

Winner Prediction of Football Match Using Machine Learning



Shailja Jadon , Aman Jain , Prathamesh Bagal , Kunal Bhatt, and Manish Rana

Abstract Over the course of this article, a simple machine learning model for the prediction of a football match winner will be discussed. An in-depth analysis and insight of this model is presented further below. And we would go about the process of building it, the relevance of the project will also be mentioned along with its business implications. Finally, the merits and flaws of the project will be discussed along with ways in which it can be improved in future.

Keywords Machine Learning · Multivariate linear regression · Football prediction · Match outcome prediction

1 Problem Description

Machine Learning has become a rather sought—after technology among young students and even industries. Machine Learning is a key solution that can answer questions of the future. Its predictive nature has appealed the masses as it removes the ambiguity from various situations where the future is unknown. Humans tend to rely on such predictions for variety of their tasks.

The prediction of the winner of a football match is a curious problem of machine learning. Here the objective is to apply certain machine learning models on existing data such that we can predict the outcome with precision. The solution of this problem

S. Jadon · A. Jain (✉) · P. Bagal · K. Bhatt · M. Rana
Thakur College of Engineering & Technology, Mumbai 400101, Maharashtra, India
e-mail: jainamanr@gmail.com

S. Jadon
e-mail: shailjajadon2001@gmail.com

P. Bagal
e-mail: prathameshbagal2908@gmail.com

K. Bhatt
e-mail: bhatakunal04@gmail.com

M. Rana
e-mail: manishrana23@gmail.com

is not just anticipated by mathematicians interested in the sport but also huge organizations that dabble in the region of betting. Even, news rooms look forward to such statistics for their audience. Thus, proving the words in the above paragraph to be true in their intent.

For this paper, we shall be discussing the problem of prediction of match outcome. We will be presenting a detailed solution in form of a project that was built using certain information about the previous seasons of the English Premier League. The given information contains various parameters such as match date, goals made etc. This dataset can be obtained from online resources.

2 Literature Survey

Machine Learning (ML) approaches have been increasingly popular for forecasting sports results over the last two decades. In this paper, the authors provide a review of studies that have used ML for predicting results in team sport, covering studies from 1996 to 2019. They have sought to answer five key research questions while extensively surveying papers in this field. This research examines which machine learning algorithms have been most often utilized in this discipline, as well as others that are beginning to emerge with promising results. Their research highlights defining characteristics of successful studies and identifies robust strategies for evaluating accuracy results in this application domain. Their study considers accuracies that have been achieved across different sports and explores the notion that results of certain team sports could be more difficult to predict than others. Finally, their study uncovers common themes of future research directions across all surveyed papers, looking for and proposing gaps and opportunities for future researchers in this domain [1].

Several efforts are targeted towards increasing the accuracy of the prediction results of the soccer match. The researchers planned numerous models via implement completely different ML algorithms. Razali et al. [2] prepared a theorem stratified model that predicts soccer results. Their model relied on the goals that each groups scored in every match. Min et al. [3] provides a dynamic system for predicting the results of football matches. This dynamic structure called the FRES system comprises of two main components: theorem supported rules and therefore the theorem network element. Therefore, the FRES methodology could be a mixture of 2 ways that job along to predict the outcomes of soccer matches.

Moreover, the FRES methodology has conjointly been introduced in-game time-series approach, that permits prediction additional sensible. Nonetheless, the FRES program needs decent professional experience so as to be controlled. Constantinou [4] has established a soccer prediction model called pi-rating to supply projections on the result of football matches, whether or not home win, draw or away winfor EPL matches throughout the 2010/2011 seasons, which mixes objective data and subjective data like team strength, team form, psychological impact and fatigue.

Jan and Lit [5] area unit increasing work by Maher on the statistical distribution, demonstrating the offensive and defensive power of the goal score distribution. Koopman and Lit area unit developing an applied math model for the study and estimation of the outcomes of soccer matches, which assumes a quantity distribution of Poisson with coefficients of intensity that adjust at random over time.

Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning by Rabindra Lamsal and Ayesha Choudhary. Cricket, particularly the Twenty20 style, has the highest level of unpredictability, with a single over having the potential to radically swing the game's momentum. With millions of people watching the Indian Premier League (IPL), creating a model for forecasting match outcomes is a real-world challenge [6].

A cricket match is influenced by a variety of circumstances, and the elements that have a major impact on the result of a Twenty20 cricket match are discovered in this study. The total weight (relative strength) of the team is determined by each player's performance on the field. To calculate points for each player in the league, a multivariate regression-based approach is provided, and the total weight of a team is computed based on the historical performance of the players who have appeared the most for the club. Finally, a dataset is modelled based on the identified seven factors which influence the outcome of an IPL match. Six machine learning models were built and utilized to forecast the outcome of each 2018 IPL match 15 minutes prior to the start of the game, just after the toss. Three of the trained models accurately predicted more than 40 matches, with the Multilayer Perceptron topping all others with a remarkable accuracy of 71.66%.

Machine learning algorithms are employed in this work to predict the outcome of soccer matches.

Although it is impossible to account for all factors that impact match outcomes, an attempt is made to identify the most important factors, and several classifiers are evaluated to tackle the problem.

Below, the literature study is summarized in form of a comparative study (Table 1).

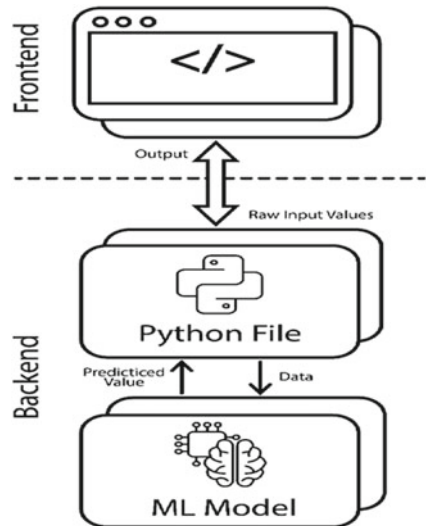
3 Methodology

To describe the approach that we have adopted for this project, it is imperative to understand the tools and resources that were available and put to use. There are two parts to this project, the frontend that was built using the simple and basic technology of HTML, CSS and bootstrap. Connecting to this is the backend, that makes it functional. It comprises of the machine learning model and some additional files to enable the model to process the input from the frontend and then provide predicted output back to the frontend. This section tailed a detail study of the same. The diagram below facilitates the understanding of how different component of the project interact with each other (Fig. 1).

Table 1 Comparative study of technical papers

Paper Title	Authors	Takeaway
Predicting outcome of soccer matches using machine learning	Albina Yezus	The study takes into account one particular league for its evaluation. It states that a simple regression model can give the results that are as good as the results given by some complex models. This study claims that a model will accuracy of 60% can be deemed a good model
The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review	Rory Bunker and Teo Susnjak	This paper claims that ANN algorithms are superior to other for predicting the outcomes of a sports match. It contains a deep analysis of various classification algorithms. It also focuses on the importance of a good dataset and mentions the challenges of finding one in the field of sports
Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning	Rabindra Lamsal and Ayesha Choudhary	This study brings to light a solution of predicting match outcomes in IPL using multivariate regression. This is key because, this is the algorithm, that will be further used by our purpose too

Fig. 1 A visual representation of how different components of the project interact with each other



4 Dataset

4.1 Description

The dataset that we are using for building this project contains information that was partly derived from online resources and partly built by intuition and common knowledge of the given situation. The available dataset included fields such as XYZ. However, this plainly would not be very sufficient to predict the match outcomes. Before we move onto pre-processing the data and use it to derive any meaningful result, we must first be fully aware of the various parameters that are present in the raw dataset (Table 2).

4.2 Exploratory Data Analysis

Even before we start building out model, it is imperative that we analyse the details in the dataset. It is important to understand the data and how well it is structured before we pre-process the data to make it useful. Given below are the screenshots of the plots that were produced by python for the given dataset. These plots eventually helped us to understand the structure of the data and give us an insight regarding the kind of pre-processing that will be required to make the dataset efficient in its use (Figs. 2 and 3).

