PREMIER LEAGUE MATCH RESULT PREDICTION USING MACHINE LEARNING

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering/Information Technology

By

Sushant (151273)
Deepanshu Rana (151375)

Under the supervision of

Mr. Amol Vasudeva

to

Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Premier League Match Result Prediction using Machine Learning"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2018 to May 2019 under the supervision of **Mr. Amol Vasudeva** (Assistant Professor (Grade-II), Computer Science & Engineering and Information Technology).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Sushant, 151273   Deepanshu Rana, 151375

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Mr. Amol Vasudeva
Assistant Professor (Grade-II)
Computer Science & Engineering and Information Technology
Dated:

ACKNOWLEDGMENT

We would firstly like to thank our supervisor Mr. Amol Vasudeva at the Department of Computer Science & Engineering and Information Technology at Jaypee University of Information Technology, where this project has been conducted. We would like to thank him for the help he has been giving throughout this work.

We have grown both academically and personally from this experience and are very grateful for having had the opportunity to conduct this study.

We are also thankful to all other faculty members for their constant motivation and helping us bring in improvements in the project.

Finally, we like to thank our family and friends for their constant support. Without their contribution it would have been impossible to complete our work.

Sushant, Deepanshu Rana
JUIT Waknaghat
May 2019

TABLE OF CONTENT

LIST OF FIGURES

ABSTRACT

The ease of the accessibility of Internet and the popularity of Machine Learning have been the prime reasons in the increase of Sports Analysis and Betting. Football being the most popular sports in the that is played in over 200 countries world is regarded as much more dynamic and complex when compared to other prevailing sports and this makes football an interesting field for research. For the development of prediction systems several methodologies and approaches are being used. In this project we predict the result of a Premier League match given a home team and an away team. The predictions are made based on various important attributes that consists of data from previous seasons of Premier League. These important attributes are very likely to decide the outcome of a match. For the purpose of predictions, we use three different algorithms namely, Logistic Regression, XGBoost and Support Vector Machines and then we select the best algorithm out of these three to predict that appropriate label. The application of these models are done on real team data and results of fixtures are gathered from http://www.football-data.co.uk/ for the seasons ranging from 2003/04 to 2018/19.

1. INTRODUCTION

1.1 INTRODUCTION

With the transcending environment of technology emanating from the Internet Technology, the implementation and generation of fascinating new ideology from the human brain has gone to its uttermost peak. Furthermore, on considering the global approach, there have been some intriguing development as the major fields in consumer products are prime example of it. Some of the other known indistinguishable transcendence in the world of the Internet have also ranked at the status where the work has been done diligently and acting as a stepping stone for success in terms of revolution with the time, and clearly the potential has been shown when it comes to sports. There are innumerable possibilities of transitions done by the internet of how it connects the dots irrespective of the genre of any sports varying from football to basketball and basketball to baseball, this infiltration among sports has led to its own distinct existence may be called as “Internet of Sports”.

One of the biggest mutation in “Internet of Sports” has not only put the financial success stories in the codex of various professional athletes but also evolved the company on the path to grandiose triumph. Activities and subjects of interest like Statistical calculation of the athletes was seldom worked in an organized and sequential manner back in the ages where these concerns were age old conundrum. With the help of the power of the Internet, the everyday analysis of these dilemmas seemed so easy to implement and simple to observe. With so much ease for people to get into work, take some of the quantum time of their lives and all kind of stuffs that seem impossible or very hard to do back, we can finally conclude that the “Internet of Sports” have been very helpful in the Fantasy Sports too!

With the passage of time, there are some massive improvements in terms of sports broadcasting, Extensible approach has been introduced where the selection of streams to catch a game has been implemented which was quite handful back in the time. To summarize the past decade in terms of “Internet of Sports”, it has transcended exponentially, in terms of fantasy sports and also for accurate forecasting in the market of prediction, which will form a primary concern of the correct working and implementation of the project.

The foremost objective of any forecasting of the prediction is accuracy with precision. The chief affairs of prediction of the outcomes must be up to the point in various aspects of fields like Business Analytics, Sports Analytics and other primary areas. And as far as tools of market prediction are considered, we can use the immense pool of power of the Internet for the forecasting of the future events which will take help in achieving the key objectives that are seemed to be observed, and the correctness in the development where the growth is transcending with every passing year. These prediction tools are being used specifically by various fortune 500 companies and some of the biggest bosses like Microsoft, IBM, Google, Amazon etc. This appearing markets is mainly having its growth due to their correctness, in all the fields of surveying market growth, which is made more efficient with the availability of large data sets.

The forecasting of sports prediction has not only used as a means of entertainment, but also helps for gameplay assessment of players, teams, leagues, and the associated results. Furthermore, unfolding decision making of respective coaches and staffs, financial success by enlarging revenue of stadium are also emitted.

Although, people hold different thoughts regarding internet betting, but regardless of these thesis of varying opinions, the absolute level of usage of this blended mixture of Internet and betting is being popular day by day and the results are quite optimistic.

1.2 PROBLEM STATEMENT

Our report has a sole objective to observe the emanating influence of the Internet Sports betting on global basis, and how the factors (age restriction, improper bookmakers, isolated transactions among people) are being examined and influenced by distinct users. Bookmakers with bad customer support affects the whole structure of the system and involvement of money initially is a bad approach. These collaborated problems which arise in betting process commonly creates menace in society and are against some of the laws and other legal standards.

Using some secure approach in transaction processes, without losing our minds on performance-precision criteria using some brute-force examination of all positive and negative aspects, we can introduce a safe, secure and maintainable online betting system with varieties of extensions.

1.3 OBJECTIVE

We will consider the teams from English Premier League, work on their datasets from past decade and develop an application for sport betting for the odds of win and loss of teams between each other. For the data analysis of the past results of all the teams, we will perform a step-by-step approach. We will predict the results which will work between 0 % to 100% of team wins and team loss(normalized to 0-1), these results get affected by the performance of individual teams in the past 10-15 years and therefore, with it, we will measure the results of all the forthcoming matches to be happened in future. This results will be valid for all the forthcoming as well as ongoing fixtures set up by the Board of Premier League in England.

We will use different classifiers for results prediction, compare the result and choose the optimal classifier for proper match prediction. Models like SVM Machine, XGBoost and Logistic Regression will be used for primary classification of the data. After all the results, we can also extend the performance and precision processes by using some advance tools.

The implemented model will be sampled and used for outcome predictions. The Structuring of the project will be performance and accuracy driven, which can be extended by using large datasets and stable data frames.

Following are the brief report contents which are available in this report:

1. Methodology, Literature Survey on prediction markets and work related market analysis.
2. Systematic Model development using some mathematical and analytical approach.
3. Algorithms implemented for the project creation.
4. Test plans formation, which will work as a part of metrics and usage of data sets.

5. Performance and accuracy analysis of the result. Precision will be calculated of each classifier for proper data prediction.

1.4 METHODOLOGY

The base of our project concentrate on the stats like home team winning or losing percentage, away team winning or losing percentage and draw results. These results will be assessed by the users to place their bets on respective teams.

Data like points criteria, past team wins and losses, total yellow and red cards acquired by a team, general behavior of referee regarding team and the referee appointed to particular matches etc. all are the parts of data sets which are gathered from the past 15 seasons. This large data sets will also help to improve the predictions by emanating more stable and accurate results so as to give more possibility of profit making than loss.

Although the high volume data sets may cause some data inconsistency, but for the long run, it will be successful. Furthermore, due to demand of instant propagation of data and information which is fulfilled by the significant part of technology, there will be a transcendence in aspect of data analysis by using tools of machine learning and accurate data sets.

1.5 ORGANIZATION

Sequential Approach carving of the project have been implemented by us. Datasets from past decades will be featured for all teams and detailed bird eye view of the project will be shown for proper understanding. For the acknowledgment of the behavior of the models on teams with the passage of time, we will use CRISP-DM Framework with some use of exploratory and spiral analysis.

Moreover, proposal of various tools and requirements to implement the betting strategies associated with the datasets. Corrective predictions is our key element and thought process the quantity of money involved will be assessed by our application. Chapter 4 will be the powerhouse of our report as the critical implementation will be taken care in it and types of model and algorithms implementation will be explained.

Chapter 5 is mainly contrasting the documentation of test plans such as data sets and the metrics.

The following chapters will mainly deal with performance analysis by in depth project scope examination. The robust mathematical as well as analytical computations will be done by emanating different inputs and the comparison will be done on the basis of behavior of the result

2. LITERATURE SURVEY

“Due to the expanding trend, sports betting system has become a fun and interesting way to bet. With the dynamic range of sports, the online betting has been a bridge among the sports fantasy and business success. But with varying diversity, the thought process of people is also becoming different. Some consider it as entertaining and a great side hobby to have, a good way mathematical statistics study improvement for long run. While many are against it and consider it as a part of actual gambling.”

(kingsandqueens.com)

“The entertaining world of online gaming apps has helped in creating a pleasant and convenient environment. The way we enter the money and some other dynamic options have made it quite user friendly and easy to implement. With this ease, the operation on these applications is suitable to all types of users.

The only thing matters that degrades the process structuring is legal issues as well as regulatory issues, and these issues vary around the globe. This problem put a direct impact on people who might be interested in doing it.”

(gamblingsites.org)

“Online sports betting is a great asset as a financial resource for schools, senior citizens and reduction of wages among states in United States of America. Other than this, some online sites providing gambling services are being accepted as a fun and leisure activity and is favored by majority of Americans. But some issues are still faced by college football due to inequity among colleges that will receive colossal share of profits comparative to

the others that will not. This will somehow create a stress of financial disproportion and the competence gap in college football.”
(stateoftheu.com)

“A little Gambling makes the games excited, and most people don’t see sports gambling as a serious problem but the consequences may vary and may shift the environment in terms of socio-economic costs. But with the evolution of new technology into the fantasy world, the government has started amending new laws for the sports gambling to prevent it going out of hands. The main issue that must be a primary objective is to improve the areas of education. To acknowledge the behaviors of college students, players and referees. And if the online gambling is being adopted, there must be more stern punishments for practicing illegal betting.”
(lawteacher.net)

“Online sports betting has already blossomed as an exponentially increasing market industry that works on a high profile resources like money, properties and other assets, not only in terms of entertainment but providing the stable resources to the regions. It has power to create opportunities for the native people by providing jobs. This will help in overall increment of the revenue for the states in US by billions in next 9-10 years. But still there will be some essence of illegal betting in this whole process as it won’t completely get eradicated from the scenarios. Not only this, there may be temporal or permanent shift of the sports from game driven to money driven process.”
(theperspective.com)

RESEARCH PAPER I:

Forecasting techniques were implemented to describe the studied event of FIFA World Cup 2006 and prediction markets were introduced for prediction of results of all the matches held in the tournament. For prediction market, they created a web interface that ran in coherence with the time the tournament took place. They calculated some virtual currency units that had certain values for every match win or loss. The trading screen interface was well evaluated.

Researchers used some datasets as a new benchmark, which happened after forecasting on historical data that has been done by them. The second benchmark was on the basis of ranking of teams by FIFA while considering random predictor as a primary one. For accuracy comparison, the percentage of correctly predicted games were reckoned by the researchers and this data was stored as hit rate (33.3% for random draws). Greater the hit data, more are the chances for a team to win a match.

RESEARCH PAPER II:

With the evolution of online sports betting industry in Kenya, researchers considered as a perfect getaway to analyze its trend and vogue for the future related concerns, whether it has a positive or negative impact on the society. The impact was analyzed thoroughly and fields of anomalies where the case were prominent were determined. This whole report was aimed to present the extent of sports betting impacting on general public of Kenya.

The study was commenced for a period of 7 months in Nairobi, Kenya. Various models like Mobile money penetration and Technology Adoption Model (TAM) were provided to trace out briefs, attitude and intentions of people to use the sports betting applications. The survey was being done on all categories of people from adolescents to elderly and their behavior were thoroughly noticed. The results were saved for the assessments of demographics vulnerabilities regarding gambling and subsequent problems.

For research methodologies, researcher deduced an overall plan to get the resultant decisions of people. Plans included various descriptive studies, with longitudinal research designs. Using descriptive research design, the trend was projected of sports betting growth and its effect on distinct groups.

Other methodologies used were quantitative approach, to quantify data in numerical quantities. These methodologies helped researchers to study the statistical measurable variables.

Sampling design categories (Frame and Technique) were referred and were used to determine the sample of population, its selection of observable sample and the tolerance of margin of error in selected sample. This study used various formulas to come up with adequate sample size.

Furthermore, some other data collection methods were used as an amplification factor for the data collection.

RESEARCH PAPER III

In this report, the project was made on the basis of real life entities. Strategies like skating plans and bet selection processes were prominent solution and were hugely involved for system development. The main aim of the research was evaluation of betting systems under consideration and to find the best one.

The sports betting introduction, its main basics, were clearly explained. Furthermore, with the introduction of this process, its transcendence in the future society was also explained like how the analysis of market study will help in overall growth of revenue, which may exceed more than 43% of the current inflow.

Solid selection process and staking optimization are the factors for tuned sports betting systems. This evidence was provided with results and proper discussion. The factors have potential to produce profits within the limited amount of time in ideal conditions.

But for real life situations, other secondary factors like probability level and uncertain time frames comes in place particularly to the specific scenario of betting applications. Overall, the above methodology executed is invaluable for determining maximized profit and reducing risk factors of specified betting system.

RESEARCH PAPER IV

This report mainly focuses on implementation of RNNs which are also called Recurrent Neural Networks and its usage. Furthermore, with using RNNs as our primary technique, there is detailed overview of the implemented distinct architectures with sequential and proper inputted data.

LSTM architectures (RNNs subset) were tuned. On the other hand, these tuned architectures were robustly tested using some defined test sets. These tuned and tested architectures were used for effective results in predictions with different strategies. The result were more favorable for 'many-to-one' strategy over 'many-to-many' with the difference in accuracy of more than 10% between them. Overall, for the classification accuracy, more sequential data was provided to increase the accuracy. Other factors like team and player information and location were also giver higher preferences for outcomes.

