PREDICTING FOOTBALL(SOCCER)
ENGLISH PREMIER LEAGUE WINNERS

By - Apoorv Upadhye
Prof -Vic Berry
Course - Boston University MET CS 664
Course Name – Artificial Intelligence

ABSTRACT

In the current sports ecosystem, a lot of times it's always about the players and managers and teams. But as the power of machines has grown, previous data can be used to predict future data and find out the parameters by pure mathematics which helps to understand which parameters are more important in the success of a team. Maybe a position that is not highly rated is the difference between the winning and losing side. So, we will be using a Machine Learning algorithm to predict the winner of the English Premier League of the current ongoing season.

Introduction

As the winning team prediction model includes a lot of reasons for a team's success, such as previous history, current form and sentiments, players quality, manager changes, and injuries, but we will focus on the previous history as playing home and away has a different impact on each team. We will try to predict the English Premier League winner for the ongoing season.

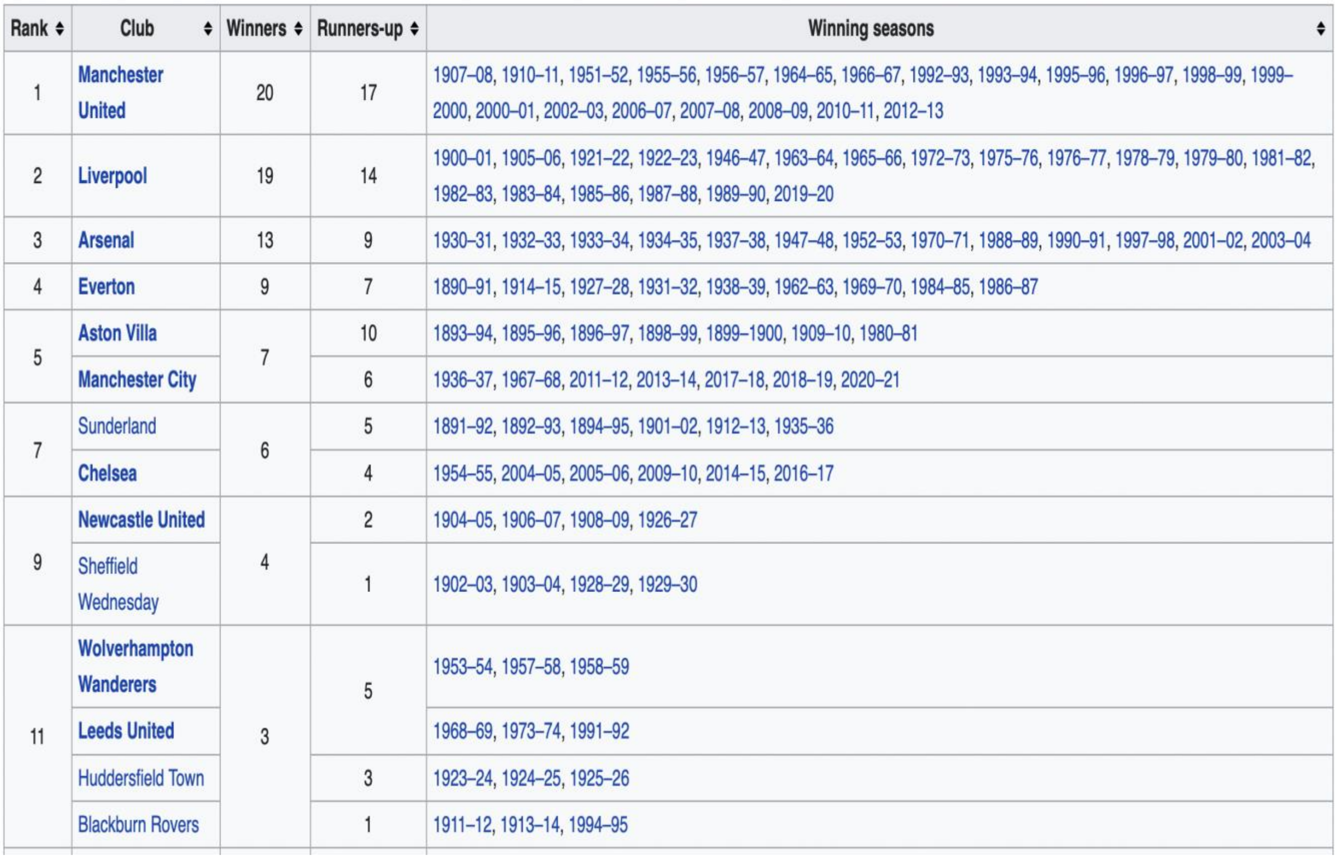
BACKGROUND

Football is a team sport played by a team of 11 players against another team of 11 players on a field.