4.3 Data Pre-processing

Data pre-processing is defined as a process of preparing and making the raw data suitable for a machine learning model. It is very rare to come across a clean and formatted data when creating a machine learning project. And while doing any operation with data, it is vital to clean it and arrange it in a formatted way.

It is necessary that we filter our raw unstructured data into a format which can be fed to models and be processed to derive meaningful results. There are various steps involved in data pre-processing such as:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Table 2 Description of features of dataset before pre-processing

Sr. No	Feature name	Meaning
1	Div	League Division
2	Date	Match Date (dd/mm/yy)
3	Home Team	Home Team
4	Away Team	Away Team
5	FTHG	Full Time Home Team Goals
6	FTAG	Full Time Away Team Goals
7	FTR	Full Time Result (H = Home Win, D = Draw, A = Away Win)
8	HTHG	Half Time Home Team Goals
9	HTAG	Half Time Away Team Goals
10	HTR	Half Time Result (H = Home Win, D = Draw, A = Away Win)
11	Referee	Match Referee
12	HS	Home Team Shots
13	AS	Away Team Shots
14	HST	Home Team Shots on Target
15	AST	Away Team Shots on Target
16	HHW	Home Team Hit Woodwork
17	AHW	Away Team Hit Woodwork
18	HC	Home Team Corners
19	AC	Away Team Corners
20	HF	Home Team Fouls Committed
21	AF	Away Team Fouls Committed
22	HO	Home Team Offsides
23	AO	Away Team Offsides
24	HY	Home Team Yellow Cards
25	AY	Away Team Yellow Cards
26	HR	Home Team Red Cards
27	AR	Away Team Red Cards
28	HBP	Home Team Bookings Points (10 = yellow, 25 = red)
29	ABP	Away Team Bookings Points (10 = yellow, 25 = red)

In our data, we have performed importing libraries, importing datasets, encoding categorical data and splitting the dataset for train and test purposes. In The figure given below we have demonstrated the dataset before pre-processing, the code that was used to pre-process the dataset and finally the outcome dataset from the code (Figs. 4 and 5).

Now the dataset comprises of fewer columns therefore making the regression easier to perform. The pre-processed dataset contains the information stated in the table below (Table 3).

