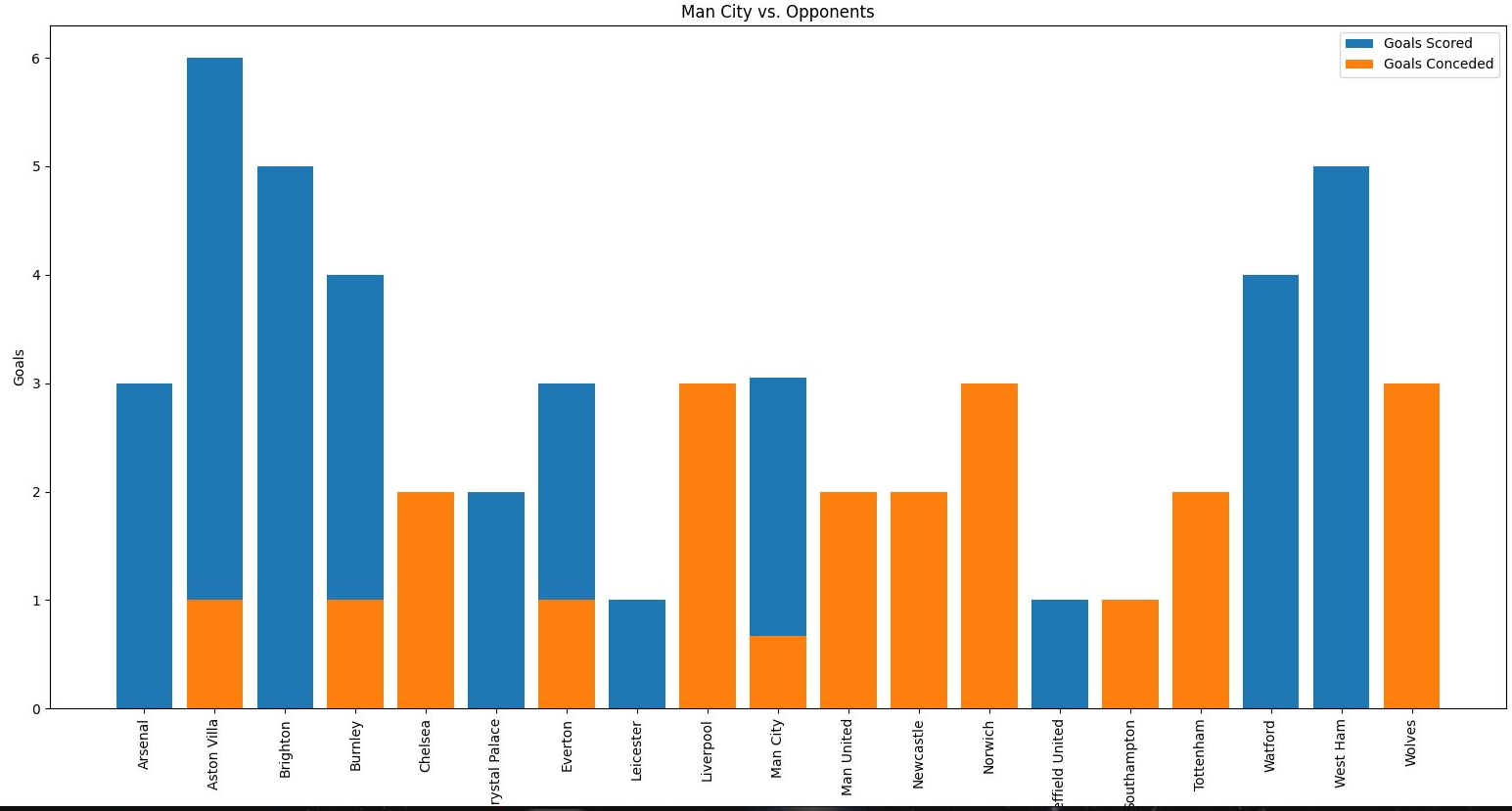
Extract and select relevant features from the cleaned and preprocessed data. Create new features that may improve the predictive power of the model, such as goal difference, home or away advantage, and form. Use domain knowledge to determine which features are most important for the prediction task.