The team has one designated goalkeeper and 10 outfield players. Outfield players are usually specialized in attacking or defending or both. A team is typically split into defenders, midfielders, and forwards, though there is no restriction on players moving anywhere on the pitch.
It is known as *soccer* in North America but is called *football* in most of the rest of the world. Its full name is Association Football.
And the World Cup is the biggest sporting event on the planet. An event that is 10 times bigger than the Olympic games in popularity and viewership.
The statistics show that certain areas are demonstrating the large potential for growth regarding the popularity of soccer: China, India, Russia, Mexico, and the U.S.A.

The most popular leagues in world football today, are:
10. MLS
9. Argentine Primera
8. Liga MX
7. Ukrainian Premier League
6. Brasileirao
5. Eredivisie
4. Serie A
3. Bundesliga
2. La Liga
1. English Premier League

The Premier League is the most-watched sports league in the world, broadcast in 212 territories to 643 million homes and a potential TV audience of 4.7 billion people. For the 2018–19 season, the average Premier League match attendance was at 38,181, second to the German Bundesliga's 43,500, while aggregated attendance across all matches is the highest of any association football league at 14,508,981.[11] Most stadium occupancies are near capacity. The Premier League ranks first in the UEFA coefficients of leagues based on performances in European competitions over the past five seasons as of 2021.[13] The English top-flight has produced the second-highest number of UEFA Champions League/European Cup titles, with five English clubs having won fourteen European trophies in total.