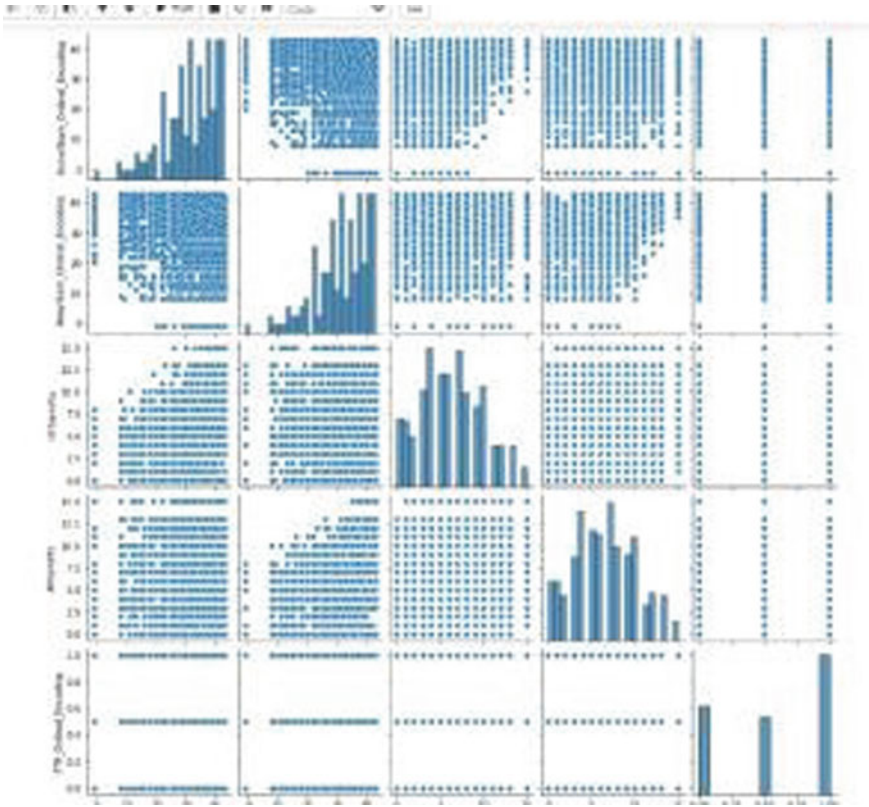


Fig. 2 Exploratory jointplot of dataset

Fig. 3 Exploratory pairplot of dataset

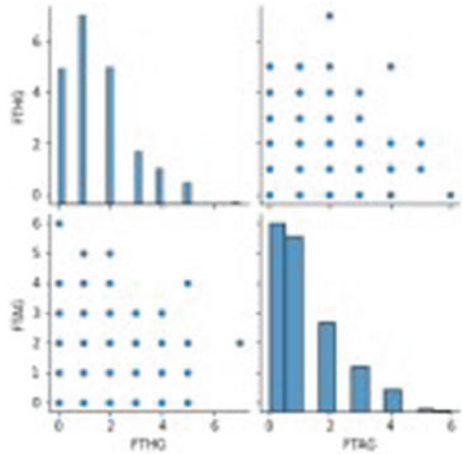




Fig. 4 Screenshot of the raw dataset prior to pre-processing



Fig. 5 Screenshot of the dataset after preprocessing

Table 3 Description of features of pre-processed dataset

Sr. No	Feature name	Meaning
1	HomeTeam_Ordinal_Encoding	Categorical Encoding for the Home Team
2	AwayTeam_Ordinal_Encoding	Categorical Encoding for the Away Team
3	HTFormPts	Recent winning/losing form of the team in the last 5 matches for the Home Team
4	ATFormPts	Recent winning/losing form of the team in the last 5 matches for the Away Team
5	FTR_Original_Encoding	Full Time Result

5 Choice of Model (Algorithms)

Predictive analytics models, which are now the most sought-after in the market, are meant to evaluate past data, uncover patterns, detect trends, and utilize that information to make predictions about future trends.

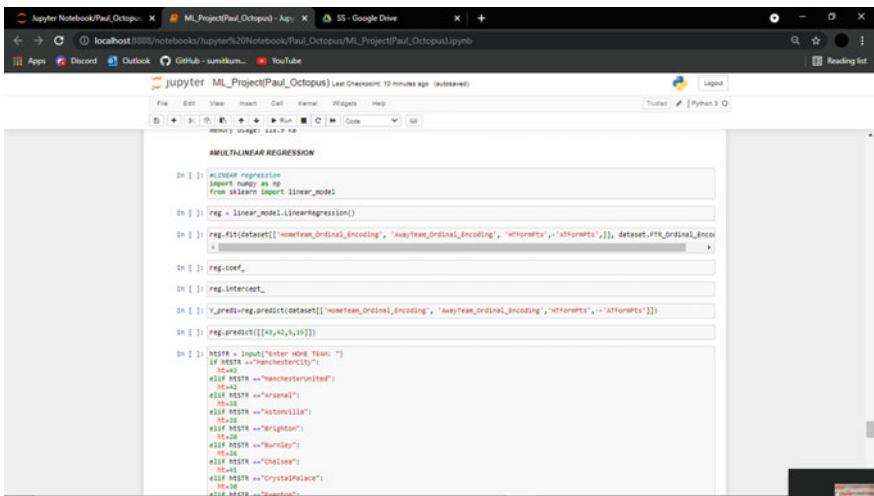
The following are a few of the most prevalent methods:

- Decision tree algorithms take data and graph it out in branches to display the possible outcomes of various decisions.

Decision trees also classify response variables and predict response variables based on previous decisions and can be used with incomplete data sets. They are easily explainable and accessible for novice data scientists.