Embedding or embedding space were modelled for distributed representation of data sets and preventing the data to get more and more complex when working in high-dimension data space. Moreover, embedding space were quite helpful to algorithms for better performance in areas such as natural language processing(NLPs). Classifiers like SOFTMAX and cross entropy error with optimizer were used in multiclass classification methods and profoundly used for measuring errors for neural networks. Some choices of hardware were also required for faster execution phases and maintaining the system composition with ongoing processes.

For Higher and complex mathematical computations and parallel computations, the CPUs and GPUs were some of the most vital and prominent factors. For different GPUs manufacturers, (Like AMD and Nvidia), specified libraries are provided (online and offline) for process simplification.

RESEARCH PAPER V

This research paper mainly deals with Data Augmentation and how concepts of generalization can prevent to arise anomalies with the imbalanced datasets using Auto

Augment. The other purposes that can be achieved using this approach is to emanate search augmentation policies from a dataset for performance improvement.

The methods for reckoning fast AutoAugment were clearly defined where Search space is evaluated firstly. For search space, we use two parameters, probability and magnitude and some of the sub-policies which are used to perform consecutive operations and the resultant is applied with the evaluated probability.

Other evaluation process includes some strategies for effective density matching to improve generalization ability using mapping of densities. Furthermore we imply some objectives to find a set of learned augmentation policies. Other techniques which includes K-fold stratified shuffling for splitting the trained datasets into different small sub datasets. Finally, we explore some other augmented policies via Bayesian Optimization.

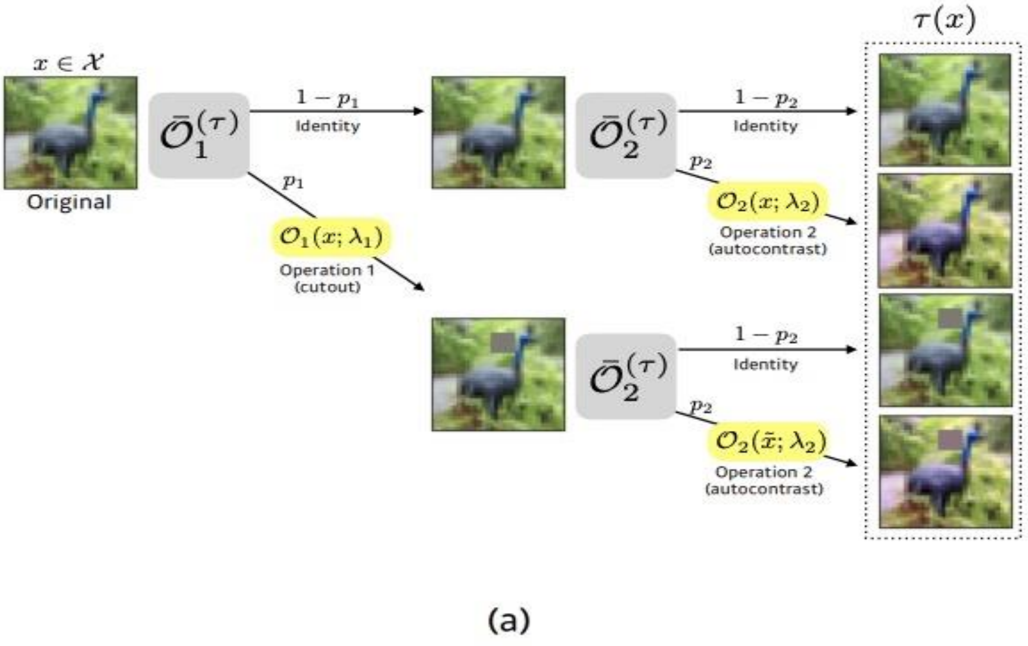

$x \in \mathcal{X}$
Original
$\bar{\mathcal{O}}_1^{(\tau)}$
$1 - p_1$
Identity
p_1
$\mathcal{O}_1(x; \lambda_1)$
Operation 1
(cutout)
$\bar{\mathcal{O}}_2^{(\tau)}$
$1 - p_2$
Identity
p_2
$\mathcal{O}_2(x; \lambda_2)$
Operation 2
(autocontrast)
$\bar{\mathcal{O}}_2^{(\tau)}$
$1 - p_2$
Identity
p_2
$\mathcal{O}_2(\tilde{x}; \lambda_2)$
Operation 2
(autocontrast)
$\tau(x)$
(a)

Figure 2.1, showing the basic structure Classification in imbalanced datasets

Hence, by proposing a automatic process of learning policies of a network, our search method has been improved and performance are compared, so as to establish a benchmark. Also for advanced architectures, concept of fast AutoAugment is used which can be slower for the former part. Furthermore, Fast AutoAugments are very reliable in areas of AutoML. This can also affect the classification of the images beyond computer vision in near future.

RESEARCH PAPER VI

In this report, Logistic Multivariate analysis was used to predict the retention of master's candidates at an outsized university in Canada. They elected Demographic such as: age, citizenship, GPA, study type, degree completion etc and also used Financial variables (funding received from internal as well as external scholarships and from research, and other teaching assistantships). Other variables were used such as freelance and divided variables for the passing and failure of the coeds within the program.

For the doctoral candidates, solely inflated length of the time, and inflated funding from all the sources were used to reckon the increment in probabilities of the graduation with the degree.

For this purpose, the researchers have used logistic regression analysis and the results were out breaking. The candidates with higher GPAs, increased length of time in program, scholarships and other factors significantly improved the chances of successful graduation Furthermore, with this study, the concluded part was the curriculum choices and attempts to choose the financial sources for students must be considered carefully and varies from programs as well by the institutions.

3. SYSTEM DEVELOPMENT

This project focuses on the results of the football fixtures for upcoming matches. Therefore, a structured approach is needed as it provides more of theoretical basis rather than the experiment basis. The basic framework that is being used is CRISP-DM framework which stands for cross-industry process for data mining. CRISP-DM is robust and hence, provides a better methodology to predict the results of the fixtures.

CRISP-DM FRAMWORK:

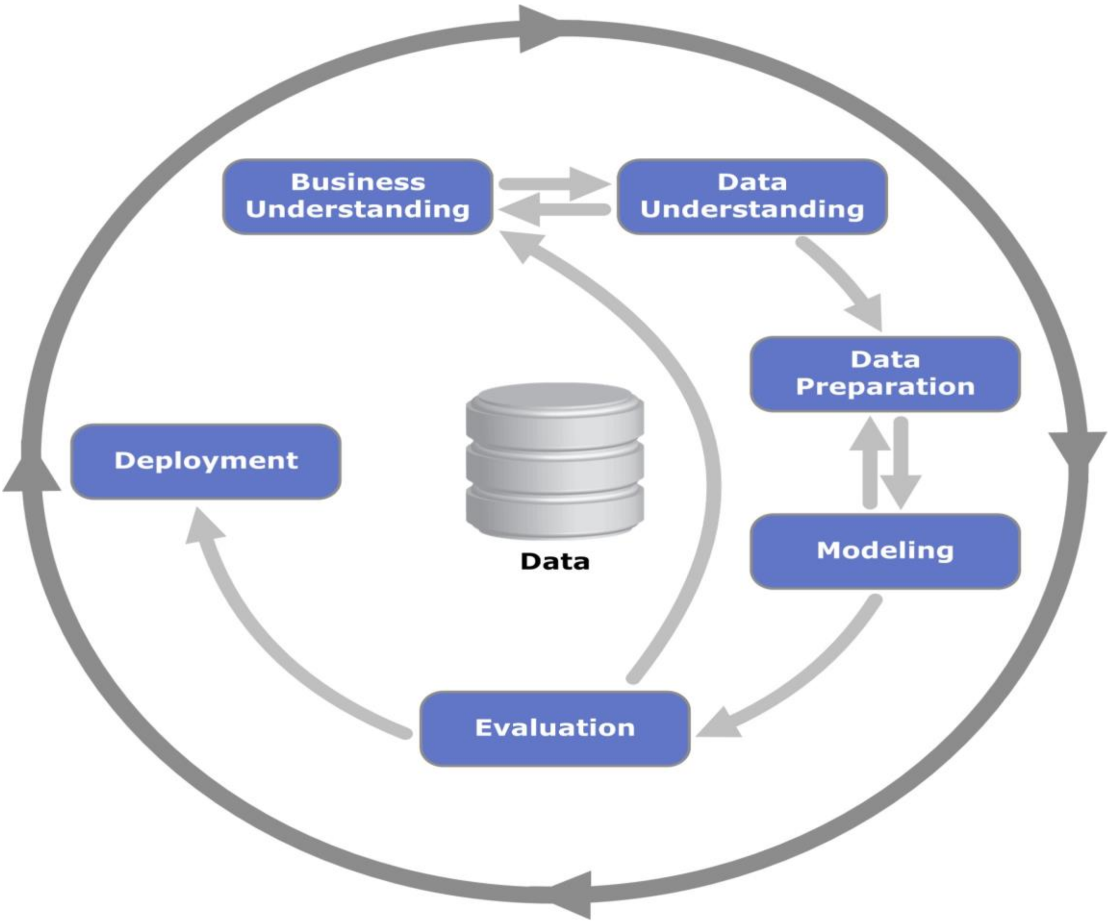


Figure 3.1, showing the basic structure of CRISP-DM framework

There are six major phases of CRISP-DM framework. CRISP-DM framework is flexible and therefore, the phases need not be used in order. The phases in the framework are dependent on each other and the arrows shown in the figure 3.1 don't show these dependencies rather they tell us about how the execution works. The datasets in data mining have a never ending observation and are represented by the outer circle of the diagram. This is basically used to enhance the structure of the system and as well as the performance also gets optimized.

As already described that the above framework is flexible and the phases need not to be applied in a particular order but the steps that are involved in these phases need to be in a specific manner. Following diagram shows the steps of the framework.

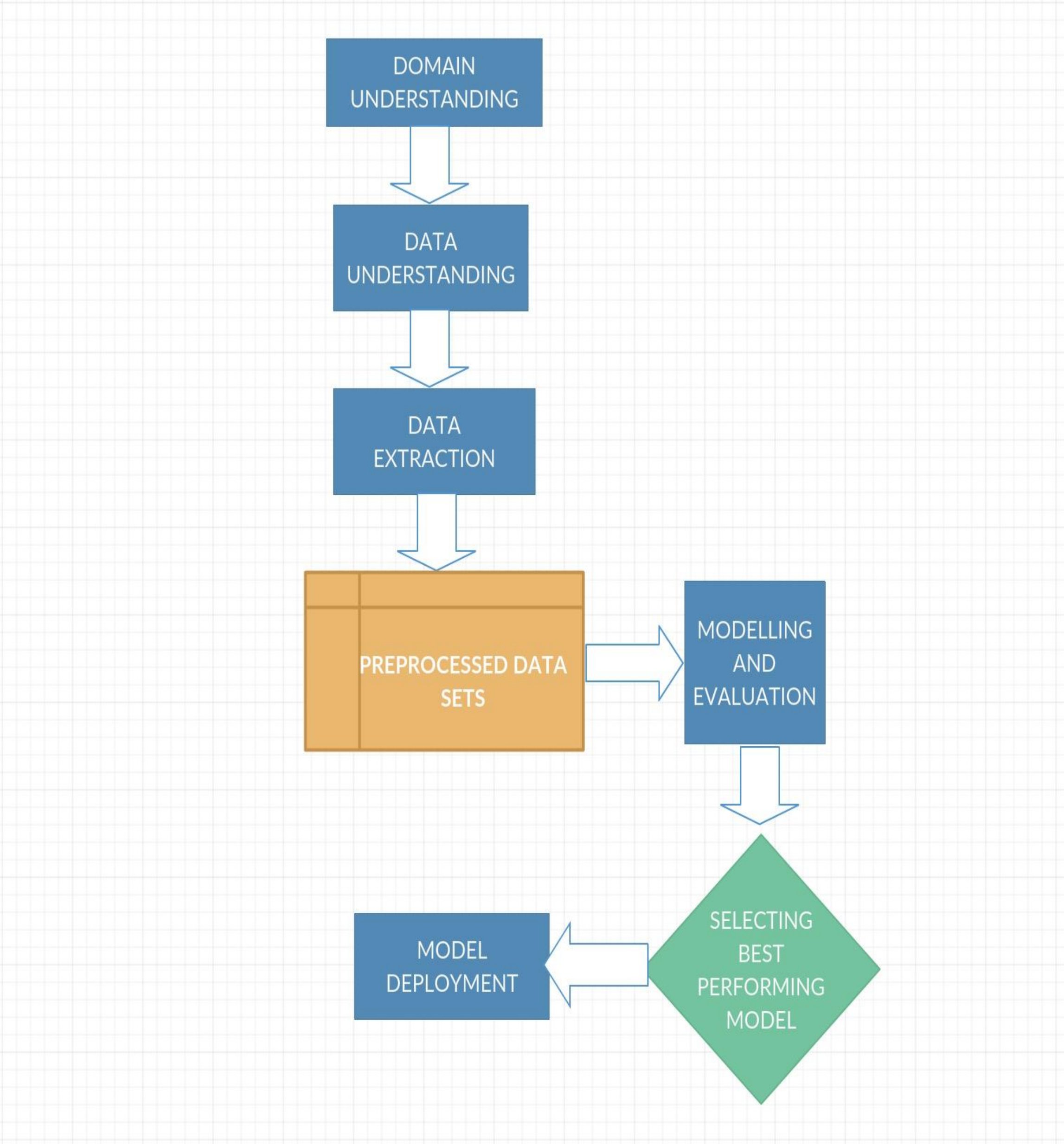


Figure 3.2, showing the flow chart of the methodology uses to complete the project

3.1 DOMAIN UNDERSTANDING

Understanding problems and the objective of the problem is the key phase in this part. How a sport is structured, its essence and what are the factors included in predicting the outcome is determined. The references by which domain comprehending works can be further extended via personal knowledge of the specific sport or reviewing the literature and papers of research.