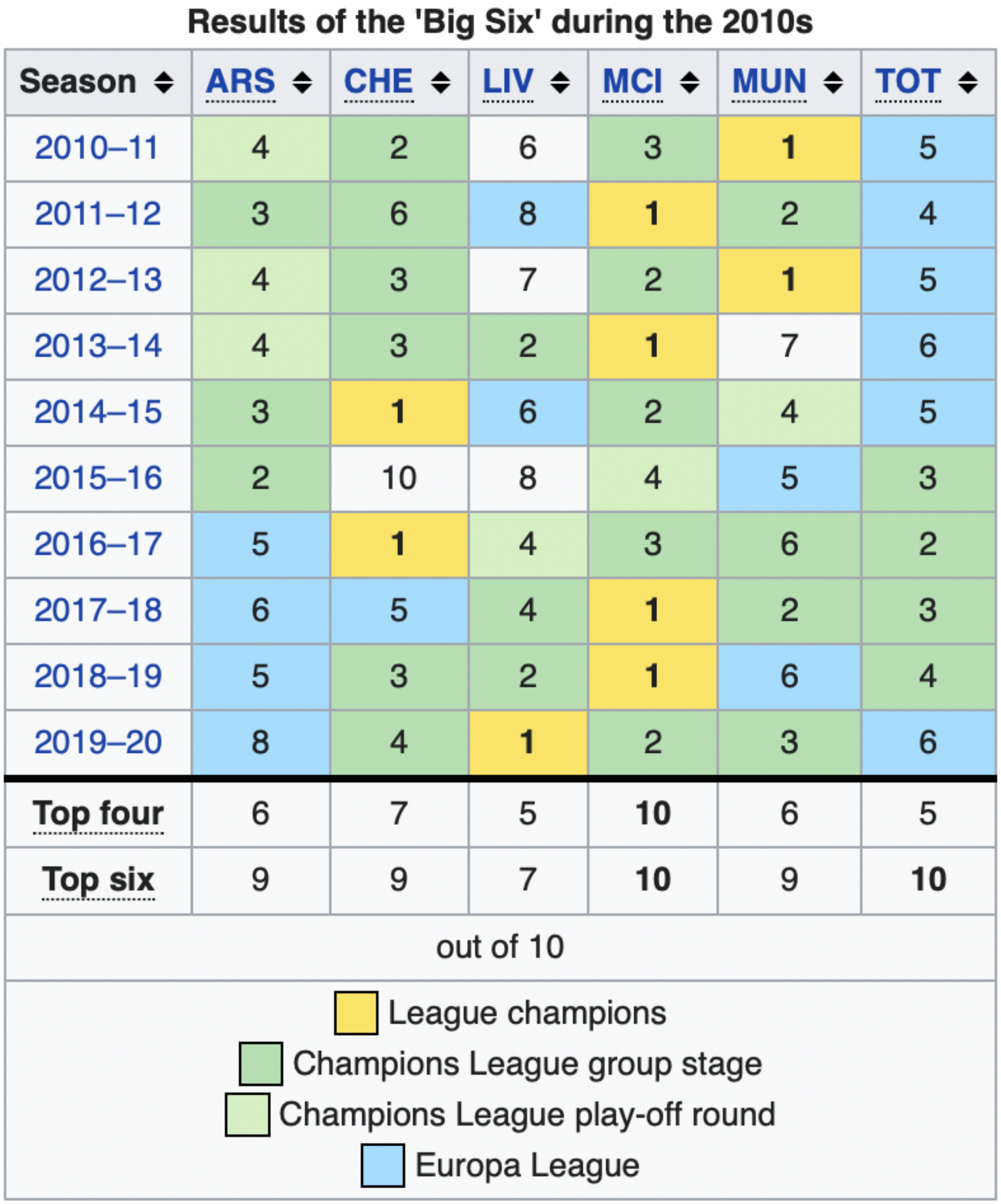


Rank	Club	Winners	Runners-up	Winning seasons
1	Manchester United	20	17	1907–08, 1910–11, 1951–52, 1955–56, 1956–57, 1964–65, 1966–67, 1992–93, 1993–94, 1995–96, 1996–97, 1998–99, 1999–2000, 2000–01, 2002–03, 2006–07, 2007–08, 2008–09, 2010–11, 2012–13
2	Liverpool	19	14	1900–01, 1905–06, 1921–22, 1922–23, 1946–47, 1963–64, 1965–66, 1972–73, 1975–76, 1976–77, 1978–79, 1979–80, 1981–82, 1982–83, 1983–84, 1985–86, 1987–88, 1989–90, 2019–20
3	Arsenal	13	9	1930–31, 1932–33, 1933–34, 1934–35, 1937–38, 1947–48, 1952–53, 1970–71, 1988–89, 1990–91, 1997–98, 2001–02, 2003–04
4	Everton	9	7	1890–91, 1914–15, 1927–28, 1931–32, 1938–39, 1962–63, 1969–70, 1984–85, 1986–87
5	Aston Villa	7	10	1893–94, 1895–96, 1896–97, 1898–99, 1899–1900, 1909–10, 1980–81
	Manchester City		6	1936–37, 1967–68, 2011–12, 2013–14, 2017–18, 2018–19, 2020–21
7	Sunderland	6	5	1891–92, 1892–93, 1894–95, 1901–02, 1912–13, 1935–36
	Chelsea		4	1954–55, 2004–05, 2005–06, 2009–10, 2014–15, 2016–17
9	Newcastle United	4	2	1904–05, 1906–07, 1908–09, 1926–27
	Sheffield Wednesday		1	1902–03, 1903–04, 1928–29, 1929–30
11	Wolverhampton Wanderers	3	5	1953–54, 1957–58, 1958–59
	Leeds United			1968–69, 1973–74, 1991–92
	Huddersfield Town		3	1923–24, 1924–25, 1925–26
	Blackburn Rovers		1	1911–12, 1913–14, 1994–95

Above is a data of total current champions in the Premier League history

The Premier League has the highest revenue of any association football league in the world, with total club revenues of €2.48 billion in 2009–10. In 2013–14, due to improved television revenues and cost controls, the Premier League clubs collectively made a net profit of more than £78 million, exceeding all other football leagues. In 2010 the Premier League was awarded the Queen's Award for Enterprise in the International Trade category for its outstanding contribution to international trade and the value it brings to English football and the United Kingdom's broadcasting industry.

The Premier League includes some of the richest football clubs in the world. Deloitte's "Football Money League" listed seven Premier League clubs in the top 20 for the 2009–10 season, and all 20 clubs were in the top 40 globally by the end of the 2013–14 season, largely because of increased broadcasting revenue. In 2019, the league generated around £3.1 billion per year in domestic and international television rights.