- Time series analysis. This predicts events through a sequence of time. You can predict future events by analyzing previous trends and extrapolating from there.
- Regression. This is a statistical analysis method that helps in data preparation. The algorithm’s capacity to sort and categorize data increases as more data is added, allowing predictions to be produced.

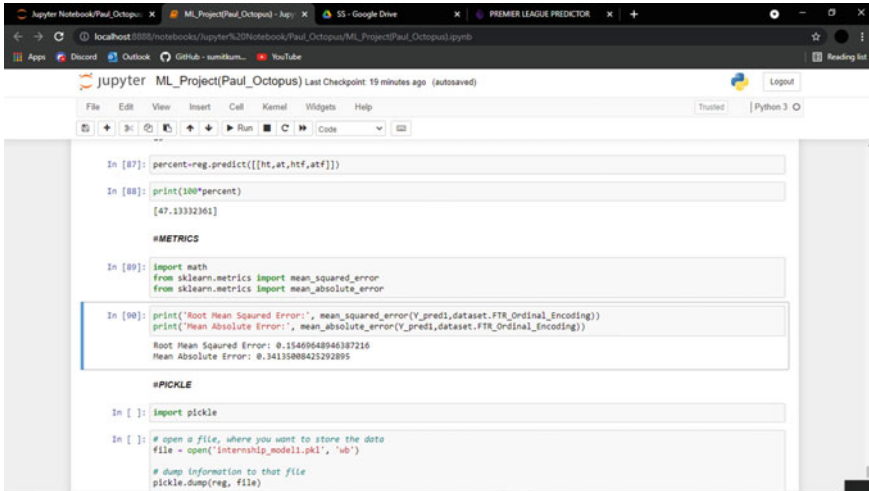
For this project, we have used multivariate linear regression due to its simple approach and accurate results in our field of exploration. Below are the code snippets of the code used to build the model (Fig. 6).



```
from sklearn.linear_model import LinearRegression
reg = LinearRegression()
reg.fit(dataset[['homeTeam_ordinal_encoding', 'awayTeam_ordinal_encoding', 'winOrDraws']], dataset['FT_ordinal_encoding'])
reg.coef_
reg.intercept_
y_pred = reg.predict(dataset[['homeTeam_ordinal_encoding', 'awayTeam_ordinal_encoding', 'winOrDraws']])
reg.predict([[13, 12, 9, 10]])

def predict(input):
    if input == "Manchester":
        return 10
    elif input == "ManchesterUtd":
        return 10
    elif input == "Arsenal":
        return 10
    elif input == "AstonVlla":
        return 10
    elif input == "Brighton":
        return 10
    elif input == "Burnley":
        return 10
    elif input == "Chelsea":
        return 10
    elif input == "CrystalPalace":
        return 10
    elif input == "Liverpool":
        return 10
```

Fig. 6 Code snippet



```
In [87]: percent-reg.predict([[ht,at,htf,atf]])

In [88]: print(100*percent)

[47.1333236]

#METRICS

In [89]: import math
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error

In [90]: print('Root Mean Squared Error:', mean_squared_error(Y_pred1,dataset.FTR_Ordinal_Encoding))
print('Mean Absolute Error:', mean_absolute_error(Y_pred1,dataset.FTR_Ordinal_Encoding))

Root Mean Squared Error: 0.15409648946387216
Mean Absolute Error: 0.34135808425292895

#PICKLE

In [ ]: import pickle

In [ ]: # open a file, where you want to store the data
file = open('interhip_model1.pkl', 'wb')

# dump information to that file
pickle.dump(reg, file)
```

Fig. 7 Code snippet

6 Testing and Training (Evaluation Metrics)

Coding is not sufficient when we talk of predictive Machine Learning models. We also need to train the model so that it can identify the relationship between the dependent and independent variables. Then we have to test out model to determine its accuracy.

$$Accuracy = \frac{No. of Correct prediction}{Total No. of Prediction Made}$$

However, in our model we do not make a train test split. This is because the relatively small amount of data available to train the model. Thus, testing is done for real time data (Fig. 7).