3.2 DATA UNDERSTANDING

The available resources could be used for comprehension of the data obtained. Some prior data may be stored in these resources which is an example of automation and online extraction.

To further enhance the user experience, user need only input the data and get desired results as output.

We will include player data, their stats in every game they played and will be separately stored in different data sets. For simultaneous prediction of goals scored by the player, we can use simple join operations among data sets. This will not only influence the precise outcomes but also extend the scope of our project.

3.3 DATA EXTRACTION

The feature subsets are created in the data extraction phase. The subsets can be any attributes like they could be goals scored or team standings. The features are divided in to subsets like the odds ratio or the values of the expert opinions by some researchers. But in this project only the internal features will be worked upon and not the expert opinions because this project doesn’t focus on sentiment analysis.

Match related features only concentrate on the arithmetic part i.e., goal difference, goals scored, whereas, external featured focus on the analytics part i.e, recent form of the teams, name of the official refreeing the match, etc. These two types of features work separately but when aggregated together give us the complete result.

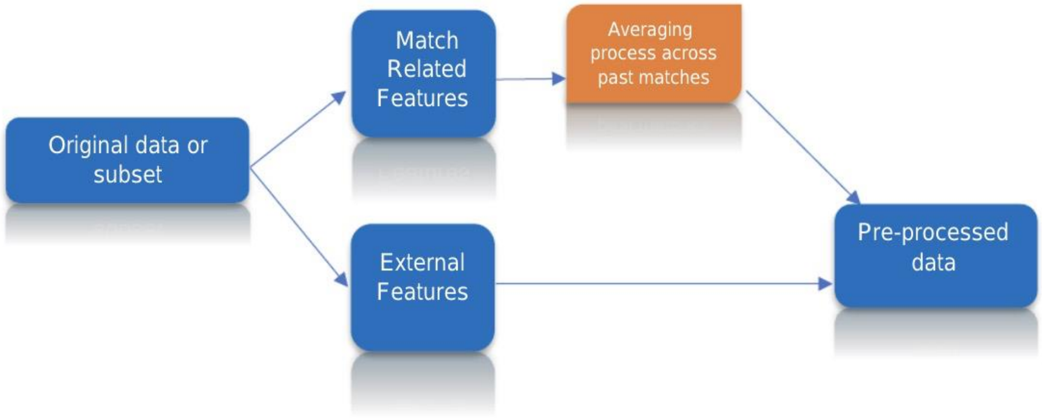


Figure 3.3, showing how the dataset is divided

3.4 MODELLING AND EVALUATION

Various research papers, journals and conferences were studied to come to a decision about which predictive model could be used for this very project.

The combination of classifiers and features will also be determined by this process.

For evaluating the models, the performance of each chosen model will be compared and confusion matrices will be created for each and every model, for the data that is balanced this king of evaluation is the best, but for highly imbalanced data, the concept of curve evaluation called Receiver Operating Characteristic (ROC) will be used because the results of upcoming fixtures are predicted on the basis on past fixtures, and the order of the training set must also be preserved. To shuffle the order of all objects cross validation techniques could also be used. Other than this, we can use spyder software which is a test suite of machine learning for instance order preservation.

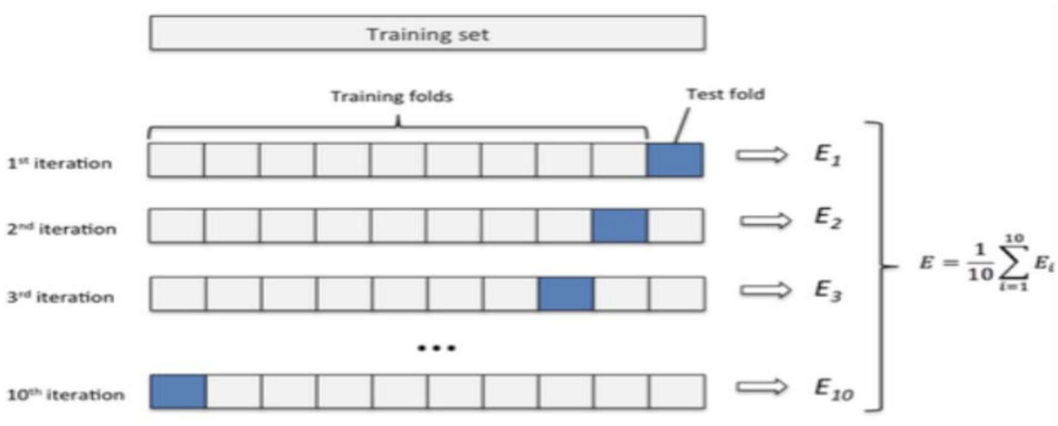


Figure 3.4, showing different iterations and training folds

3.5 MODEL DEPLOYMENT

Adjustment of training set and test set is done, and with the automated process the new data is obtained and added to the database either manually by end-user or automatically in matched one. With the calculations, new training sets are made with new predictions. The results are returned back to the end user. The learning model is also updated with the training set continuously with some time period and must receive input data dynamically.

4. ALGORITHMS

When it comes to predicting the outcome of football matches a large number of variables have to be taken into consideration and therefore it requires an algorithm which can relate all the variables in a manner such that the outcome that we get is the optimum one. For the specific problem of predicting a match result given a set of variables including the previous match statistics of all the teams one could make use of ANN too. A large number of training set examples are required for Deep Learning and hence the name, otherwise good results won't come. In sport predictions the number of features available is way much larger than the training examples that are present. For this problem the data present consists of previous season statistics of all the matches containing the amount of goals scored by each team and other particular attributes.

To start the prediction process, identification is the first step. The models that learn, datasets, model evaluation process and specific challenges that could very well hinder the process need to be identified.

There are two types of ML algorithms:

1. **Supervised Learning**
2. **Unsupervised Learning**

Just like its name, in supervised learning particular sets of input and output data are present. While, in unsupervised learning grouping of the data is done followed by learning through input data only. To predict the outcome of a football match supervised learning is the better option as it takes in account the results and statistics of previous fixtures too. The results of previous fixtures and the statistics of those fixtures act as the i/p-o/p pair for supervised learning model.

Following are the three algorithms that are used for predicting the outcome of a specific football fixture:

1. **Logistic Regression**
2. **Support Vector Machine (SVM)**
3. **XGBoost**

4.1 Logistic Regression

Logistic regression is based on the usage of a logistic function. It is a statistical model. What this logistic function does is that it models a binary variable. Also if we consider Regression Analysis then Logistic Regression can be termed as binomil regression, used primarily for the estimation of the parametrs of the logistic model. In terms of math, the dependent variable of the model can attain two values 1 or 0 and these values are indicated by an indicator variable. Also for the variables having value 1, if combined make up for the logarithm of odds. A variable which is independent of others can itself be of two minds, i.e., continuous or binary where binary means segregation of two different labels and the value stored in the dependent variable is indicated by the indicator variable. Unit logit is used by the log of odds which means "**log**istic un**it"** for the purpose of measurement.

Basically, Logsitic Regression is used for probability calculation. This probability is related to the occurrence of an event. The data given as input to the models is fitted to the Logistic curve for prediction. For e.g., the chance of someone having a cardiac arrest could be determind by various attributes of that very person such as, age, sex, body to mass index, etc. Rather than just being used in the sports prediction, it is also used in marketing and by businesses to predict the chance of a person buying a certain product.

4.1.1 Logistic Model

To know the working of Logistic Regression, primarily the consideration of the arguments given to the model is done and then coeff. Estimation is done by the help of the data that is given. Now suppose a model with two variabls, x1 and x2. The given variables can either be continuous or indicator functions for variables that are binary. The log-odds or the Logarithm of odds (denoted by *l*) can be written as below:

$$l = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

By explaining the basic Log Function, Logistic Regression could be understood better. The logistic function that is being talked about is nothing but a sigmoid function. Here, input is *t* such that $t \in \mathbf{R}$, and output is received is between 1 and 0.

$$\sigma(t) = e^t / (e^t + 1) = 1 / (1 + e^{-t})$$

The logistic function where $-6 < t < 6$ is represented graphically in the figure 4.1.

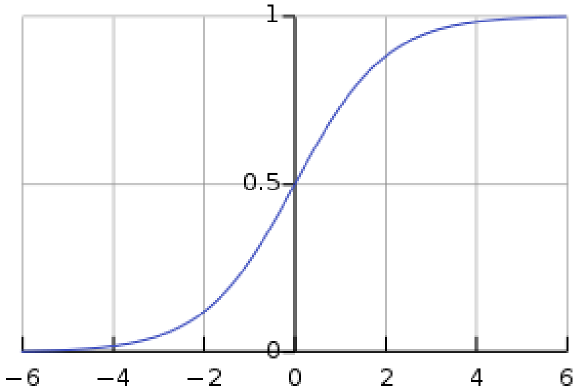


Figure 4.1, showing the Logistic curve

x here is a variable which is explanatory therefore, it could be supposed that *t* is a linear function of the *x* and hence, *t* is represented as follows:

$$t = \beta_0 + \beta_1 x$$

The function now changes to:

$$p(x) = 1 / (1 + e^{-(\beta 0 + \beta 1x)})$$

What is worth noting here is that the ***p(x)*** is written as the probability of the dependent variable and it gives only the success rate rather than the failure rate.

Logistic Function if inversed can be written as:

$$g(p(x)) = logit\, p(x) = ln(p(x) / (1 - p(x))) = \beta_0 + \beta_1 x$$

Now, exponenting both the LHS and RHS, we get:

$$p(x) / (1 - p(x)) = e^{\beta 0 + \beta 1x}$$

4.1.2 Odds

When we calculate the exponential function of logistic regression then we can almost equate the equivalency of the odds of the variable and the exponential function. A particular function which is knows as Logit function could be used to link the probability and linear expression function. Logistic Regression could also be apllied on the with the help of conversion of Logit to odds ratio.

In terms of math, dependent variable’s odds can be found out as shown below:

$$Odds = e^{\beta 0 + \beta 1x}$$

The odds ratio is calculated below:

$$OR = odds(x + 1) / odds(x) = e^{\beta 0 + \beta 1(x + 1)} / e^{\beta 0 + \beta 1x} = e^{\beta 1}$$

4.2 Support Vector Machine

The Support Vector Machines (**SVMs** or also sometimes known as the **Support Vector Network**) are supervised learning models. Support Vector Machines have with themselves some associated learning algorithms which could be used for two things i.e., Classification and Regression Analysis. Basic functionality of SVM is that it does the categorization of the input data and any other new data when comes it puts it into already formed categories. This is achieved by plotting the data points on the graph and sperating them using hyper-plane such that there is enough and visible gap between the two classes. Any new data is then mapped according to its attributes in its particular class without affecting the other class in any way.

Linear data is easily classified but for classification of non-linear data SVM maps the input data’s high-dimensional space. This is done by SVM using **Kernel Trick**.

SVM uses the n-dimensional space to plot the data points according to the attributes (where number of featues is represented by the variable ‘n’). The coordinates of each datapoint tells us about its value. Hence, the classification’s achieved by a hyper-plane that is created to differentiate among two different categories.

Figure 4.2 is the perfect representation of the n-dimensional data space with datapoints plotted on it and a hyper-plane separating them. Coordinates of each datapoint are represented by support vectors.

Mostly, SVM gives a time complexity of O(n^3) and a space complexity of O(n^2). The training set size can further be reduced with the help of decrementation of different aspects given.

METHOD THEORY

For non-linear case, we use a non-linear function to map data with higher dimensional space, next to create a custom hyper plane in space, so following function can be defined as follows:

$$f(x) = Sgn\left[\sum_{i=1}^{n} \alpha_i y_i < \varphi(x_i), \varphi(x) > +b\right]$$

using some inner product for optimal classification, we transform the function into a new further advanced and optimised function as follows:

$$W(\alpha) = \sum_{i=1}^{1} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{1} \alpha_i \alpha_j y_i K(x_i, x_j)$$

After some derivations and calculations, we get the modified version of comprehended classification function which is described as follows:

$$f(x) = Sgn\left[\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right]$$

here (b) is the classification threshold and the value is achievable from any one of the support vector.

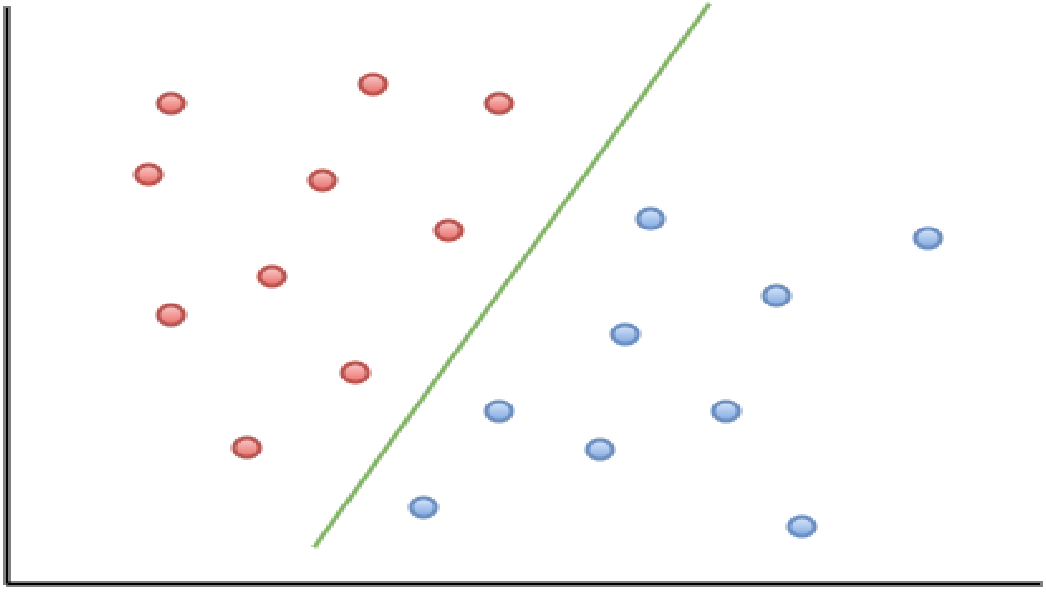


Figure 4.2, showing two classes differentiated by a hyper-plane

4.2.1 Hyper Planes

SVM’s main working depends on classification of the classes and classifying them in the best manner is not so easy of a task therefore, the hyper-plane has to be chosen in an optimum manner.