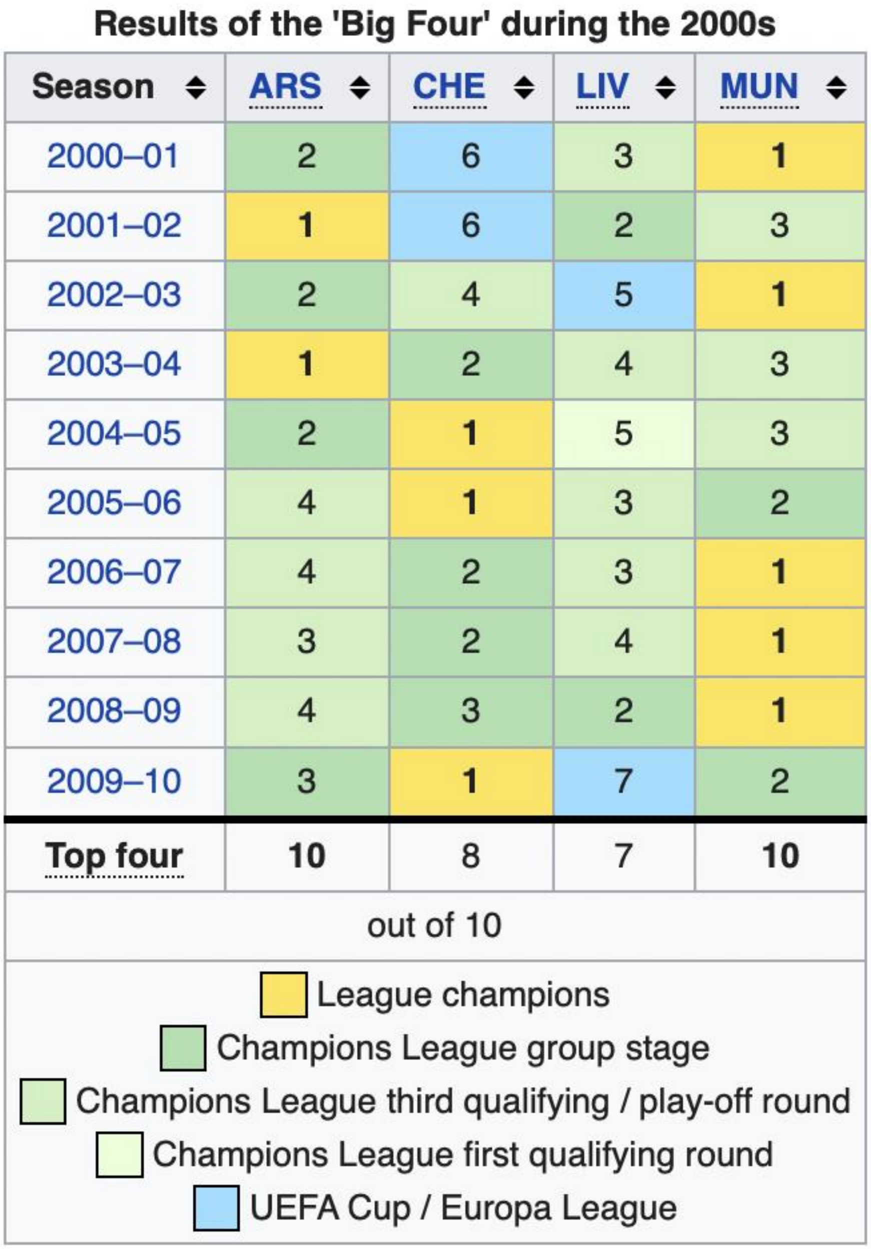
Results of the 'Big Six' during the 2010s

Season	ARS	CHE	LIV	MCI	MUN	TOT
2010–11	4	2	6	3	**1**	5
2011–12	3	6	8	**1**	2	4
2012–13	4	3	7	2	**1**	5
2013–14	4	3	2	**1**	7	6
2014–15	3	**1**	6	2	4	5
2015–16	2	10	8	4	5	3
2016–17	5	**1**	4	3	6	2
2017–18	6	5	4	**1**	2	3
2018–19	5	3	2	**1**	6	4
2019–20	8	4	**1**	2	3	6
Top four	6	7	5	**10**	6	5
Top six	9	9	7	**10**	9	**10**
out of 10						

🟨 League champions
🟩 Champions League group stage
🟩 Champions League play-off round
🟦 Europa League

Data set of teams that ruled the 2010-2019 years and most top podium finishes.

A combination of 6 teams had the most say of finishing in the top 4



Results of the 'Big Four' during the 2000s

Season	ARS	CHE	LIV	MUN
2000–01	2	6	3	1
2001–02	1	6	2	3
2002–03	2	4	5	1
2003–04	1	2	4	3
2004–05	2	1	5	3
2005–06	4	1	3	2
2006–07	4	2	3	1
2007–08	3	2	4	1
2008–09	4	3	2	1
2009–10	3	1	7	2
Top four	10	8	7	10
out of 10				

🟨 League champions
🟩 Champions League group stage
🟩 Champions League third qualifying / play-off round
⬜ Champions League first qualifying round
🟦 UEFA Cup / Europa League

Data set of teams that ruled the 2000-2009 years and most top podium finishes.

A combination of 4 teams had the most say of finishing in the top 4

Now the algorithm that we will be focusing on to get the prediction will be described

1. Logistic Regression

- ➔ The following algorithm is a classification algorithm, and which helps to predict the outcome as a class.
- ➔ The following algorithm is helpful for binary and linear classification problems.
- ➔ It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes.