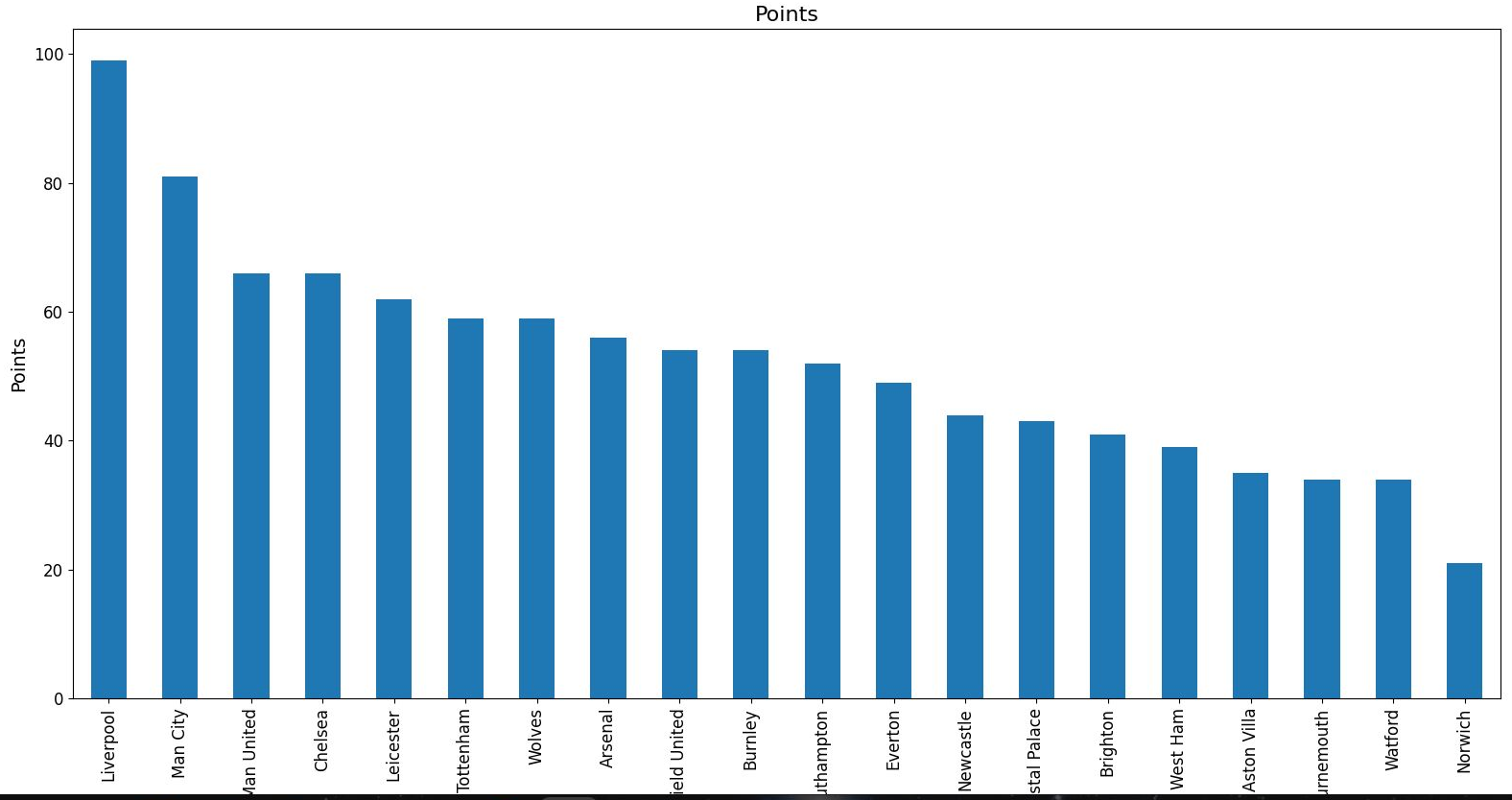
EXPLORATORY DATA ANALYSIS

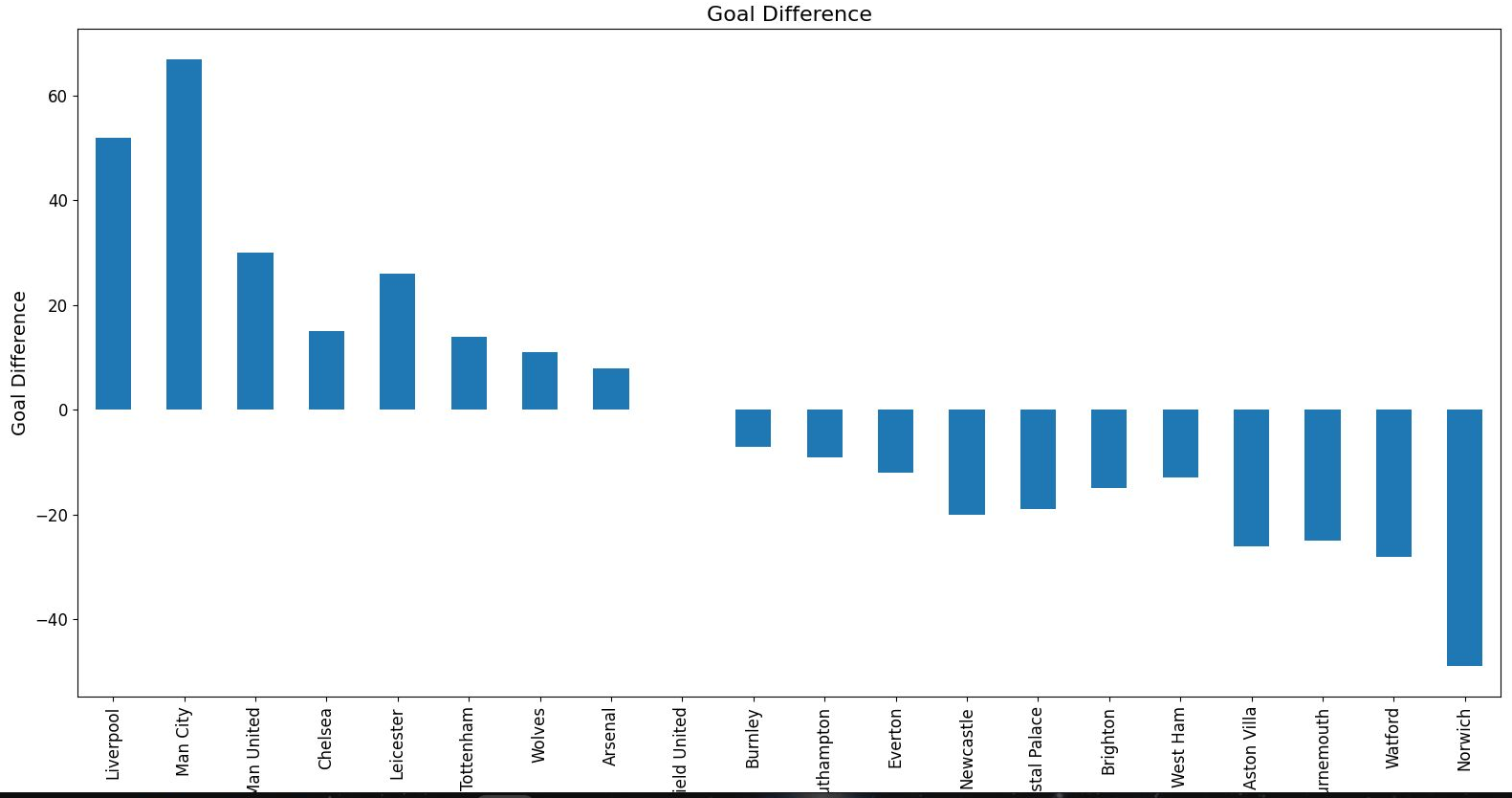
We plot different plots for thorough team wise analysis for different seasons.