Now, to better understand the concept of hyper planes:

- **Scenario – 1:** The hyper plane B classifies the two given classes better than A or C.

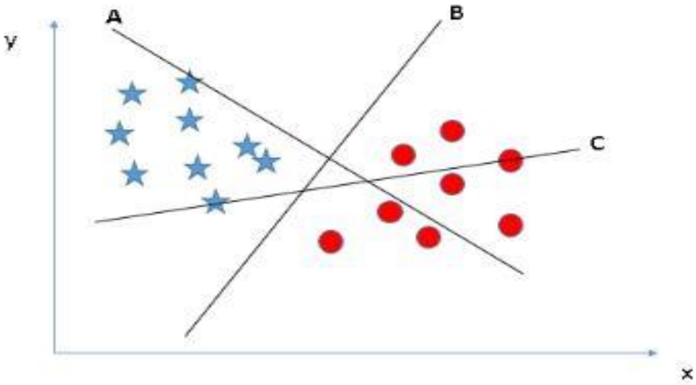


Figure 4.3, showing three different hyper-planes for scenario-1

- **Scenario – 2:** Here C does the best job as it keeps the largest gap with the two classes and therefore, is a better classification of the classes.

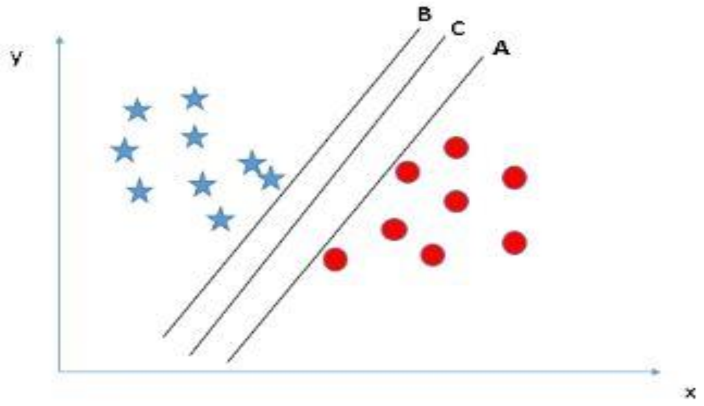


Figure 4.4, showing three different hyper-planes for scenario-2

- **Scenario – 3:** Following data is not linear therefore, it becomes difficult to classify them.

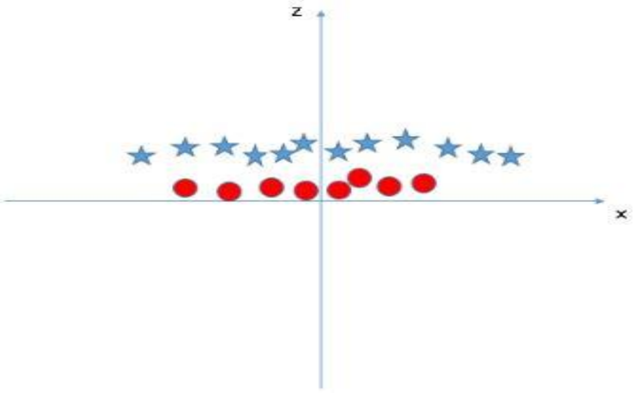


Figure 4.5, showing a non-linear hyper-plane that differentiates the two classes

SVM can even classify non-linear data and for this purpose it adds a new feature to the space i.e., the z-axis where, $z = x^2 + y^2$

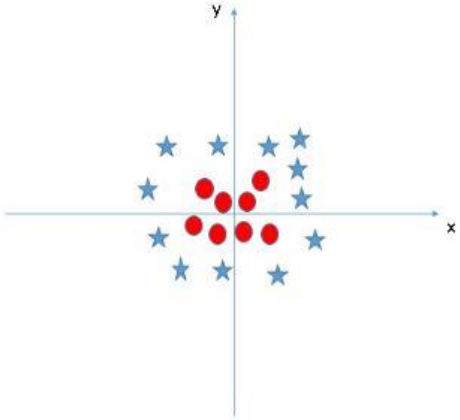


Figure 4.6, showing the addition of a new feature

Kernel trick solves this problem :

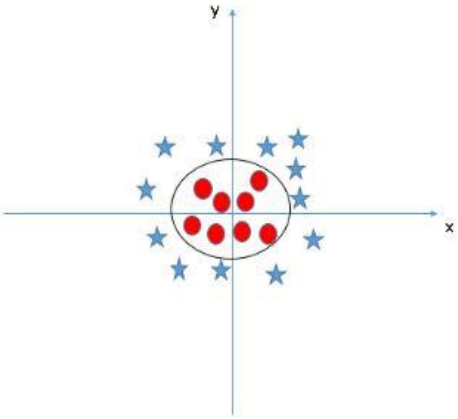


Figure 4.7, showing a new hyper-plane created using a Kernel trick

SVM TESTING AND TRAINING FRAMEWORK

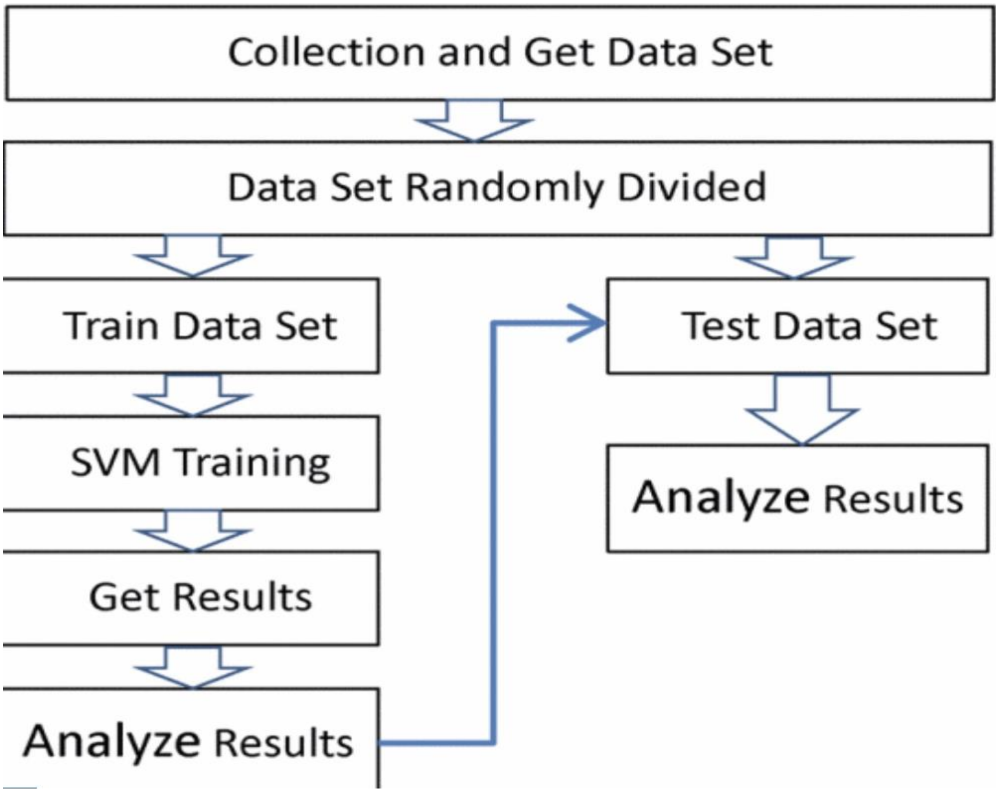


Figure 4.8, showing an SVM training and testing framework

4.3 XGBoost

XGBoost is an applied Machine Learning algorithm which comes under the category of ensemble learning. It uses a certain process in which, numerous deviants of dataset are created and different prediction techniques are combined to predict the reuslts.

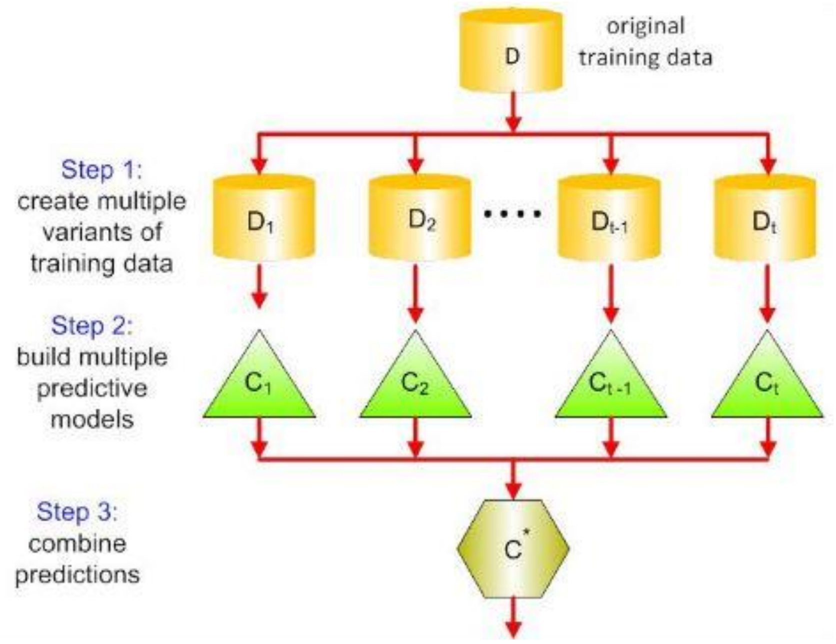

D
original training data
Step 1: create multiple variants of training data
D_1
D_2
....
D_{t-1}
D_t
Step 2: build multiple predictive models
C_1
C_2
C_{t-1}
C_t
Step 3: combine predictions
C^*

Figure 4.9, showing how XGBoost works

Base Learners that are the models which form the ensemble could be either from a different learning algorithm or even from the same learning algorithm. The two of the ensemble learners that are most widely used are Boosting and Bagging. Several statistical models can use these techniques but, decision trees make use of them the most.

4.3.1 Bagging

While choice trees are a standout amongst the most effortlessly interpretable models, they display profoundly factor conduct. Consider a solitary preparing dataset that we haphazardly split into two sections. Presently, we should utilize each part to prepare a choice tree with the end goal to get two models.

While fitting of these models, they would yield diverse outcomes. Choice trees are said to be related with high fluctuation because of this conduct. Sacking or boosting accumulation lessens the change in any learner. A few choice trees which are created in parallel, frame the base learners of stowing method.

4.3.2 Boosting

In boosting, the trees are constructed successively with the end goal that each ensuing tree expects to lessen the blunders of the past tree. Each tree gains from its forerunners and updates the leftover mistakes. Henceforth, the tree that becomes next in the grouping will gain from a refreshed rendition of the residuals.

As opposed to packing strategies like Random Forest, in which trees are developed to their most extreme degree, boosting makes utilization of trees with less parts. Such little trees, which are not profound, are very interpretable. Parameters like the quantity of trees or cycles, the rate at which the angle boosting learns, and the profundity of the tree, could be ideally chosen through approval strategies like k-overlap cross approval. Having countless may prompt overfitting. In this way, it is important to deliberately pick the halting criteria for boosting.

Boosting comprises of three basic steps:

- An underlying model F0 is characterized to foresee the objective variable y. This model will be related with a lingering (y – F0)
- Another model h1 is fit to the residuals from the past advance
- Presently, F0 and h1 are consolidated to give F1, the supported adaptation of F0. The mean squared blunder from F1 will be lower than that from F0:

  $$F_1(x) <\text{-}\ F_0(x) + h_1(x)$$

- To enhance the execution of F1, we could demonstrate after the residuals of F1 and make another model F2:

  $$F_2(x) <\text{-}\ F_1(x) + h_2(x)$$

- This should be possible for 'm' iterations, until the point when residuals have been limited however much as could reasonably be expected:

  $$F_m(x) <\text{-}\ F_{m-1}(x) + h_m(x)$$

Here, the added substance learners don't do anything to the capacities made in the past iterations. Rather, they bestow data of their own to cut down the blunders.

The main working of tree is based on the score on each of the leaf, which is unlike the regular decision trees.

The final prediction is summed up and the final score is evaluated. This final score helps in selection of Ensemble Model of the specific trees and the correctness and precision is maintained with Gradient tree boosting.

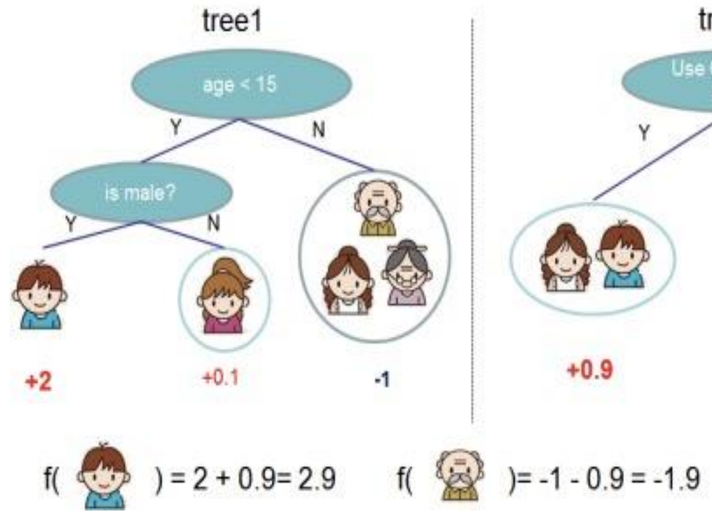


Figure 4.10, showing the Tree Ensemble Model. Calculation of score determines the tree to be chosen.