2. Support Vector Machines

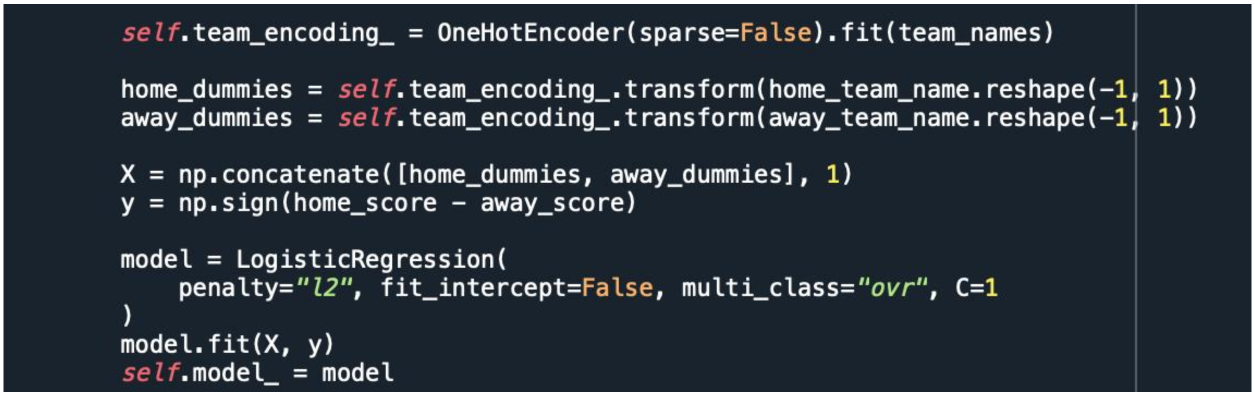
- ➔ Support Vector Machines are a set of supervised learning algorithms that can be used for regression, outliers' detection, and most importantly classification as well.
- ➔ This algorithm works wonderfully in the case of high dimensional spaces and can also be effective where several dimensions or features are greater than the number of samples.
- ➔ The kernel functions help the SVM to solve the inner product between two points in an appropriate feature space.
- ➔ Some famous kernels used to get the inner product are.
  1. Gaussian Radial Basis Function
  2. Sigmoid Kernel
  3. Bessel Function Kernel

Methodology

We need to train three models with a binary target that represents the match results. When a model is classified as draw, the match result is 1, or else it is 0

It will be represented in three different classes depending on the outcomes. We would be using a one-hot encoder

As the English Premier League contains 20 teams that are playing home and away. Individual entry can be 1 or 0 as a vector where each team is encoded with a length of 40. The top 20 entries represent the home teams and the next 20 as the away teams. So, to encode a match between two teams, must put 1 at the right place. We can follow this by using scikit-learn.

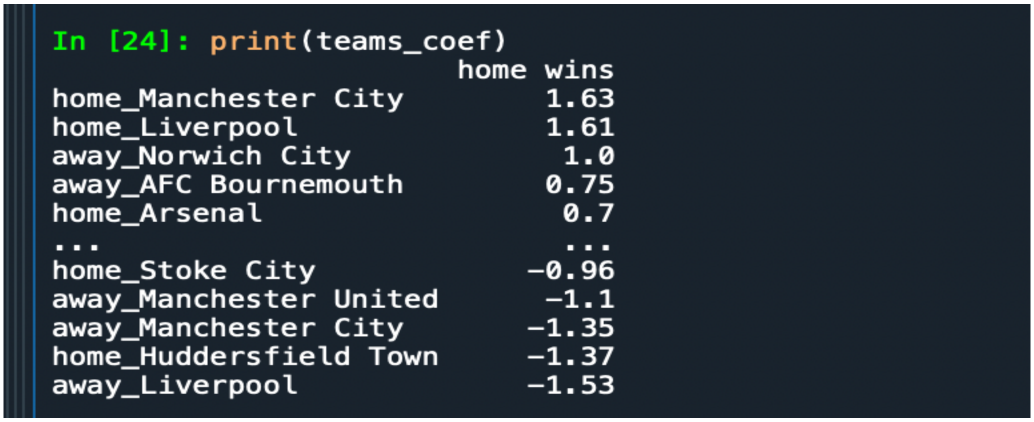


```
self.team_encoding_ = OneHotEncoder(sparse=False).fit(team_names)

home_dummies = self.team_encoding_.transform(home_team_name.reshape(-1, 1))
away_dummies = self.team_encoding_.transform(away_team_name.reshape(-1, 1))

X = np.concatenate([home_dummies, away_dummies], 1)
y = np.sign(home_score - away_score)

model = LogisticRegression(
    penalty="l2", fit_intercept=False, multi_class="ovr", C=1
)
model.fit(X, y)
self.model_ = model
```

Past matches are encoded to train the model. We can use the last three seasons. Having a home and away team and we can encode them to get our features.
For example, we have the past 200 matches, then home and away dummies are the arrays with 200 rows and 20 columns with 0 and 1. We can concatenate these two arrays to get a feature that each of the logistic models will use