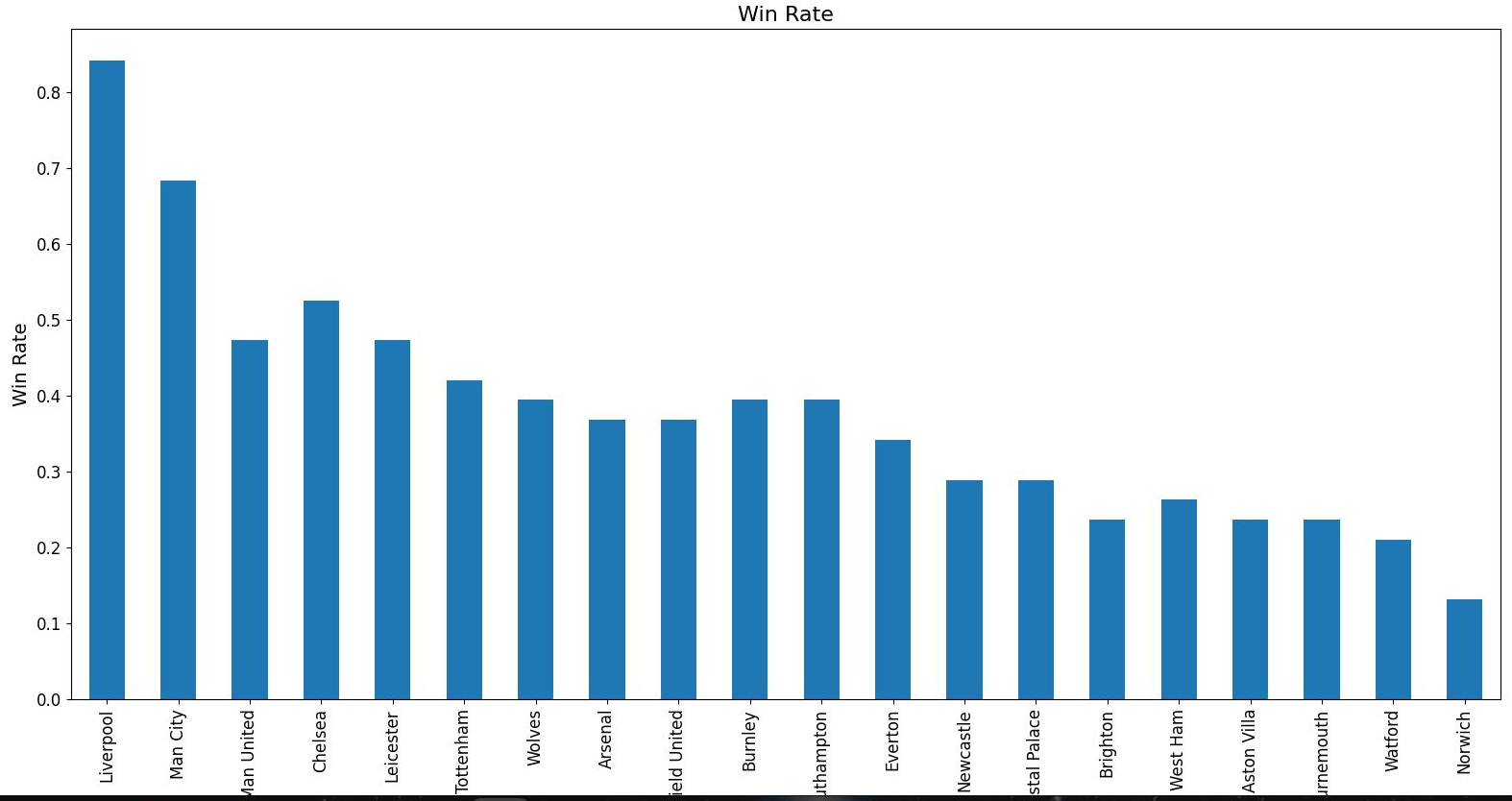
We have attached some examples of all types of graphs we used:-


Activities
matplotlib
May 7 20:38
Figure 1
x=55.25 y=10.5
Man City
Man United
Liverpool
Tottenham
Chelsea
Leicester
Arsenal
West Ham
Aston Villa
Leeds
Everton
Brighton
Burnley
Southampton
Crystal Palace
Fulham
West Brom
Sheffield United
Goal Difference
Points
40
20
0
−20
−40
20
30
40
50
60
70
80


Man City vs. Opponents
Goals Scored
Goals Conceded
Goals
0
1
2
3
4
5
6
Arsenal
Aston Villa
Brighton
Burnley
Chelsea
Everton
Leicester
Liverpool
Man City
Man United
Newcastle
Norwich
Tottenham
Watford
West Ham
Wolves


Points
Points
100
80
60
40
20
0
Liverpool
Man City
Man United
Chelsea
Leicester
Tottenham
Wolves
Arsenal
Burnley
Everton
Newcastle
Brighton
West Ham
Aston Villa
Watford
Norwich