For calculation of the final score of different trees, we sum up the gradient and second level gradient stats on each and every leaf. For obtaining the quality score, the resultant score from all leaves is evaluated.

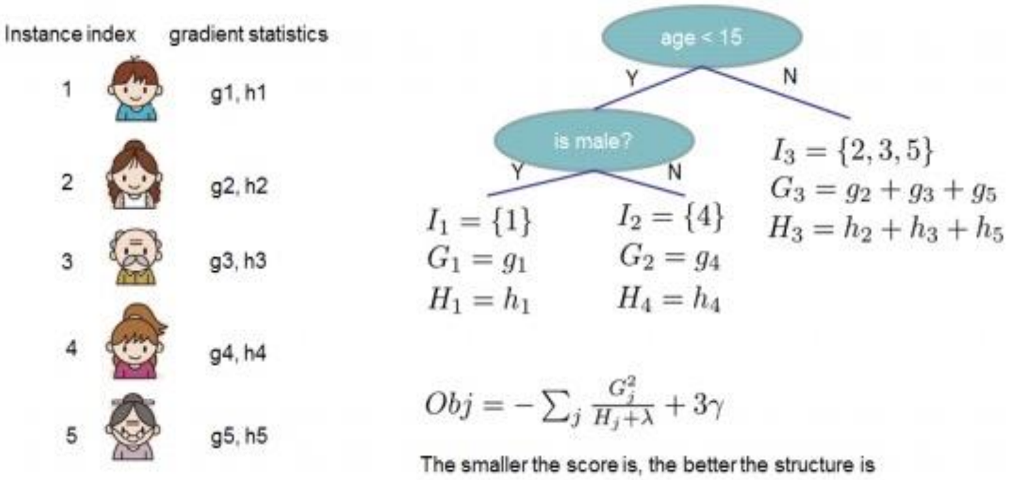


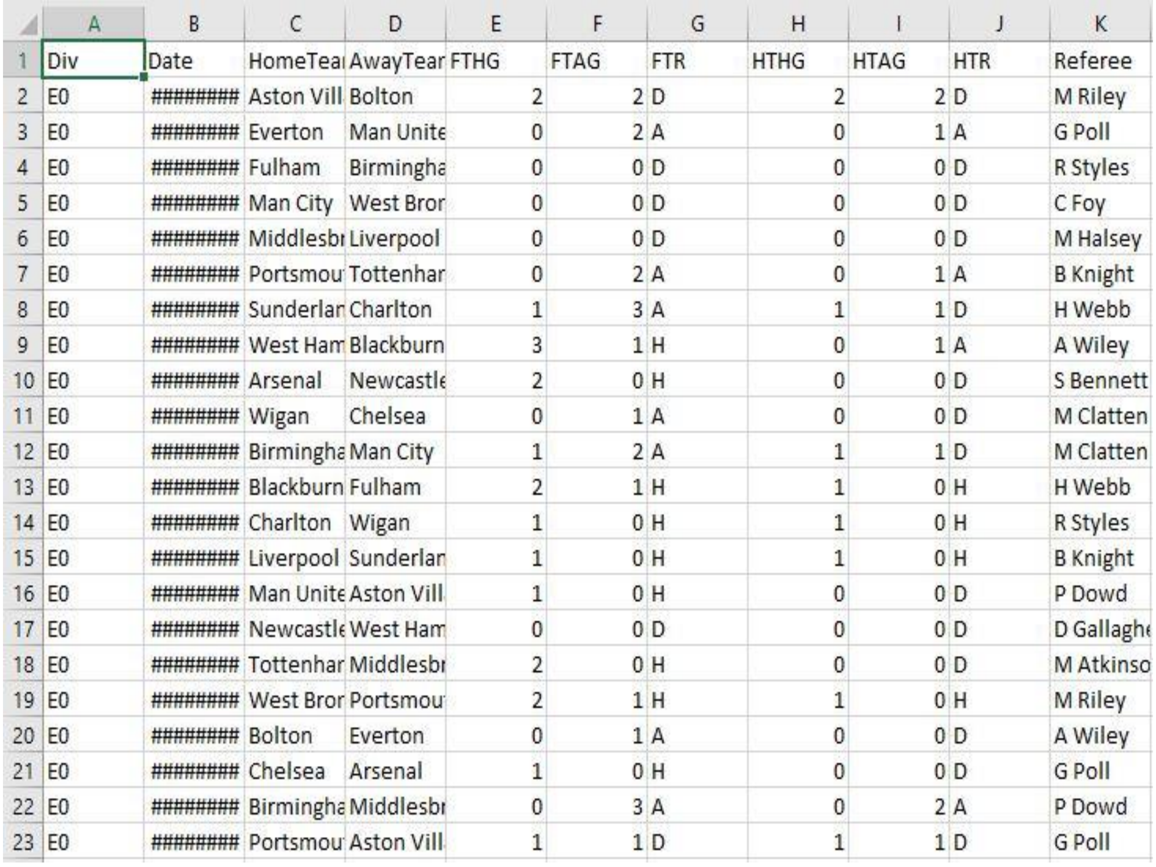
Figure 4.11, showing the structure score calculation.

5. TEST PLAN

5.1 Data Set

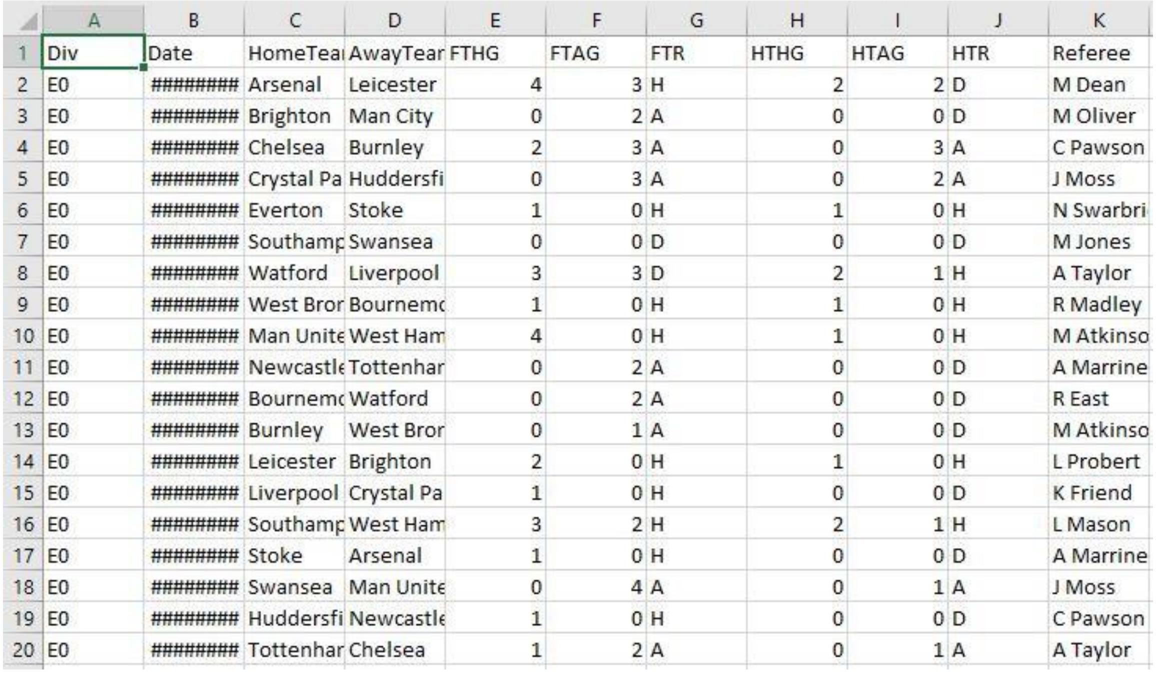
Football is uncertain in its own way and therefore, a lot of features need to be considered to correctly predict the outcome of a ficture. A number of factors can decide a game of football for ex, Home Team always has better chances of winning but there chances could be lowered if the away team is leading the league table or is in a very good run of form say, 5 or 10 matches gone unbeaten. Therefore, all these factors have to be considered to train the models for better predictions. To make the predictions even better huge amount of data has to be fed to the learning models and therefore, to achieve this the datasets of previous 12 Premier League season were collected.

The data set screenshot of the Premier League season 2005-06 is shown in the figure 5.1 and the data set of the Premier League season 2017-18 is shown in the figure 5.2.



	A	B	C	D	E	F	G	H	I	J	K
1	Div	Date	HomeTeai	AwayTear	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee
2	E0	########	Aston Vill	Bolton	2	2	D	2	2	D	M Riley
3	E0	########	Everton	Man Unite	0	2	A	0	1	A	G Poll
4	E0	########	Fulham	Birmingha	0	0	D	0	0	D	R Styles
5	E0	########	Man City	West Bror	0	0	D	0	0	D	C Foy
6	E0	########	Middlesbr	Liverpool	0	0	D	0	0	D	M Halsey
7	E0	########	Portsmou	Tottenhar	0	2	A	0	1	A	B Knight
8	E0	########	Sunderlar	Charlton	1	3	A	1	1	D	H Webb
9	E0	########	West Ham	Blackburn	3	1	H	0	1	A	A Wiley
10	E0	########	Arsenal	Newcastle	2	0	H	0	0	D	S Bennett
11	E0	########	Wigan	Chelsea	0	1	A	0	0	D	M Clatten
12	E0	########	Birmingha	Man City	1	2	A	1	1	D	M Clatten
13	E0	########	Blackburn	Fulham	2	1	H	1	0	H	H Webb
14	E0	########	Charlton	Wigan	1	0	H	1	0	H	R Styles
15	E0	########	Liverpool	Sunderlar	1	0	H	1	0	H	B Knight
16	E0	########	Man Unite	Aston Vill	1	0	H	0	0	D	P Dowd
17	E0	########	Newcastle	West Ham	0	0	D	0	0	D	D Gallaghe
18	E0	########	Tottenhar	Middlesbr	2	0	H	0	0	D	M Atkinso
19	E0	########	West Bror	Portsmou	2	1	H	1	0	H	M Riley
20	E0	########	Bolton	Everton	0	1	A	0	0	D	A Wiley
21	E0	########	Chelsea	Arsenal	1	0	H	0	0	D	G Poll
22	E0	########	Birmingha	Middlesbr	0	3	A	0	2	A	P Dowd
23	E0	########	Portsmou	Aston Vill	1	1	D	1	1	D	G Poll

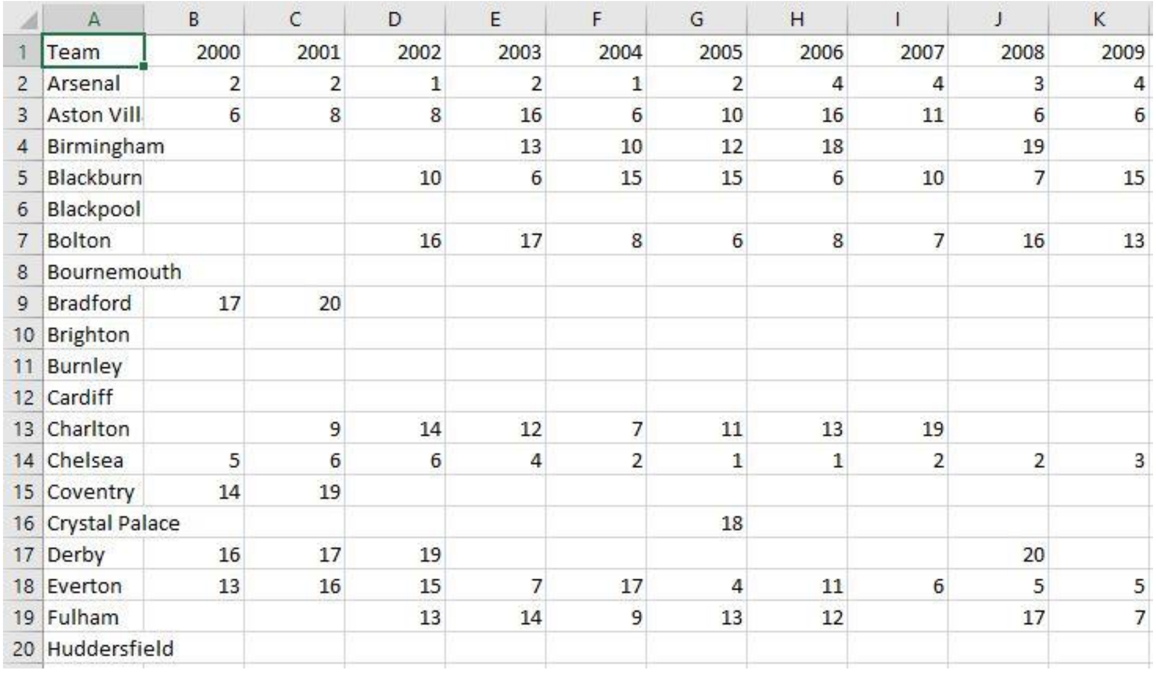
Figure 5.1, showing the dataset of Premier League season 2005/06



	A	B	C	D	E	F	G	H	I	J	K
1	Div	Date	HomeTeai	AwayTear	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee
2	E0	########	Arsenal	Leicester	4	3	H	2	2	D	M Dean
3	E0	########	Brighton	Man City	0	2	A	0	0	D	M Oliver
4	E0	########	Chelsea	Burnley	2	3	A	0	3	A	C Pawson
5	E0	########	Crystal Pa	Huddersfi	0	3	A	0	2	A	J Moss
6	E0	########	Everton	Stoke	1	0	H	1	0	H	N Swarbri
7	E0	########	Southamp	Swansea	0	0	D	0	0	D	M Jones
8	E0	########	Watford	Liverpool	3	3	D	2	1	H	A Taylor
9	E0	########	West Bror	Bournemc	1	0	H	1	0	H	R Madley
10	E0	########	Man Unite	West Ham	4	0	H	1	0	H	M Atkinso
11	E0	########	Newcastle	Tottenhar	0	2	A	0	0	D	A Marrine
12	E0	########	Bournemc	Watford	0	2	A	0	0	D	R East
13	E0	########	Burnley	West Bror	0	1	A	0	0	D	M Atkinso
14	E0	########	Leicester	Brighton	2	0	H	1	0	H	L Probert
15	E0	########	Liverpool	Crystal Pa	1	0	H	0	0	D	K Friend
16	E0	########	Southamp	West Ham	3	2	H	2	1	H	L Mason
17	E0	########	Stoke	Arsenal	1	0	H	0	0	D	A Marrine
18	E0	########	Swansea	Man Unite	0	4	A	0	1	A	J Moss
19	E0	########	Huddersfi	Newcastle	1	0	H	0	0	D	C Pawson
20	E0	########	Tottenhar	Chelsea	1	2	A	0	1	A	A Taylor

Figure 5.2, showing the dataset of Premier League season 2017/18

Figure 5.3 shows the league standing of all the teams that have participated in atleast one Premier League since 2000 year wise. Blank spaces represent the absence of the corresponding team in that year's Premier League season.