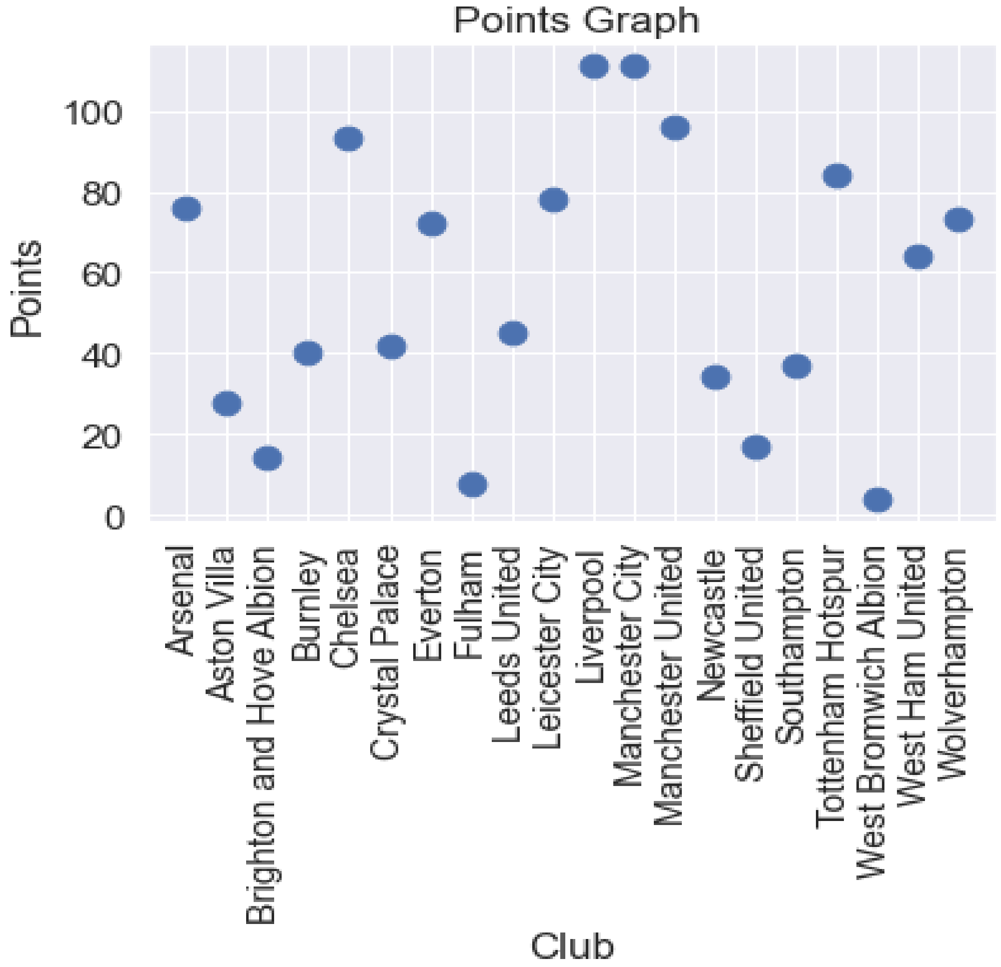
For instance

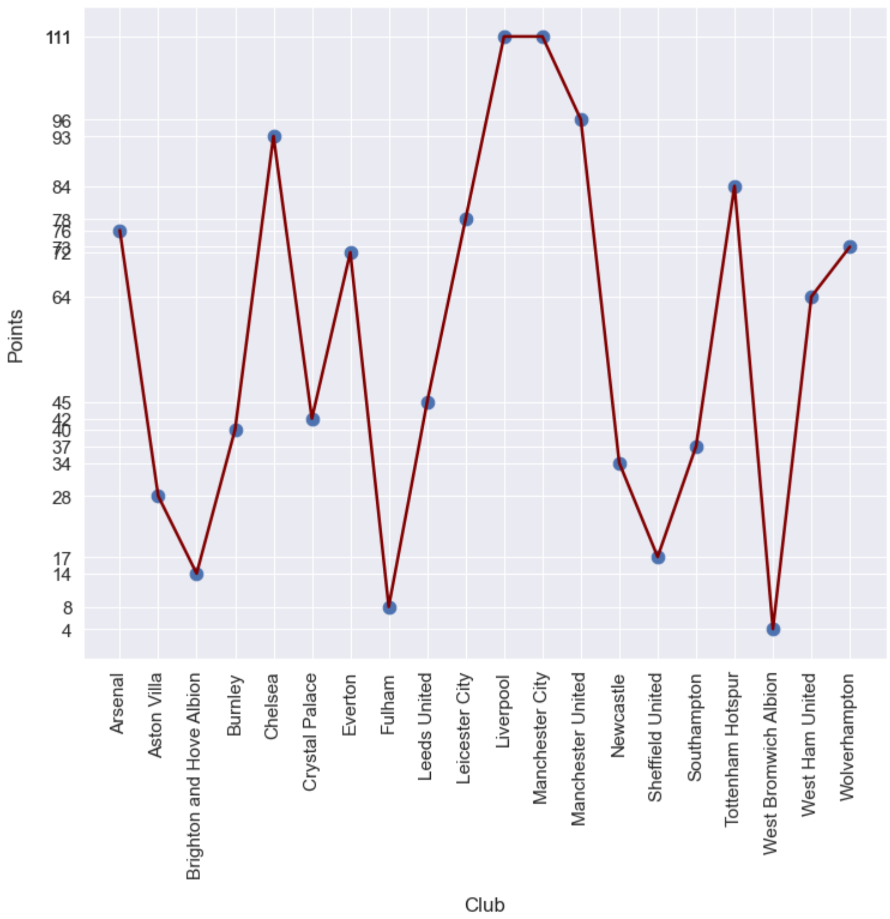
Liverpool has been good at playing away but also playing home as the coefficient is 1.61. **Liverpool while playing away hurts the chance the home team wins** as its away coefficient is very high in negative numbers.
When Stoke City plays home the coefficient
is also negative (-0.96) meaning that they have a negative impact
on their chance of winning.
After getting the following data and fitting it in logistic regression, we predict the future outcomes between two teams, one playing home and away.
After fitting all the teams in the season 2021/22 season in the following prediction Logistic Algorithm. We loop the whole list with matches played by the club week wise



