Football Match Prediction Machine Learning Model

Yuvraj Tilotia
MAT501 – Applied Mathematics and Artificial Intelligence
2200518@uad.ac.uk

Abstract - A random forest classifier (RFC) is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [1]. This model is helpful when you have a large amount of data, thus protecting from over or underfitting. We have already seen machine learning models which use the team strengths to determine the result of a football match. This model aims to determine the same but only using the data like the time of day when the match is played, the round of matches when the match is played, the day of the week, and the date the match is played on. This report provides an overview of how the data was acquired, the methodology used for scraping the data, and the training model decision-making process thus explaining more about RFC and the conclusions drawn from the results.

Keywords – RFC (Random Forest Classifier), Football Match prediction, Methodologies.

1. INTRODUCTION

As one of the most popular sports on the planet, football has always been followed very closely by many people. In recent years, new types of data have been collected for many games in various countries, such as play-by-play data including information on each shot, passes made in a match, possession statistics, breaks in play, Expected Goals(xG), etc. Football matches can be difficult to predict, with surprises often popping up. It is an interesting example as matches have fixed lengths. It also possesses a single type of scoring event: goals that can happen an infinite number of times during a match. The possible outcomes for a team taking part in a football match are win, lose or draw. It can therefore seem quite straightforward to predict the outcome of a game.

1.a) Existing Models

The traditional predictive methods have simply used match results to evaluate team performance and build statistical models to predict the results of future games. They utilize the vast amount of data that is generally available on official websites. This includes metrics as follows – [2]

1. General Statistics -> Total Shots, Shots on Target, Chances Created, Corners, Shots Inside the Box, etc.
2. Individual Player Statistics -> Aerial Duels, Clearances, Interceptions, Tackles, Accurate Passes, Crosses, etc.
3. Unique Statistics -> xG Data, Passing Maps, Pressing by PPDA, Average Player Positions, Heat Maps, etc.

Fig 1.1 - Individual Player Statistics

1.b) Objectives

his project aims to predict the result of football matches by utilizing a unique set of datasets. The dataset includes information about the time of day the match was played, the round of matches that are being played, the day of the week, and the date the match is played on. We would be using data from the 2017-18 season onwards until 01/01/2022 as the training data and then try to predict the match results of the remainder of the 2021-22 season. We would also try to see on how many occasions the model guessed the right result using some machine learning metrics.

To generate predictions, there are some objectives that we need to fulfill: Firstly, we need to find good-quality data and sanitize it to be used in our models. To do so, we will need to find suitable data sources. This will allow us to have access to a more unique set of statistics to use, compared to most of the past research that has made the predictions using team strength and opponent strength.

An important part of this project will be to build a suitable Machine Learning training and testing pipeline to be able to test by using the dataset. Finally, the model will be assessed against the real match result, and thus provide an accuracy score to measure the performance of our model.

2. SYSTEM ARCHITECTURE AND DESIGN PROCESS

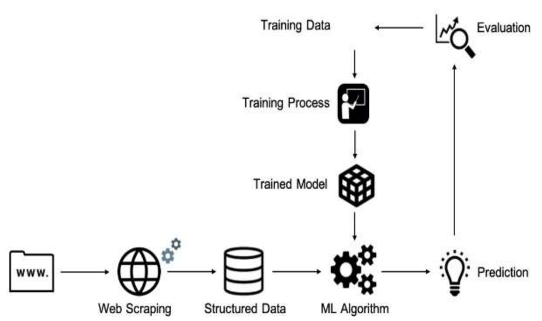

Fig 2.1 – System Architecture

2.a) Data Scraping and Modelling

The overall architecture can be divided into 2 parts – Data Gathering and Training the Model. In the first part i.e., Data Gathering, we web scrape data from a website called "www.fbref.com" [3] which has been recording English Premier League data from the day the league was formed. This data includes data like the Time of the Match, Round of Matches, Day of the Week, and Date of the Match which we will be used to predict the football match result. The data also includes traditional data like xG, Total Shots, Shots on Target, Distance Covered, Free Kicks, etc. which has been the go-to data for any football match result prediction. The process includes using a python library called "BeautifulSoup" which is used to scrape data from an HTML page. First, we use an initial webpage from where we can scrape data. On the "FBREF" website the information that we require is stored under the table class called "stats_table". By using the BeautifulSoup library, we can automatically make the scraping loop move to previous seasons and their respective stats_table.

```
standings_url = "https://fbref.com/en/comps/9/2021-2022/2021-2022-Premier-League-Stats"

for year in years:
    data = requests.get(standings_url)
    soup = BeautifulSoup(data.text)
    standings_table = soup.select('table.stats_table')[0]

    links = [l.get("href") for l in standings_table.find_all('a')]
    links = [l for l in links if '/squads/' in l]
    team_urls = [f"https://fbref.com{l}" for l in links]

    previous_season = soup.select("a.prev")[0].get("href")
    standings_url = f"https://fbref.com{previous_season}"
```