	A	B	C	D	E	F	G	H	I	J	K
1	Team	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
2	Arsenal	2	2	1	2	1	2	4	4	3	4
3	Aston Vill	6	8	8	16	6	10	16	11	6	6
4	Birmingham				13	10	12	18		19	
5	Blackburn			10	6	15	15	6	10	7	15
6	Blackpool										
7	Bolton			16	17	8	6	8	7	16	13
8	Bournemouth										
9	Bradford	17	20								
10	Brighton										
11	Burnley										
12	Cardiff										
13	Charlton		9	14	12	7	11	13	19		
14	Chelsea	5	6	6	4	2	1	1	2	2	3
15	Coventry	14	19								
16	Crystal Palace						18				
17	Derby	16	17	19						20	
18	Everton	13	16	15	7	17	4	11	6	5	5
19	Fulham			13	14	9	13	12		17	7
20	Huddersfield										

Figure 5.3, showing the dataset of each team's standings in seasons from 2000 to 2018

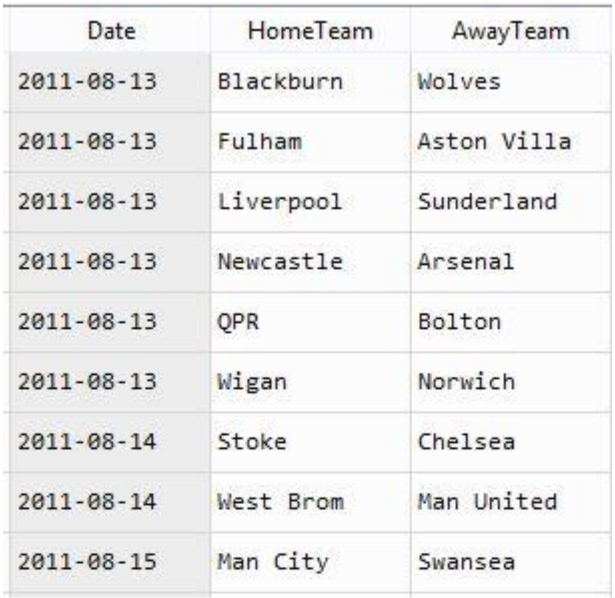
5.2 Metrics

The dates of all the match days in different datasets of different seasons were given in different formats therefore, they have to be converted into a similar single format so that they could be used whenever need be.

Following code in Python was used to do so.

```
FMT = "%d/%m/%y"
FMT1 = "%d/%m/%Y"
def parse_date(date):
    if date == '':
        return None
    else:
        return dt.strptime(date, FMT).date()
def parse_date_other(date):
    if date == '':
        return None
    else:
        return dt.strptime(date, FMT1).date()
```

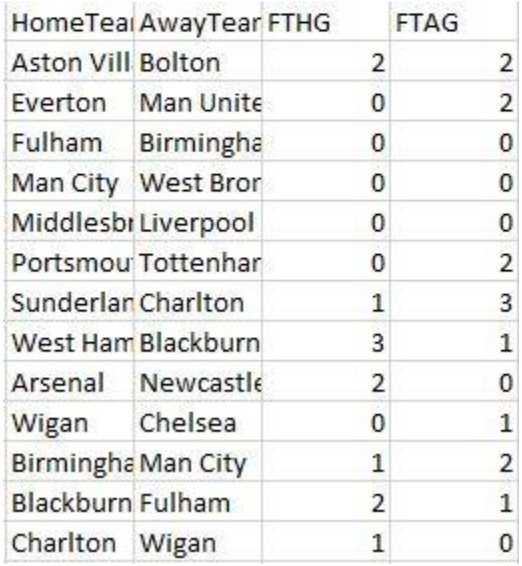
Finally, all the dates were converted to format YYYY-MM-DD as shown in the figure 5.4.



Date	HomeTeam	AwayTeam
2011-08-13	Blackburn	Wolves
2011-08-13	Fulham	Aston Villa
2011-08-13	Liverpool	Sunderland
2011-08-13	Newcastle	Arsenal
2011-08-13	QPR	Bolton
2011-08-13	Wigan	Norwich
2011-08-14	Stoke	Chelsea
2011-08-14	West Brom	Man United
2011-08-15	Man City	Swansea

Figure 5.4, showing the data format for fixtures

The goals scored are represented numerically with number equals to the exact number of goals scored by a particular team as shown in the figure 5.5.



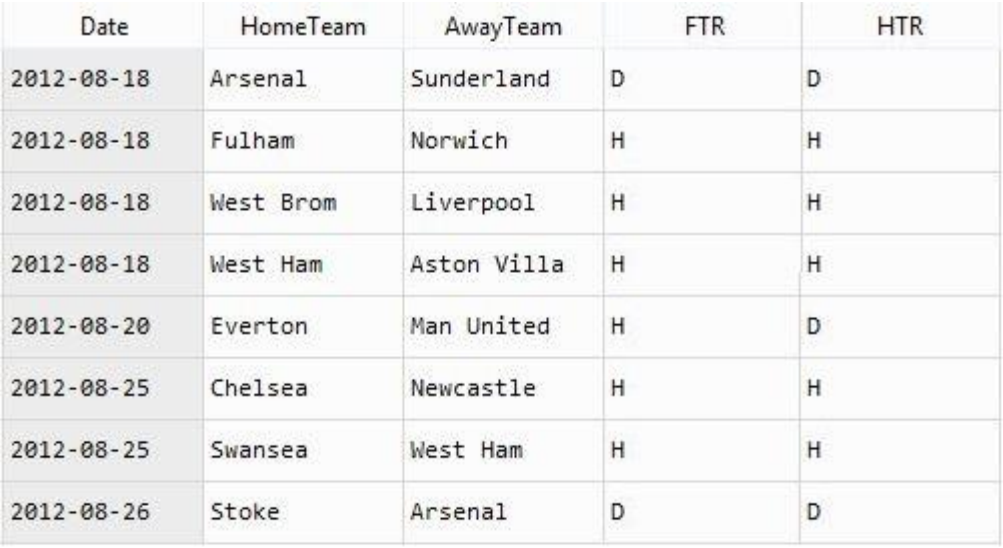
HomeTeai	AwayTear	FTHG	FTAG
Aston Vill	Bolton	2	2
Everton	Man Unite	0	2
Fulham	Birmingha	0	0
Man City	West Bror	0	0
Middlesbı	Liverpool	0	0
Portsmou	Tottenhar	0	2
Sunderlar	Charlton	1	3
West Ham	Blackburn	3	1
Arsenal	Newcastle	2	0
Wigan	Chelsea	0	1
Birmingha	Man City	1	2
Blackburn	Fulham	2	1
Charlton	Wigan	1	0

Figure 5.5, showing the goals scored at Full-time by home and away teams

In a football match either the home team wins, loses or draws the match and keeping in mind these three parameters following metrics were chosen to represent the result of a match.

The haft-time and full-time result of the match is represented as:

- ‘H’ if the home team is winning,
- ‘A’ if the away team is winning or
- ‘D’ if the match is drawing as shown in the figure 5.6.

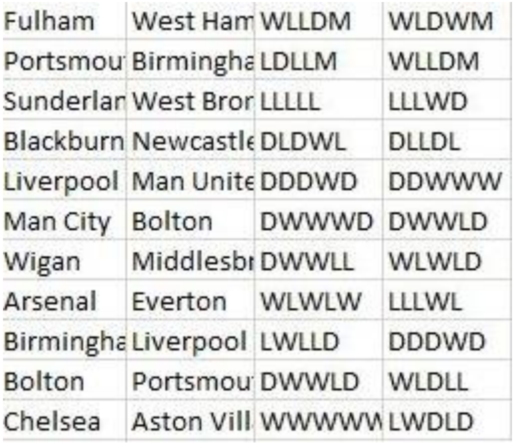


Date	HomeTeam	AwayTeam	FTR	HTR
2012-08-18	Arsenal	Sunderland	D	D
2012-08-18	Fulham	Norwich	H	H
2012-08-18	West Brom	Liverpool	H	H
2012-08-18	West Ham	Aston Villa	H	H
2012-08-20	Everton	Man United	H	D
2012-08-25	Chelsea	Newcastle	H	H
2012-08-25	Swansea	West Ham	H	H
2012-08-26	Stoke	Arsenal	D	D

Figure 5.6, showing the full-time and half-time result of a fixture

Another feature that can affect the outcome of a football match is the form of the team. Even if the team is bottom of the table but has a very good form on going let’s say, 10 matches unbeaten then that team is highly likely to come with at least a point from the current match. Therefore, following metrics were used to define the form of a given team. The form represented here displays the results of the last 5 matches that the team has played.

- ‘W’ is used if the team has won a match,
- ‘D’ is used if the team has drawn the match,
- ‘L’ is used if the team has lost the match,
- ‘M’ is used if the result of a match is not known i.e., the team hasn’t played that match yet.



Fulham	West Ham	WLLDM	WLDWM
Portsmou	Birmingha	LDLLM	WLLDM
Sunderlan	West Bror	LLLLL	LLLWD
Blackburn	Newcastle	DLDWL	DLLDL
Liverpool	Man Unite	DDDWD	DDWWW
Man City	Bolton	DWWWD	DWWLD
Wigan	Middlesbr	DWWLL	WLWLD
Arsenal	Everton	WLWLW	LLLWL
Birmingha	Liverpool	LWLLD	DDDWD
Bolton	Portsmou	DWWLD	WLDLL
Chelsea	Aston Vill	WWWWWW	LWDLD

Figure 5.7, showing the last 5 match form of home and away teams

5.3 Test Setup and Plan

1. The data of the seasons from 2005-06 to 2017-18 were collected as have already been shown in the figures 5.1, figure 5.2 and figure 5.3.
2. The data had to be made fit so that it could be fed to learning models and data pre-processing techniques were used to do so:
    a. The date formats of all the datasets were made similar.
    b. The goals scored and conceded by all teams were calculated by the following code.

```
def get_goals(playing_statistics):
    GC = get_goalsConceded(playing_statistics)
    GS = get_goalsScored(playing_statistics)
    j = 0
    HTGS = []
    ATGS = []
    HTGC = []
    ATGC = []

    for i in range(380):
        ht = playing_statsistic.iloc[i].HomeTeam
        at = playing_statistics.iloc[i].AwayTeam
        HTGS.append(GS.loc[ht][j])
        ATGS.append(GS.loc[at][j])
        HTGC.append(GC.loc[ht][j])
        ATGC.append(GC.loc[at][j])

        if ((i + 1)% 10) == 0:
            j = j + 1

    playing_statsistic ['HTGS'] = HTGS
    playing_statsistic ['ATGS'] = ATGS
    playing_statsistic ['HTGC'] = HTGC
    playing_statsistic ['ATGC'] = ATGC

    return playing_statsistic
```

c. Aggregate points of the teams were calculated as below:

```
def get_points(match_result):
    match_result_points = match_result.applymap(get_points)
    for i in range(2,39):
        match_result_points[i] = match_result_points[i] + match_result_points[i-1]

    match_result_points.insert(column =0, loc = 0, value = [0*i for i in range(20)])
```

```
        return match_result_points
```

d. Form of the teams was calculated as below:

```
def get_form(playing_statistics,num):
    form = get_match_result(playing_statistics)
    final_form = form.copy()
    for i in range(num,39):
        form_final[i] = "
        j = 0
        while j < num:
            final_form [i] = final_form [i] + form[i-j]
            j = j + 1
    return final_form
```

e. Finally, the scraped and cleaned datasets were put together into one data frame. This final data frame was then saved into a csv file.