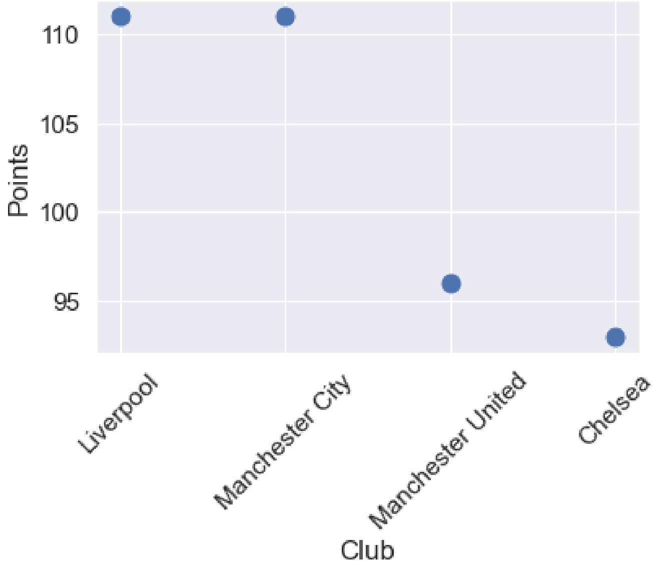
```
In [24]: print(teams_coef)
                             home wins
home_Manchester City              1.63
home_Liverpool                    1.61
away_Norwich City                  1.0
away_AFC Bournemouth              0.75
home_Arsenal                       0.7
...                                ...
home_Stoke City                  -0.96
away_Manchester United            -1.1
away_Manchester City             -1.35
home_Huddersfield Town           -1.37
away_Liverpool                   -1.53
```