Fig 2.2 – Scraping

The next step is to make a csv file to store the first webpage data and then append the new data from the next web pages into the same file. To help with the careful scraping of data from the website, after each instance of stats_table scraping, the program sleeps for 3 seconds. This helps in protecting the scrape request from getting denied by servers.

```
        team_data["Season"] = year
        team_data["Team"] = team_name
        all_matches.append(team_data)
        time.sleep(3)

match_df = pd.concat(all_matches)
match_df.columns = [c.lower() for c in match_df.columns]
match_df.to_csv("matches.csv")
```

Fig 2.3 – Saving data to a csv file

2.b) Random Forest Classifier

In this model, we used Random Forest Classifier (RFC) as the training machine learning algorithm.

In definition, a random forest is a classifier consisting of a collection of tree-structured classifiers {h(x, Θk), k=1, ...} where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [4].

Random forests are an effective tool in prediction. Because of the Law of Large Numbers, they do not overfit. Injecting the right kind of randomness makes them accurate classifiers and regressors. Furthermore, the framework in terms of the strength of the individual predictors and their correlations gives insight into the ability of the random forest to predict. Using out-of-bag estimation makes concrete the otherwise theoretical values of strength and correlation.

Features ->

1. To improve accuracy, the RFC has randomness injected, which minimizes the correlation ρ while also maintaining strength.
2. Another added advantage is that it's relatively robust to outliers and noise. This is helpful when some Football matches give unexpected results.
3. While it is simple and easily parallelized, it is fast too. When dealing with large datasets, faster computation is one of the most desirable features of any machine learning algorithm.

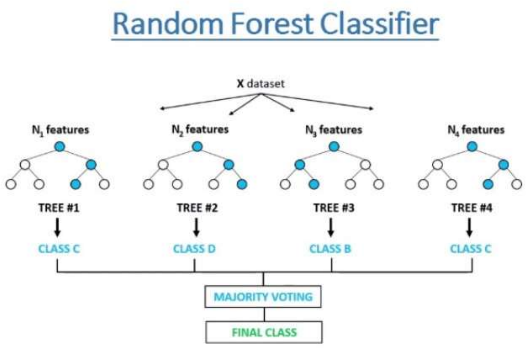

Fig 2.4 – Random Forest Classifier

2.c) Training the model

The second part of the working model is Training. The process involved using the "sklearn" library in python which has all sub-libraries for our model. The first step is to load the data and then there is a requirement to refine the data according to our needs. This involves removing redundancies, replacing object type datatypes with numeric form, and assigning categorical values to data like Time of Match, Round of Matches, Day of the Week, and Date of the Match which will be used to train the data with a Classifier machine learning algorithm. The next step is to determine the training and test data. The training data includes all matches played from the 2017-18 season onwards until 31/12/2021. The test data includes all matches played after 01/01/2022. This is also the date on which we will measure our accuracy score. The next step is to train the model on the training dataset and then test it on the test dataset.

```
matches = pd.read_csv("matches.csv", index_col=0)
del matches["comp"]
del matches["notes"]
matches["date"] = pd.to_datetime(matches["date"])
matches["target"] = (matches["result"] == "W").astype("int")
matches["venue_code"] = matches["venue"].astype("category").cat.codes
matches["opp_code"] = matches["opponent"].astype("category").cat.codes
matches["hour"] = matches["time"].str.replace(":.+", "", regex=True).astype("int")
matches["day_code"] = matches["date"].dt.dayofweek

rf = RandomForestClassifier(n_estimators=1000000, min_samples_split=10, random_state=100)
train = matches[matches["date"] < '2022-01-01']
test = matches[matches["date"] > '2021-01-01']
predictors = ["venue_code", "opp_code", "hour", "day_code"]
rf.fit(train[predictors], train["target"])
preds = rf.predict(test[predictors])
acc = accuracy_score(test["target"], preds)
combined = pd.DataFrame(dict(actual=test["target"], predicted=preds))
acc
```