7 Result and Discussion

In the snippet below, we see the front-end of the project also we see a sample of how we can use the webpage for predicting the winner of a football match.

The model gives us a result that is accurate 70% times. This can be considered as a good model since it is not only simple but it also takes into account various factors of a match. The data is condensed into five columns from an overwhelming number of columns initially. The features were well used to predict the results (Fig. 8).

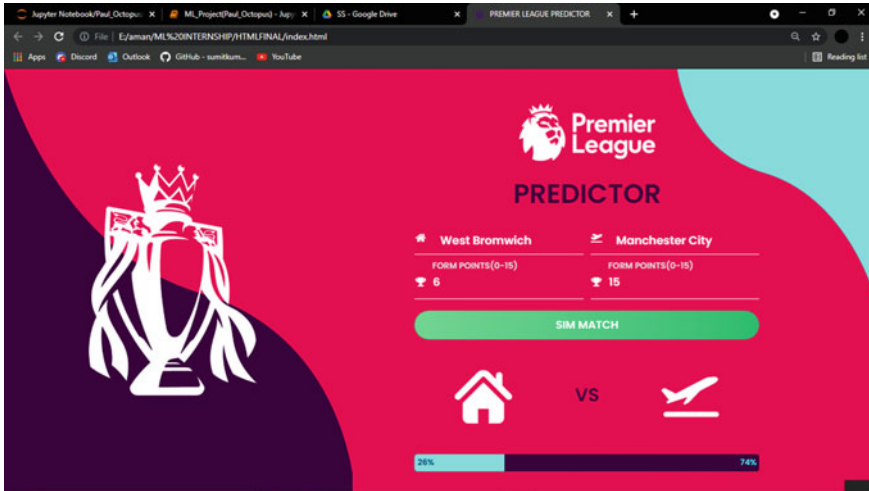


Fig. 8 Demo of the software

8 Future Scope

No project can ever be truly called complete. There is always room to do more and to do better. Such is the case of our project. In our project we have implemented a Linear regression however, other models can also be tried. We can also adopt a more advanced dataset that can help us to make better predictions. Despite several attempts we couldn't find a dataset that included the players of each team and hence we would have to work further to form a dataset in which the key players of each team are encoded and that would become an influential factor in determining the winning chances of each team. We can also develop the predictor to also predict the number of goals that would be made by each team. However, for such computation a larger dataset would be required.

9 Conclusion

To conclude, this paper has presented deep insights into what goes into making a project that amalgamates machine learning and web development. We discussed various studies to justify our choice of model which is multivariate regression. This study also presents all the steps undertaken in pre-processing, model development and training and testing of the model.

Acknowledgements Developing this project wouldn't have been possible without the support of the faculties at TCET. We are thankful to our mentor Dr. Manish Rana who has guided us throughout the technical paper. We are grateful to receive your valuable input that has enhanced this paper above

the mediocrity. We grateful for the opportunity, to present this paper. We are also thankful to the authors and resources which have provided a strong base for our research and project development.

References

1. Yezus A Predicting outcome of soccer matches using machine learning. Term paper at Saint-Petersburg State University
2. Razali N, Mustapha A, Clemente FM, Ahmad MF, Salamat MA (2018) Pattern analysis of goals scored in Malaysia super league. *Indonesian J Electr Eng Comput Sci* 11(2):718–724
3. Min B, Kim J, Choe C, Eom H, McKay RB (2008) A compound framework for sports results prediction: a football case study. *Knowl-Based Syst* 21(7):551–562
4. Anthony C, Fenton N, Neil M (2012) Pi-football: a Bayesian network model for forecasting association football match outcomes. *Knowl-Based Syst* 36:322–339
5. Jan KS, Lit R (2015) A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *J Roy Stat Soc Ser A (Stat Soc)* 178(1):167–186
6. Azeman AA, Mustapha A, Mostafa SA, Abu Salim SWG, Jubair MA, Hassan MH (2020) Football match outcome prediction by applying three machine learning algorithms. *Int J Emerg Trends Eng Res* 8(1)