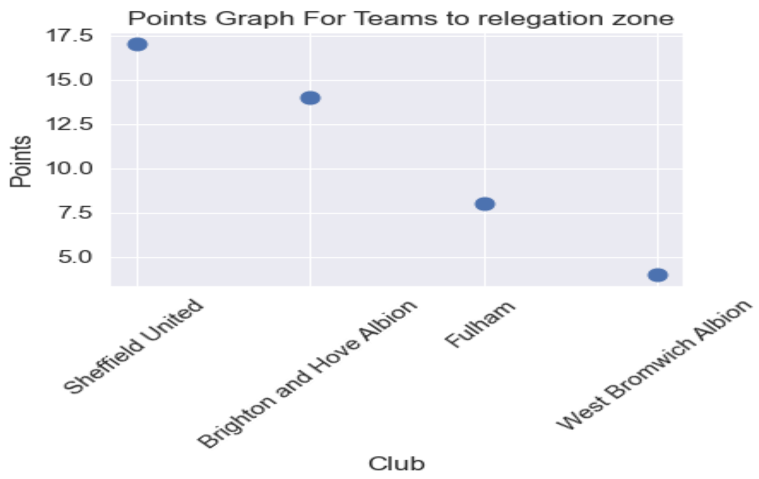
RESULTS

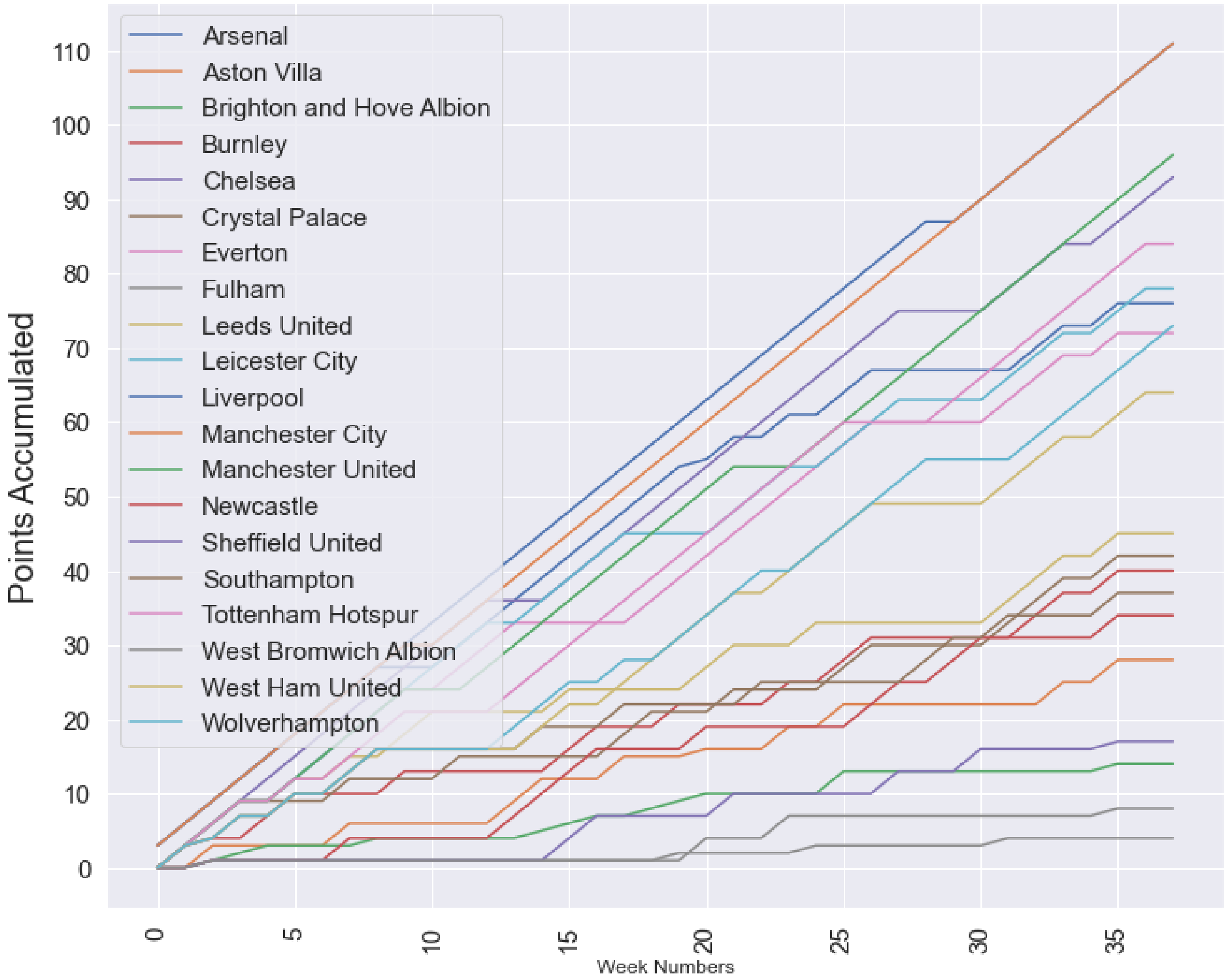

Points Graph
Points
100
80
60
40
20
0
Arsenal
Aston Villa
Brighton and Hove Albion
Burnley
Chelsea
Crystal Palace
Everton
Fulham
Leeds United
Leicester City
Liverpool
Manchester City
Manchester United
Newcastle
Sheffield United
Southampton
Tottenham Hotspur
West Bromwich Albion
West Ham United
Wolverhampton
Club



Points Graph For Teams going ahead in Champions League







The following chart is a line chart, showing every club’s week wise points accumulated

Conclusion

After applying Logistic Regression and comparing data with Support Vector Machines

1. Manchester City and Liverpool have the best odds to win the English Premier League
2. With City and Liverpool, Chelsea and Manchester United come close and qualifies for the UEFA Champions League
3. Meanwhile, the bottom 3 clubs are relegated every year to be sent to lower leagues and now three teams are promoted.
4. Expected clubs to be relegated are Fulham, Brighton and Hove Albion, and West Bromwich Albion

Finally, as we have predicted our winners, maybe in the future we can use different features like injuries or managers changes to get a better outcome on the results but as injuries remain unpredictable but maybe a dataset about the injuries and new players' or managers signing can change an outlook of the squad.

REFERENCES

1. *226 countries can't be wrong | Bleacher Report | Latest ...,* https://bleacherreport.com/articles/82112-226-countries-cant-be-wrong.12
2. *Logistic Regression - an overview | ScienceDirect Topics,* https://www.sciencedirect.com/topics/computer-science/logistic-regression.12