Fig 2.5 – Training the model

3. RESULT AND CONCLUSION

The model accuracy of prediction for the remainder of the 2021-22 season turned out to be 69.4%. The results are very positive as we are only using data like the Time of the Match, Round of Matches, Day of the Week, and Date of the Match.

```
rf.fit(train[predictors], train["target"])
preds = rf.predict(test[predictors])
acc = accuracy_score(test["target"], preds)
combined = pd.DataFrame(dict(actual=test["target"], predicted=preds))
acc

0.6940298507462687
```

Fig 3.1 – Accuracy

The F1 score of the model tells the accuracy by combining the precision and recall scores of the model.

```
from sklearn.metrics import f1_score
f1_score(test["target"], preds)
```

```
0.5472392638036809
```

Fig 3.2 – F1 score

We also found the feature importance of the model. The results were – 1. Time of Match (59%), 2. Round of Matches (20.2%), 3. Day of the Week (15.5%), 4. Date of the Match (5.3%).

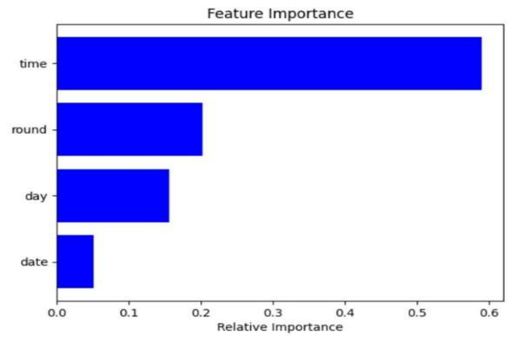

Fig 3.2 – Feature Importance

We also got the classification report from the model. The results are below –

```
              precision    recall  f1-score   support

           0       0.84      0.71      0.77       862
           1       0.47      0.65      0.55       344

    accuracy                           0.69      1206
   macro avg       0.65      0.68      0.66      1206
weighted avg       0.73      0.69      0.71      1206
```

Fig 3.3 – Classification Report

The confusion matrix of the model which told us the number of correct predictions made by the model was also formed. The results are as follows –

```
array([[614, 121],
       [248, 223]], dtype=int64)
```

Fig 3.4 – Confusion Matrix

In conclusion, we first developed a web scraping system, which retains the data of football matches from the 2017-18 season onwards to the 2021-22 season. Then, we employed Random Forest Classifier prediction technique. In continuation, we carried out training and testing which was aimed to show the efficiency and effectiveness of the proposed system. Finally, we used the accuracy score, and F1 score to measure the prediction accuracy of the model and other metrics like the Classification Report, Confusion Matrix, and F1 score to see how the model is working.

4. FUTURE WORK

The current model uses a unique dataset to check their involvement in predicting a football match result. In the future, using such unique datasets like social media tracking to measure how much influence does a positive or negative social media campaign makes over the result of a football match. This work can also be an entry to understand the importance of the group psychology of the club involved in the match and how it affects the match results.

5. REFERENCES

[1] Random Forest Classifier (no date). Available at:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn-ensemble-randomforestclassifier (Accessed: 17 November 2021).

[2] Traditional data used to make prediction (2020). Available at:

https://medium.com/@drsmukherjee/analyzing-a-football-match-using-data-where-is-the-data-what-is-the-data-and-what-can-we-177fac9095fb (Accessed: 28 November 2021).

[3] Premier League data for 2021-22 season (no date). Available at:

https://fbref.com/en/comps/9/2021-2022/2021-2022-Premier-League-Stats (Accessed: 7 December 2021).

[4] Research paper on Random Forest (2001). Available at:

https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf (Accessed: 28 December 2021).