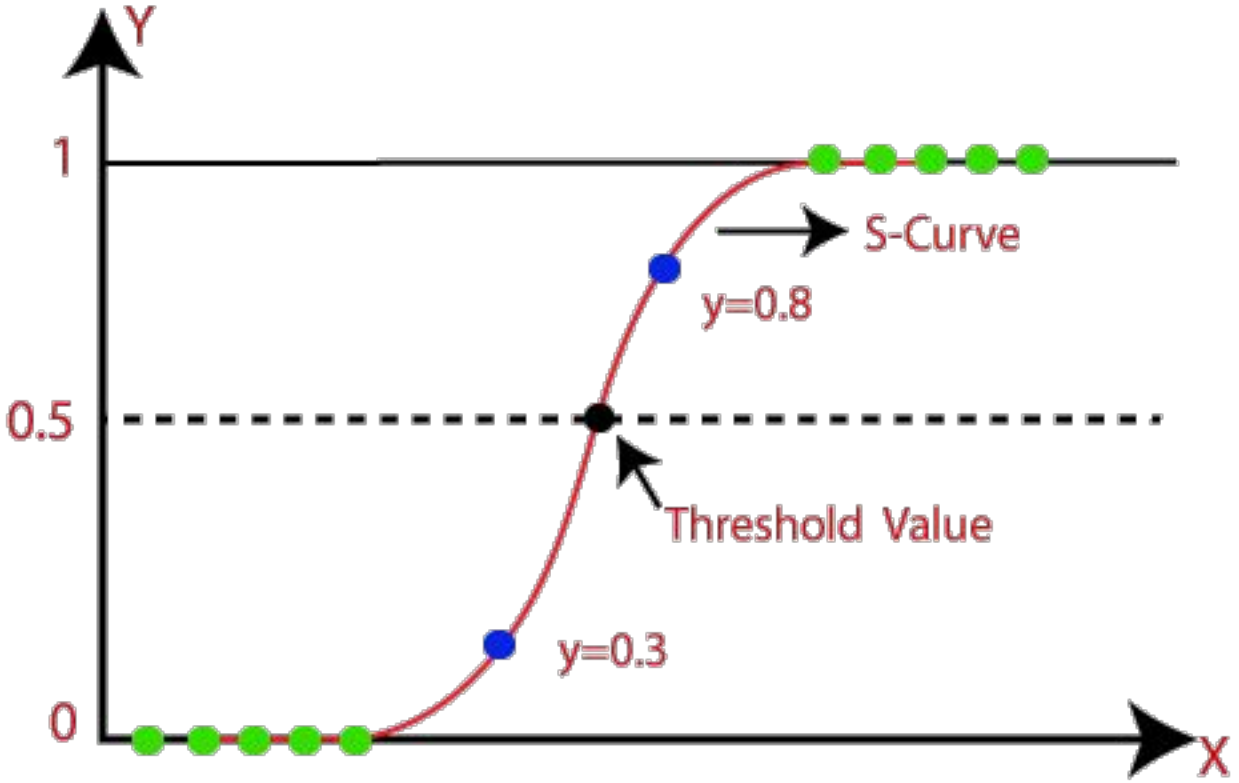

Goal Difference
Goal Difference
60
40
20
0
−20
−40
Liverpool
Man City
Man United
Chelsea
Leicester
Tottenham
Wolves
Arsenal
Burnley
Everton
Newcastle
Brighton
West Ham
Aston Villa
Watford
Norwich


Win Rate
Win Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Liverpool
Man City
Man United
Chelsea
Leicester
Tottenham
Wolves
Arsenal
Burnley
Everton
Newcastle
Brighton
West Ham
Aston Villa
Watford
Norwich