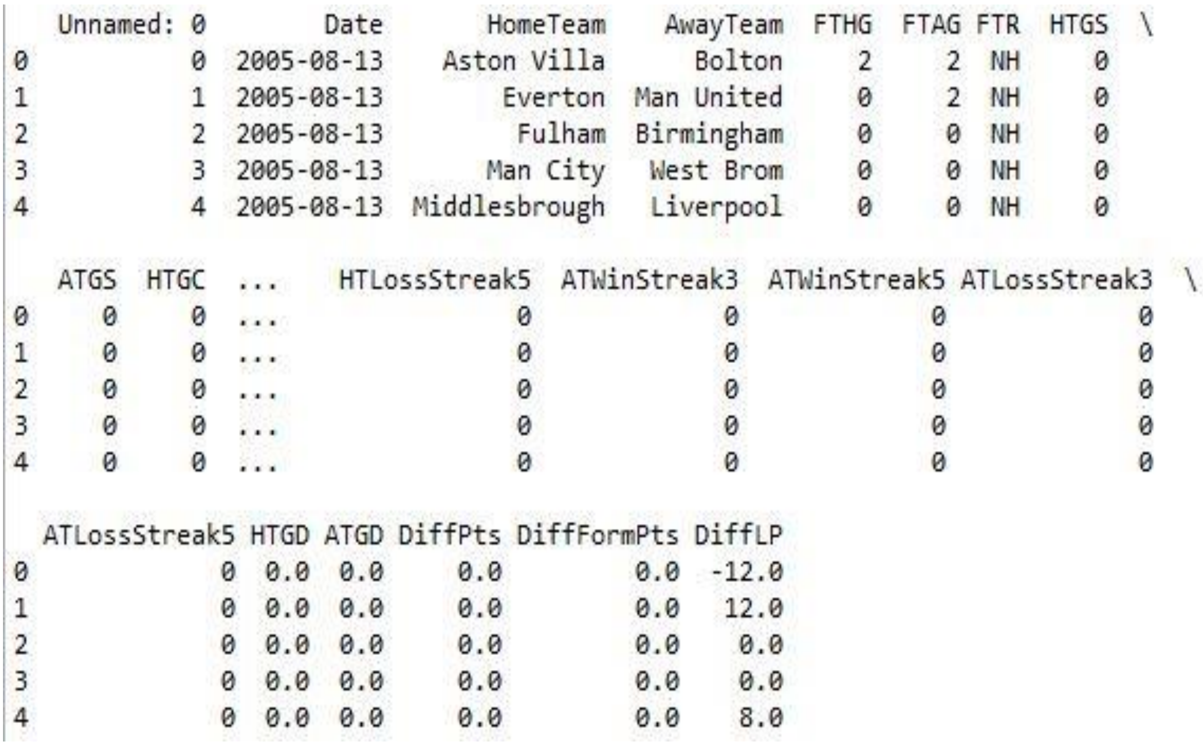
3. Last data frame was split into two sets i.e., training and testing. They have 12 features and 1 target i.e., the winning team(Home (H)/Not Home (NH)).
4. The three algorithms described in chapter 4 were then fed the whole data and made to learn the trends.
5. The three algorithms were then evaluated and compared.
6. Logistic Regression came as the standout performer which is explained how in the Section 6, Results and Performance.
7. The dataset of the current Premier League season i.e. the 2018/19 was obtained.
8. Lastly for the application purpose the best model i.e. Logistic Regression is used to make predictions for the 38th matchweek of Premier League season 2018/19
9. User can input the names of any two teams from the 2018/19 season and the result provided will be the predicted outcome of the match along the with the probability of that team winning the match.

6. RESULTS AND PERFORMANCE ANALYSIS

6.1 Import Dependencies

In this section, that final dataset and all the important modules that will be used are imported.

```
loc = "C:/Users/Sushant Vashisht/Desktop/4th Year Project/Datasets/"

data = pd.read_csv(loc + 'final_dataset.csv')

display(data.head())
```



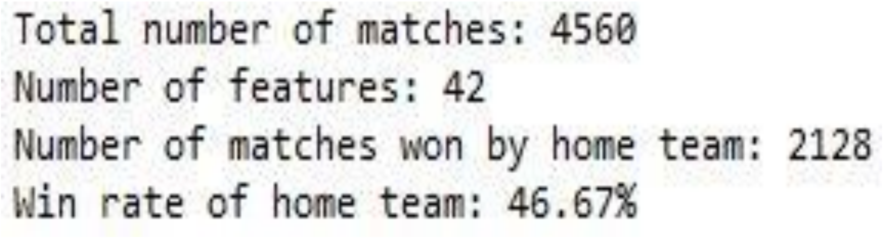
```
   Unnamed: 0        Date      HomeTeam     AwayTeam  FTHG  FTAG FTR  HTGS  \
0           0  2005-08-13   Aston Villa       Bolton     2     2  NH     0
1           1  2005-08-13       Everton   Man United     0     2  NH     0
2           2  2005-08-13        Fulham   Birmingham     0     0  NH     0
3           3  2005-08-13      Man City    West Brom     0     0  NH     0
4           4  2005-08-13  Middlesbrough    Liverpool     0     0  NH     0

   ATGS  HTGC  ...    HTLossStreak5  ATWinStreak3  ATWinStreak5  ATLossStreak3  \
0     0     0  ...                0             0             0              0
1     0     0  ...                0             0             0              0
2     0     0  ...                0             0             0              0
3     0     0  ...                0             0             0              0
4     0     0  ...                0             0             0              0

  ATLossStreak5 HTGD ATGD DiffPts DiffFormPts DiffLP
0             0  0.0  0.0     0.0         0.0  -12.0
1             0  0.0  0.0     0.0         0.0   12.0
2             0  0.0  0.0     0.0         0.0    0.0
3             0  0.0  0.0     0.0         0.0    0.0
4             0  0.0  0.0     0.0         0.0    8.0
```

Figure 6.1, showing the first five rows of the main dataframe

6.2 Data Exploration

In this section, the total number of matches, total number of features, number of matches won by the home teams and the win rate are calculated.

```
n_matches = data.shape[0]

n_homewins = len(data[data.FTR == 'H'])

win_rate = (float(n_homewins) / (n_matches)) * 100

print ("Total number of matches: {}".format(n_matches))

print ("Number of features: {}".format(n_features))

print ("Number of matches won by home team: {}".format(n_homewins))

print ("Win rate of home team: {:.2f}%".format(win_rate))
```



```
Total number of matches: 4560
Number of features: 42
Number of matches won by home team: 2128
Win rate of home team: 46.67%
```

Figure 6.2, showing the win rate of the home team

For the matrix of scatter plots, we will import pandas.plotting.scatter_matrix. All the arguments will be configured based on user’s choice. Following is the code for scatter plotting of attributes with the figure for example:

```
from pandas.tools.plotting import scatter_matrix

scatter_matrix(data[['HTGD','ATGD','HTP','ATP','DiffFormPts','DiffLP']],

figsize=(10,10))
```

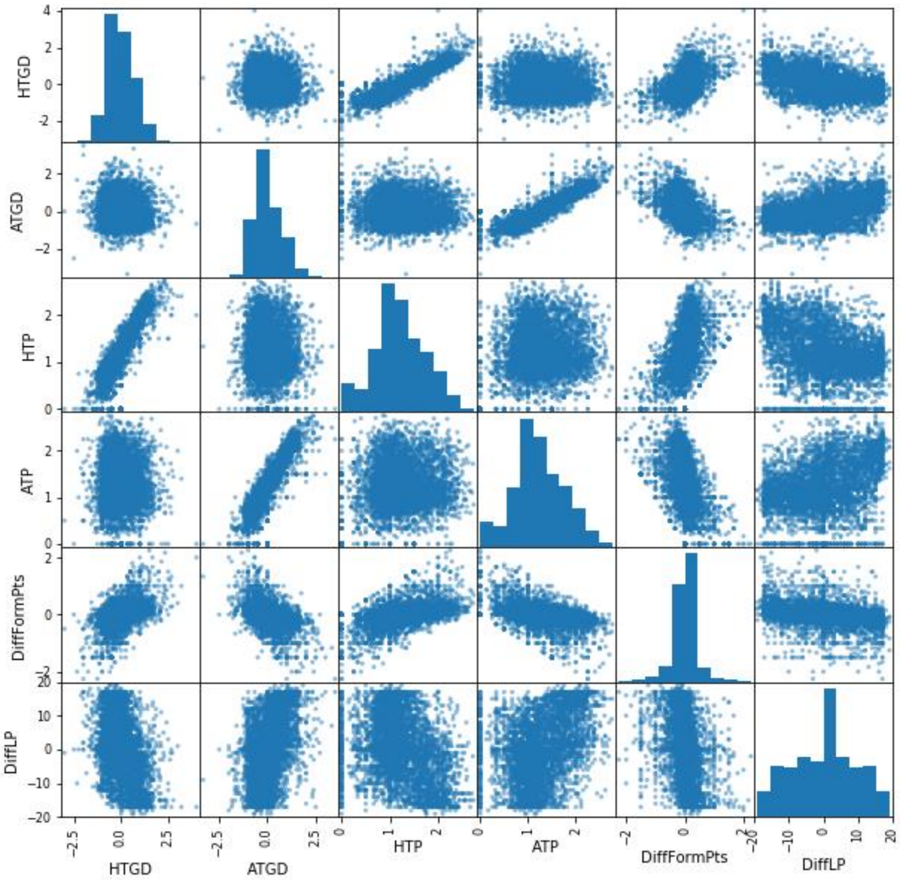


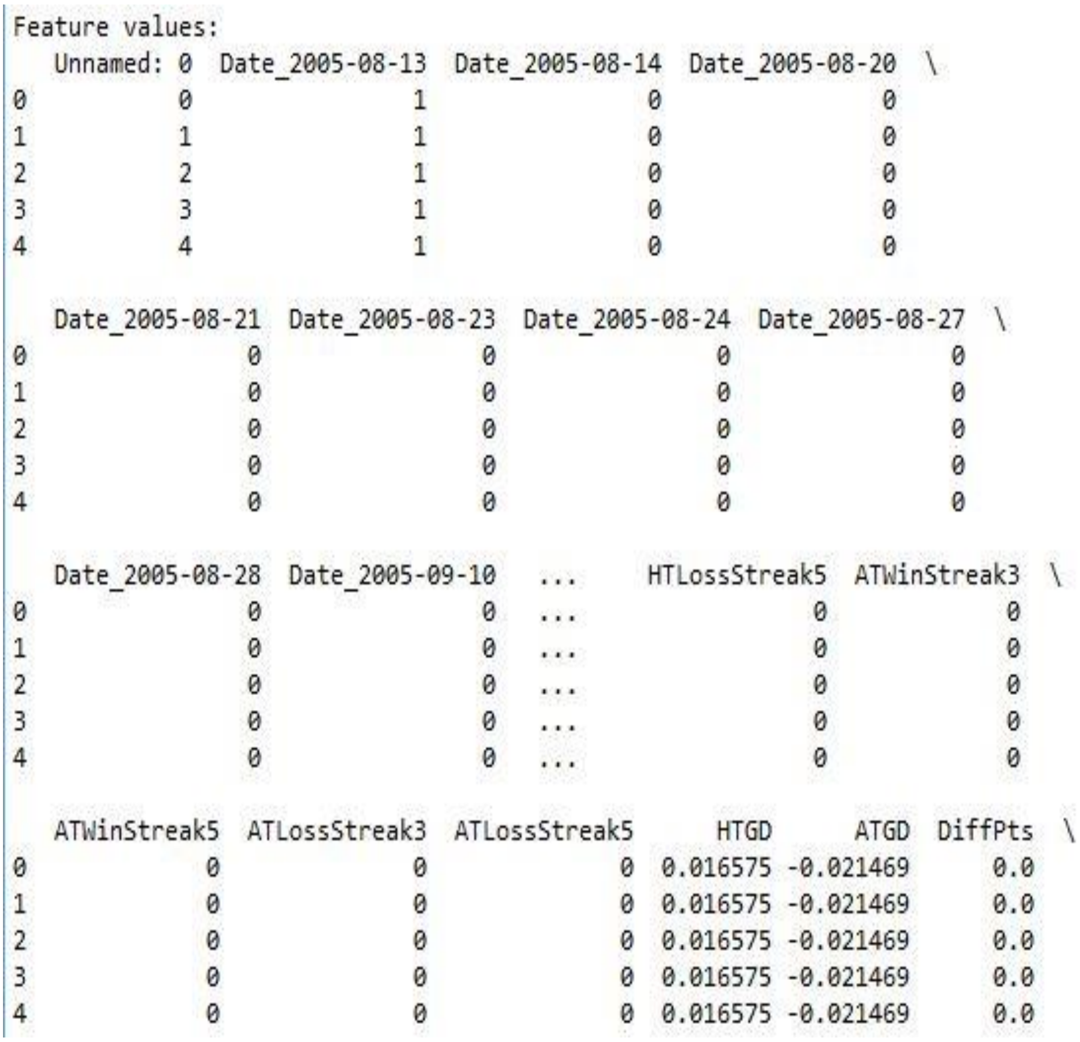
Figure 6.3, shows the plot of different attributes with one another

6.3 Preparing the Data

In this section, the data is prepared such that it could finally be splitted into testing and training datasets that will be used to train and test the models.

```
def preprocess_features(X):
    output = pdDataFrame(index = X.index)
    for col, colata in X.iteritems():
        if coldata.dtype == object:
            coldata = pd.get_dummies(col_data, prefix = col)
        output = output.join(col_data)
    return output
```

```
print ("\nFeatur values:")

display(X_all.hed())

from sklearn.cross_validation import traintestsplit

X_train, X_test, y_train, y_test = trai_test_split(X_all, y_all,

                                        testsize = 50,

                                        random_state = 2,

                                        stratiy = y_all)
```



```
Feature values:
   Unnamed: 0  Date_2005-08-13  Date_2005-08-14  Date_2005-08-20  \
0           0                1                0                0
1           1                1                0                0
2           2                1                0                0
3           3                1                0                0
4           4                1                0                0

   Date_2005-08-21  Date_2005-08-23  Date_2005-08-24  Date_2005-08-27  \
0                0                0                0                0
1                0                0                0                0
2                0                0                0                0
3                0                0                0                0
4                0                0                0                0

   Date_2005-08-28  Date_2005-09-10  ...     HTLossStreak5  ATWinStreak3  \
0                0                0  ...                 0             0
1                0                0  ...                 0             0
2                0                0  ...                 0             0
3                0                0  ...                 0             0
4                0                0  ...                 0             0

   ATWinStreak5  ATLossStreak3  ATLossStreak5      HTGD      ATGD  DiffPts  \
0             0              0              0  0.016575 -0.021469      0.0
1             0              0              0  0.016575 -0.021469      0.0
2             0              0              0  0.016575 -0.021469      0.0
3             0              0              0  0.016575 -0.021469      0.0
4             0              0              0  0.016575 -0.021469      0.0
```