APPENDIX: DATASET EXAMPLES

	date	time	comp	round	day	venue	result	gf	ga	opponent	xg	xga	poss	attendance	captain
1	11/08/2017	19:45	Premier League	Matchweek 1	Fri	Home	W	4	3	Leicester City	2.5	1.5	68	59387	Petr ÄŒech
0	11/08/2017	19:45	Premier League	Matchweek 1	Fri	Away	L	3	4	Arsenal	1.5	2.5	32	59387	Wes Morgan
0	12/08/2017	17:30	Premier League	Matchweek 1	Sat	Away	W	2	0	Brighton	1.9	0.3	77	30415	Vincent Kompany
0	12/08/2017	12:30	Premier League	Matchweek 1	Sat	Away	D	3	3	Watford	2.6	2.1	54	20407	Jordan Henderson
1	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Home	L	2	3	Burnley	1.5	0.6	62	41616	Gary Cahill
0	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Away	W	3	2	Chelsea	0.6	1.5	38	41616	Tom Heaton
2	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Home	W	1	0	Stoke City	0.6	0.4	60	39045	Phil Jagielka
0	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Home	L	0	3	Huddersfield	1.1	1.5	56	25448	Jason Puncheon
0	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Away	L	0	1	West Brom	0.5	1.3	69	25011	Simon Francis
0	12/08/2017	12:30	Premier League	Matchweek 1	Sat	Home	D	3	3	Liverpool	2.1	2.6	46	20407	Heurelho Gomes
0	12/08/2017	17:30	Premier League	Matchweek 1	Sat	Home	L	0	2	Manchester City	0.3	1.9	23	30415	Bruno
0	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Away	W	3	0	Crystal Palace	1.5	1.1	45	25448	Tommy Smith
0	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Home	D	0	0	Swansea City	2	0.3	60	31447	Steven Davis
0	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Away	D	0	0	Southampton	0.3	2	40	31447	Leon Britton
0	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Away	L	0	1	Everton	0.4	0.6	40	39045	Ryan Shawcross
0	12/08/2017	15:00	Premier League	Matchweek 1	Sat	Home	W	1	0	Bournemouth	1.3	0.5	31	25011	Jake Livermore
1	13/08/2017	16:00	Premier League	Matchweek 1	Sun	Home	W	4	0	West Ham	2.1	0.5	55	74928	Antonio Valencia
0	13/08/2017	13:30	Premier League	Matchweek 1	Sun	Away	W	2	0	Newcastle Utd	2.5	0.8	72	52077	Hugo Lloris
0	13/08/2017	13:30	Premier League	Matchweek 1	Sun	Home	L	0	2	Tottenham	0.8	2.5	28	52077	Jonjo Shelvey
0	13/08/2017	16:00	Premier League	Matchweek 1	Sun	Away	L	0	4	Manchester Utd	0.5	2.1	45	74928	Mark Noble
2	19/08/2017	12:30	Premier League	Matchweek 2	Sat	Away	W	4	0	Swansea City	3	0.4	58	20862	Antonio Valencia
2	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Home	W	1	0	Crystal Palace	2.5	0.7	71	53138	Jordan Henderson
2	19/08/2017	17:30	Premier League	Matchweek 2	Sat	Away	L	0	1	Stoke City	1.5	0.7	76	29459	Petr ÄŒech
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Home	L	0	1	West Brom	1.3	0.9	66	19619	Tom Heaton
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Home	W	2	0	Brighton	2	0.2	45	31902	Wes Morgan
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Away	L	0	1	Liverpool	0.7	2.5	29	53138	Jason Puncheon
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Home	L	0	2	Watford	1	2.4	55	10501	Andrew Surman
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Away	L	2	3	Southampton	2	2.1	34	31424	Mark Noble
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Away	W	2	0	Bournemouth	2.4	1	45	10501	Heurelho Gomes
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Away	L	0	2	Leicester City	0.2	2	55	31902	Bruno
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Home	W	3	2	West Ham	2.1	2	66	31424	Steven Davis
1	19/08/2017	12:30	Premier League	Matchweek 2	Sat	Home	L	0	4	Manchester Utd	0.4	3	42	20862	Federico FernÃ¡ndez
1	19/08/2017	17:30	Premier League	Matchweek 2	Sat	Home	W	1	0	Arsenal	0.7	1.5	24	29459	Ryan Shawcross
1	19/08/2017	15:00	Premier League	Matchweek 2	Sat	Away	W	1	0	Burnley	0.9	1.3	34	19619	Jake Livermore
1	20/08/2017	16:00	Premier League	Matchweek 2	Sun	Home	L	1	2	Chelsea	0.7	0.7	68	73587	Hugo Lloris
2	20/08/2017	16:00	Premier League	Matchweek 2	Sun	Away	W	2	1	Tottenham	0.7	0.7	32	73587	CÃ©sar Azpilicueta
1	20/08/2017	13:30	Premier League	Matchweek 2	Sun	Away	L	0	1	Huddersfield	0.7	0.3	48	24128	Jamaal Lascelles
1	20/08/2017	13:30	Premier League	Matchweek 2	Sun	Home	W	1	0	Newcastle Utd	0.3	0.7	52	24128	Tommy Smith
1	21/08/2017	20:00	Premier League	Matchweek 2	Mon	Home	D	1	1	Everton	1.1	0.6	64	49108	Vincent Kompany
4	21/08/2017	20:00	Premier League	Matchweek 2	Mon	Away	D	1	1	Manchester City	0.6	1.1	36	49108	Phil Jagielka
2	26/08/2017	12:30	Premier League	Matchweek 3	Sat	Away	W	2	1	Bournemouth	1.4	0.5	70	10419	Vincent Kompany
3	26/08/2017	17:30	Premier League	Matchweek 3	Sat	Home	W	2	0	Leicester City	2.8	0.9	69	75021	Antonio Valencia