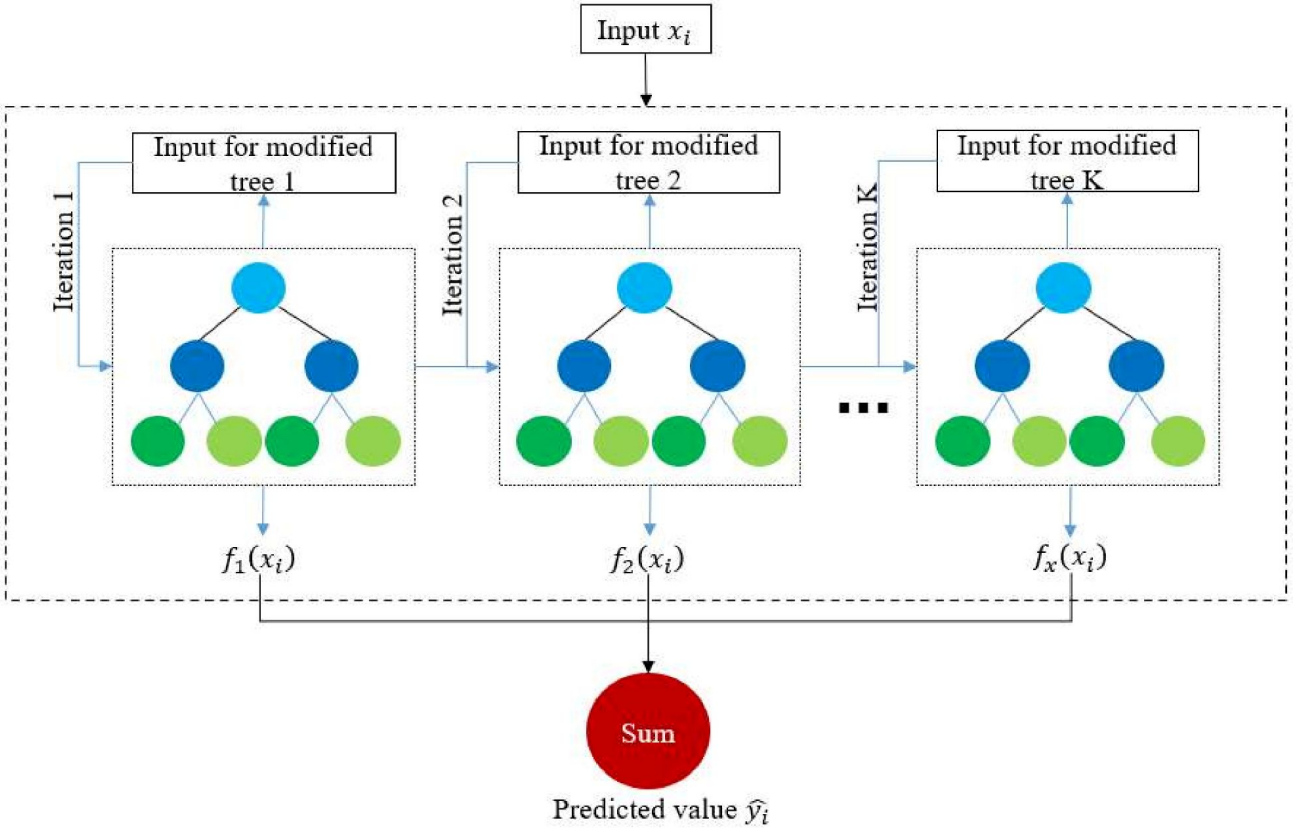
MODELS WE USED

1. Logistic Regression
2. XGBoost
3. Support Vector Machines

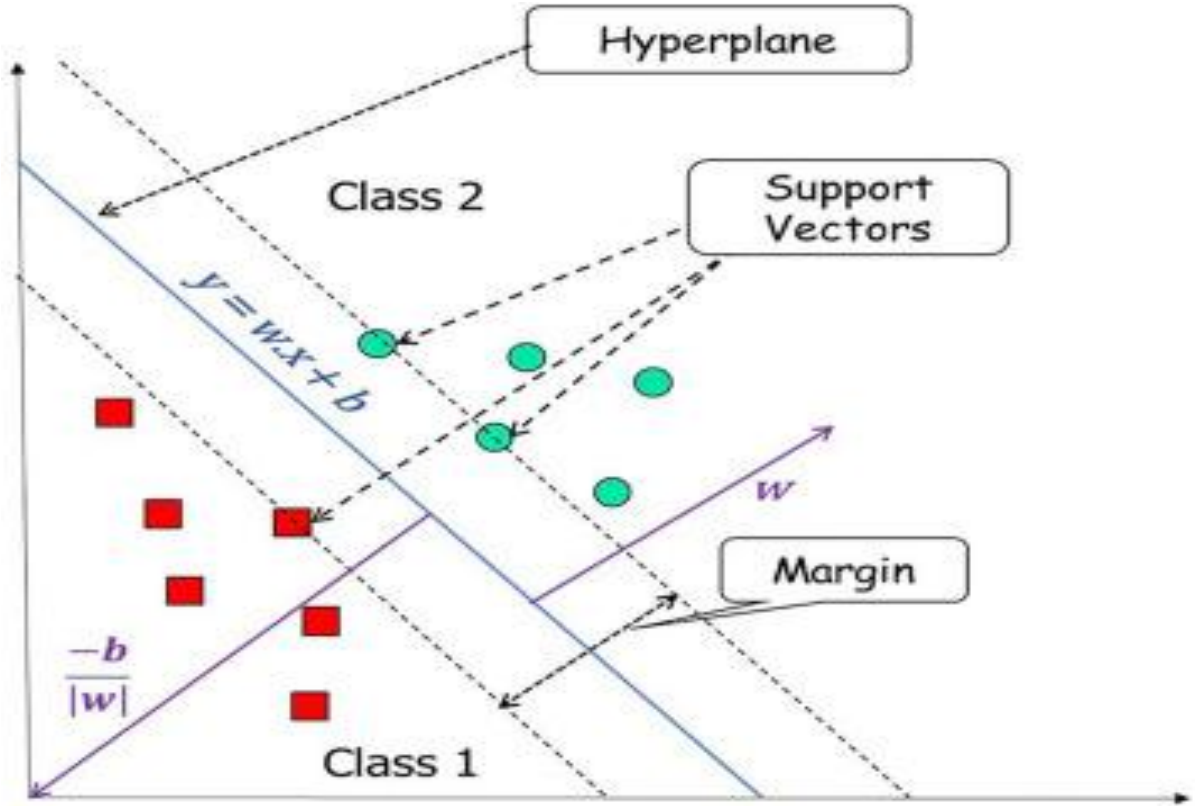
Logistic Regression Classifier



XGBoost classification (Extreme gradient boosting)