Fig 1: Screen capture of Raw Data

date	time	day	round	venue	result	opponent	team
11/08/2017	19:45	Fri	Matchweek 1	Home	W	Leicester City	Arsenal
11/08/2017	19:45	Fri	Matchweek 1	Away	L	Arsenal	Leicester City
12/08/2017	17:30	Sat	Matchweek 1	Away	W	Brighton	Manchester City
12/08/2017	12:30	Sat	Matchweek 1	Away	D	Watford	Liverpool
12/08/2017	15:00	Sat	Matchweek 1	Home	L	Burnley	Chelsea
12/08/2017	15:00	Sat	Matchweek 1	Away	W	Chelsea	Burnley
12/08/2017	15:00	Sat	Matchweek 1	Home	W	Stoke City	Everton
12/08/2017	15:00	Sat	Matchweek 1	Home	L	Huddersfield	Crystal Palace
12/08/2017	15:00	Sat	Matchweek 1	Away	L	West Brom	Bournemouth
12/08/2017	12:30	Sat	Matchweek 1	Home	D	Liverpool	Watford
12/08/2017	17:30	Sat	Matchweek 1	Home	L	Manchester City	Brighton and Hove Albion
12/08/2017	15:00	Sat	Matchweek 1	Away	W	Crystal Palace	Huddersfield Town
12/08/2017	15:00	Sat	Matchweek 1	Home	D	Swansea City	Southampton
12/08/2017	15:00	Sat	Matchweek 1	Away	D	Southampton	Swansea City
12/08/2017	15:00	Sat	Matchweek 1	Away	L	Everton	Stoke City
12/08/2017	15:00	Sat	Matchweek 1	Home	W	Bournemouth	West Bromwich Albion
13/08/2017	16:00	Sun	Matchweek 1	Home	W	West Ham	Manchester United
13/08/2017	13:30	Sun	Matchweek 1	Away	W	Newcastle Utd	Tottenham Hotspur
13/08/2017	13:30	Sun	Matchweek 1	Home	L	Tottenham	Newcastle United
13/08/2017	16:00	Sun	Matchweek 1	Away	L	Manchester Utd	West Ham United
19/08/2017	12:30	Sat	Matchweek 2	Away	W	Swansea City	Manchester United
19/08/2017	15:00	Sat	Matchweek 2	Home	W	Crystal Palace	Liverpool
19/08/2017	17:30	Sat	Matchweek 2	Away	L	Stoke City	Arsenal
19/08/2017	15:00	Sat	Matchweek 2	Home	L	West Brom	Burnley
19/08/2017	15:00	Sat	Matchweek 2	Home	W	Brighton	Leicester City
19/08/2017	15:00	Sat	Matchweek 2	Away	L	Liverpool	Crystal Palace
19/08/2017	15:00	Sat	Matchweek 2	Home	L	Watford	Bournemouth
19/08/2017	15:00	Sat	Matchweek 2	Away	L	Southampton	West Ham United
19/08/2017	15:00	Sat	Matchweek 2	Away	W	Bournemouth	Watford
19/08/2017	15:00	Sat	Matchweek 2	Away	L	Leicester City	Brighton and Hove Albion
19/08/2017	15:00	Sat	Matchweek 2	Home	W	West Ham	Southampton
19/08/2017	12:30	Sat	Matchweek 2	Home	L	Manchester Utd	Swansea City
19/08/2017	17:30	Sat	Matchweek 2	Home	W	Arsenal	Stoke City
19/08/2017	15:00	Sat	Matchweek 2	Away	W	Burnley	West Bromwich Albion
20/08/2017	16:00	Sun	Matchweek 2	Home	L	Chelsea	Tottenham Hotspur
20/08/2017	16:00	Sun	Matchweek 2	Away	W	Tottenham	Chelsea
20/08/2017	13:30	Sun	Matchweek 2	Away	L	Huddersfield	Newcastle United
20/08/2017	13:30	Sun	Matchweek 2	Home	W	Newcastle Utd	Huddersfield Town
21/08/2017	20:00	Mon	Matchweek 2	Home	D	Everton	Manchester City
21/08/2017	20:00	Mon	Matchweek 2	Away	D	Manchester City	Everton
26/08/2017	12:30	Sat	Matchweek 3	Away	W	Bournemouth	Manchester City
26/08/2017	17:30	Sat	Matchweek 3	Home	W	Leicester City	Manchester United

Fig 2: Screen capture of Clean Data