Support Vector Machines



Model Enhancement

- Methods used for hyperparameter tuning:

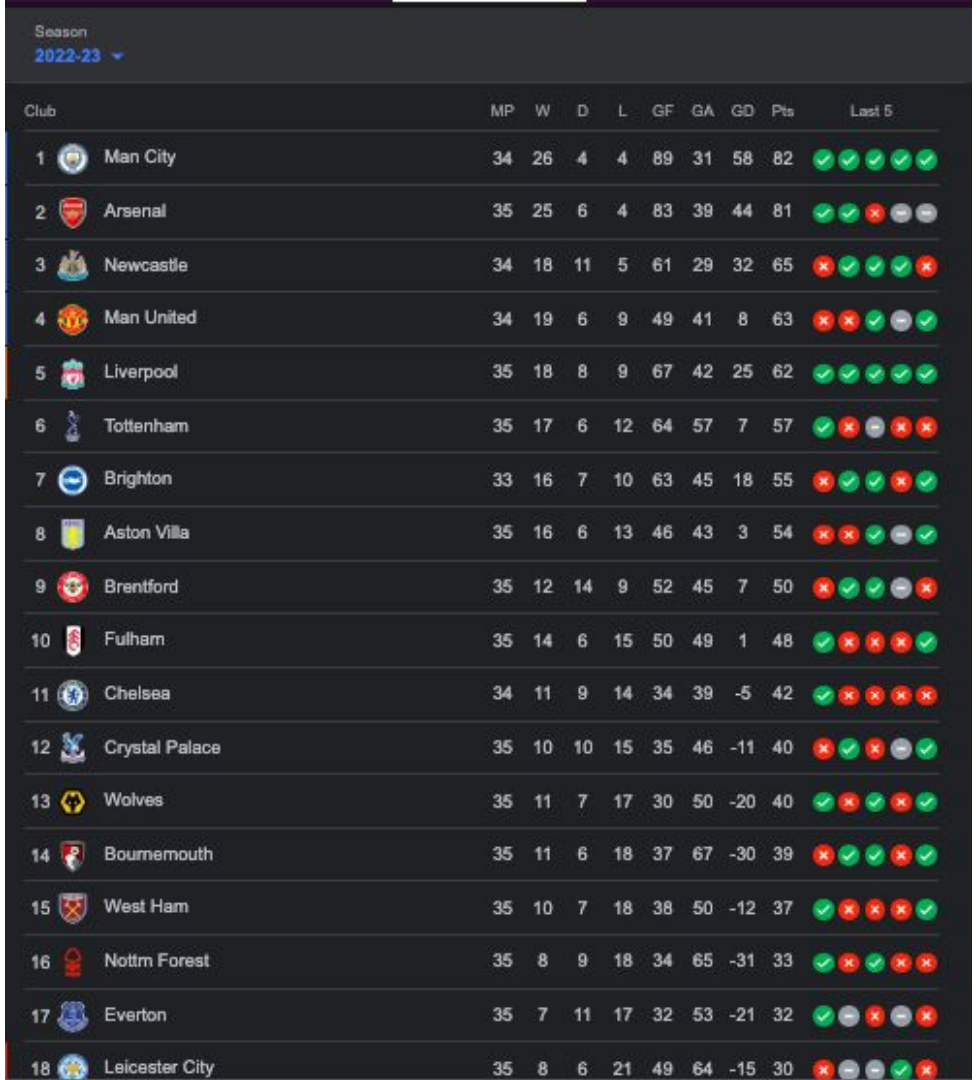
- Grid Search - iterative in the defined search space; computationally expensive; thorough.

- Random Search - randomly selects parameters; computationally efficient; may miss a better set of parameters.

Prediction and Results

```
team,spi,off.,def.,goal diff.,proj. pts.pts.,Every position,re
Man. City67 pts,93.1,3.0,0.3,62,88,,–,>99%,58%
Arsenal73 pts,84.4,2.5,0.6,48,87,,–,>99%,42%
Newcastle56 pts,82.2,2.2,0.5,32,72,,–,85%,<1%
Man. United56 pts,79.2,2.3,0.7,12,71,,–,79%,<1%
Tottenham53 pts,75.1,2.1,0.8,14,65,,–,18%,<1%
Brighton46 pts,81.3,2.3,0.6,20,62,,<1%,10%,<1%
Liverpool44 pts,83.8,2.6,0.6,24,62,,<1%,7%,–
Aston Villa47 pts,74.7,2.0,0.7,0,57,,<1%,<1%,–
Brentford43 pts,73.9,2.1,0.8,6,54,,<1%,<1%,–
Chelsea39 pts,80.8,2.0,0.5,-1,51,,<1%,<1%,–
Fulham39 pts,65.2,1.8,0.9,-6,49,,<1%,<1%,–
Crystal Palace33 pts,71.8,1.8,0.7,-9,45,,<1%,<1%,–
West Ham30 pts,71.0,1.8,0.7,-16,40,,7%,<1%,–
Wolves31 pts,64.1,1.6,0.9,-22,39,,9%,–,–
Leeds United29 pts,63.0,1.9,1.1,-20,37,,20%,–,–
Everton27 pts,60.9,1.7,1.0,-25,35,,41%,–,–
Leicester25 pts,65.8,1.8,0.9,-16,34,,49%,–,–
Southampton23 pts,61.8,1.6,0.9,-32,31,,80%,–,–
```

Actual Table as of 8th May,23.



Season
2022-23

	Club	MP	W	D	L	GF	GA	GD	Pts	Last 5
1	Man City	34	26	4	4	89	31	58	82	W W W W W
2	Arsenal	35	25	6	4	83	39	44	81	W W L D D
3	Newcastle	34	18	11	5	61	29	32	65	L W W W L
4	Man United	34	19	6	9	49	41	8	63	L L W D W
5	Liverpool	35	18	8	9	67	42	25	62	W W W W W
6	Tottenham	35	17	6	12	64	57	7	57	W L D L L
7	Brighton	33	16	7	10	63	45	18	55	L W W L W
8	Aston Villa	35	16	6	13	46	43	3	54	L L W D W
9	Brentford	35	12	14	9	52	45	7	50	L W W D L
10	Fulham	35	14	6	15	50	49	1	48	W L L L W
11	Chelsea	34	11	9	14	34	39	-5	42	W L L L L
12	Crystal Palace	35	10	10	15	35	46	-11	40	L W L D W
13	Wolves	35	11	7	17	30	50	-20	40	W L W L W
14	Bournemouth	35	11	6	18	37	67	-30	39	L W W L W
15	West Ham	35	10	7	18	38	50	-12	37	W L L L W
16	Nottm Forest	35	8	9	18	34	65	-31	33	W L W L L
17	Everton	35	7	11	17	32	53	-21	32	W D L D L
18	Leicester City	35	8	6	21	49	64	-15	30	L D D W L