Figure 6.4, shows the first five rows of the preprocessed dataframe

6.4 Training and Evaluating Models

In this section, the models that have been listed in the section 4, Algorithms, are finally initialized and used on our dataset.

```
from time import time

from sklearn.metrics import f1score

def train_classifier(clf, Xtrain, y_train):

    start = time()
    clf.fit(Xtrain, y_train)
    end = time()

    print ("Traned model in {:.4f} seconds".format(end - start))

def predictlabels(clf, features, target):

    start = time()
    y_pred = clf.predict(featres)
    display(ypred)
    end = time()

    print ("Made predictins in {:.4f} seconds.".format(end - start))

    return f1_score(targt, y_pred, pos_label='H'), sum(target == y_pred) /
float(len(y_pred)



def train_predit(clf, Xtrain, y_train, Xtest, y_test):

    print ("Training a {} using a training set size of {}. . ".format(clf._class__._name__,
len(X_train))
```

```
    train_classifer(clf, X_train, y_train)

    f1, acc = predictlabels(clf, X_train, y_train)

    print (f1, acc)

    print ("F1 score ad accuacy score for training set: {:.4f} , {:.4f}.".format(f1 , acc))

    f1, acc = predict_labels(lf, X_test, y_test)

    print ("F1 score and acuracy score for test set: {:.4f} , {:.4f}.".format(f1 , acc))

clf_A = LogisticRegresson(random_state = 42)

clf_B = SVC(random_sate = 912, kernel='rbf')

clf_C = xgb.XGBClassfier(seed = 82)

train_predict(clf_A, Xtrain, y_train, X_test, y_test)
print ''
train_predict(clf_B, Xtran, y_train, Xtest, y_test)
print ''
train_predict(clf_C, Xtrain, y_train, Xtest, y_test)
```

The confusion matrices for the training and testing set corresponding to the three models used are as follows:

1. Logistic Regression

- **Training Set**

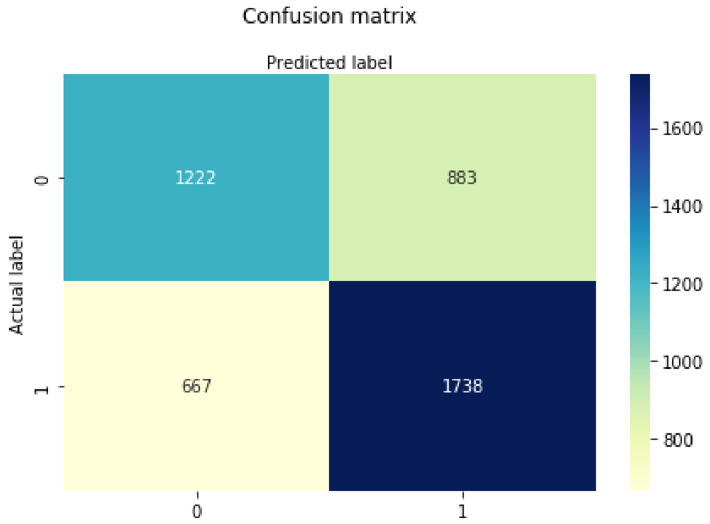


Figure 6.5, shows the confusion matrix for training set of Logistic Regression

- **Test Set**

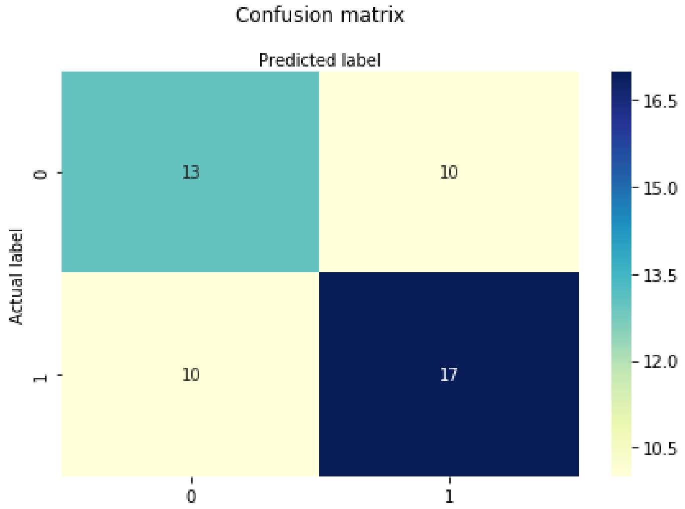


Figure 6.6, shows the confusion matrix for test set of Logistic Regression

2. Support Vector Machine

- **Training Set**

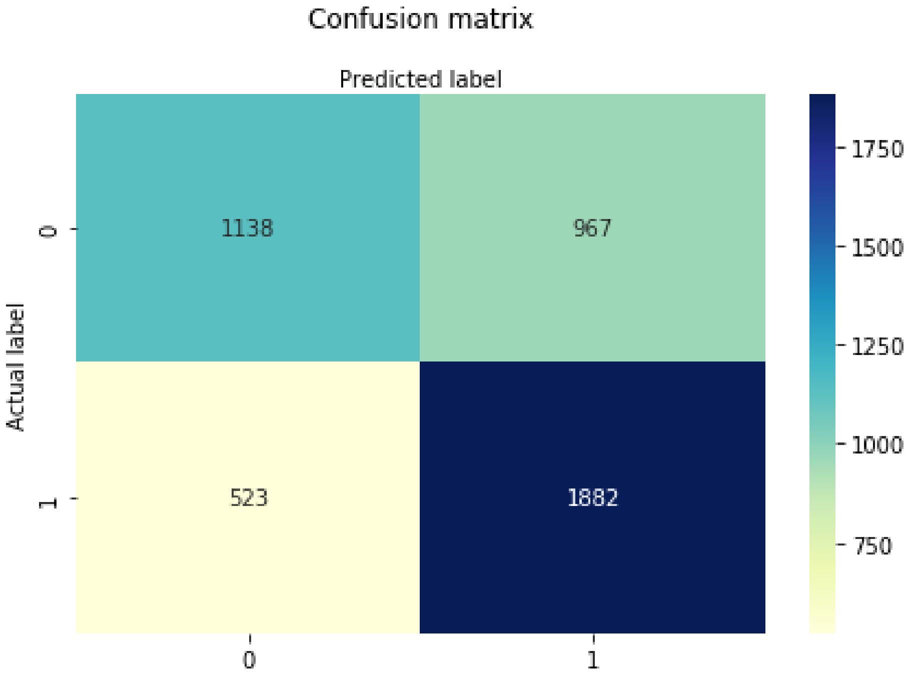


Figure 6.7, shows the confusion matrix for training set of SVM

- **Test Set**

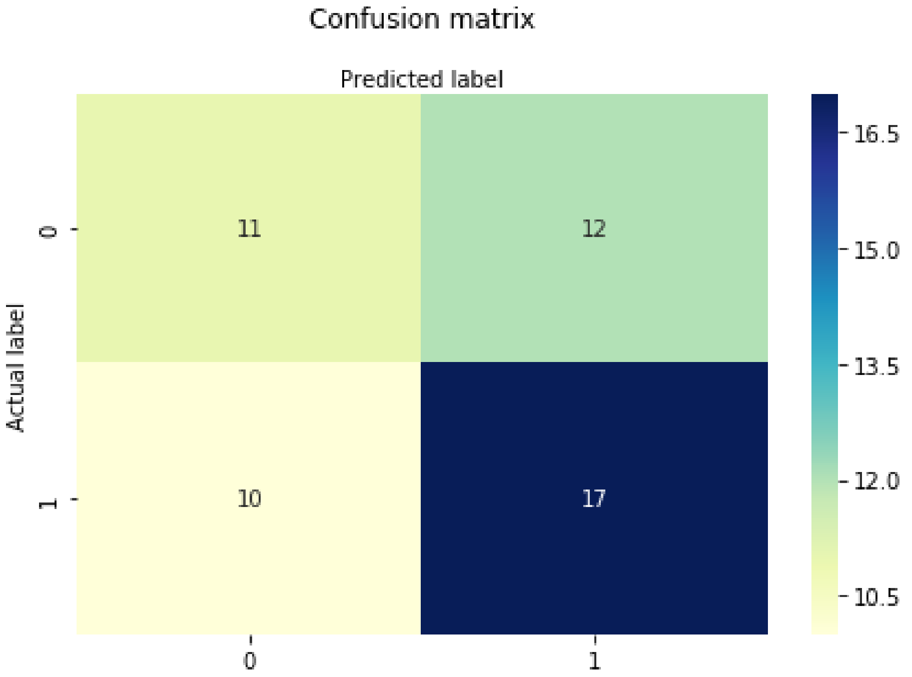


Figure 6.8, shows the confusion matrix for test set of SVM

3. XGBoost

- **Training Set**

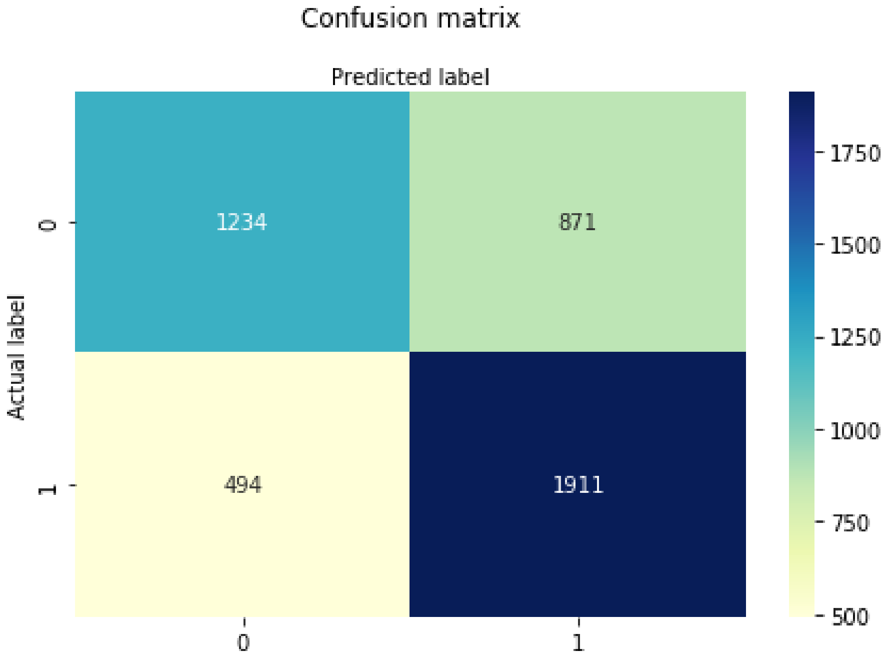


Figure 6.9, shows the confusion matrix for training set of XGBoost

- **Test Set**

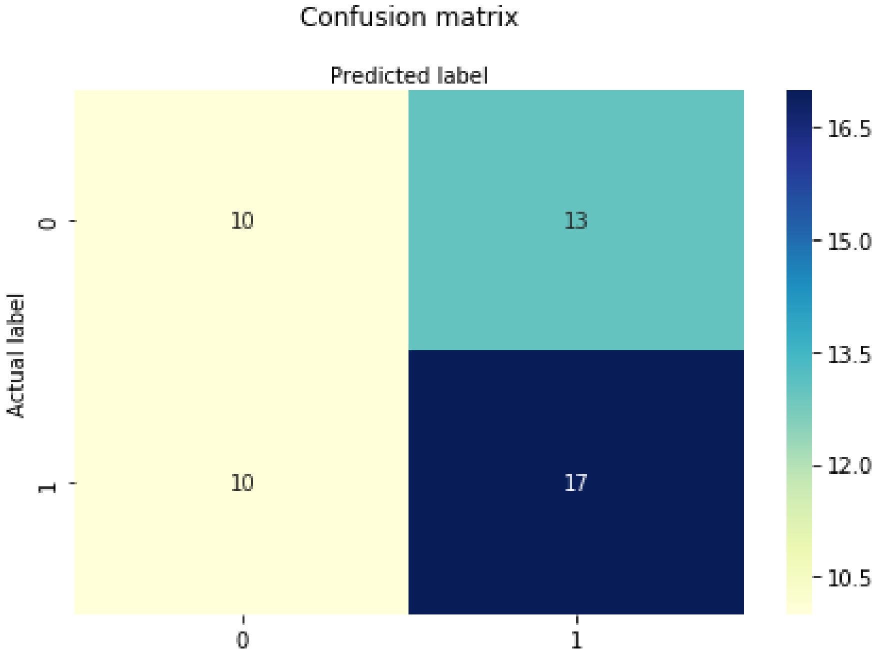
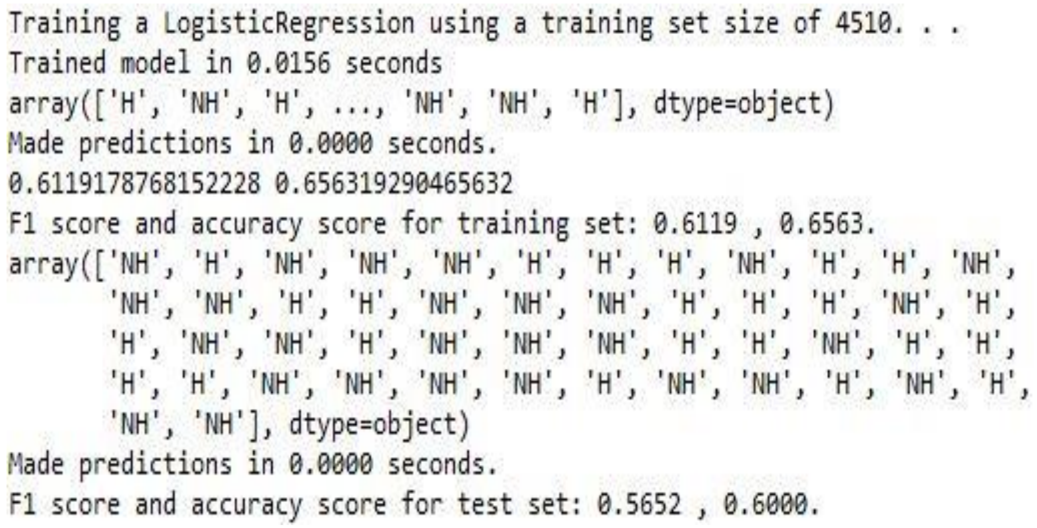


Figure 6.10, shows the confusion matrix for test set of XGBoost

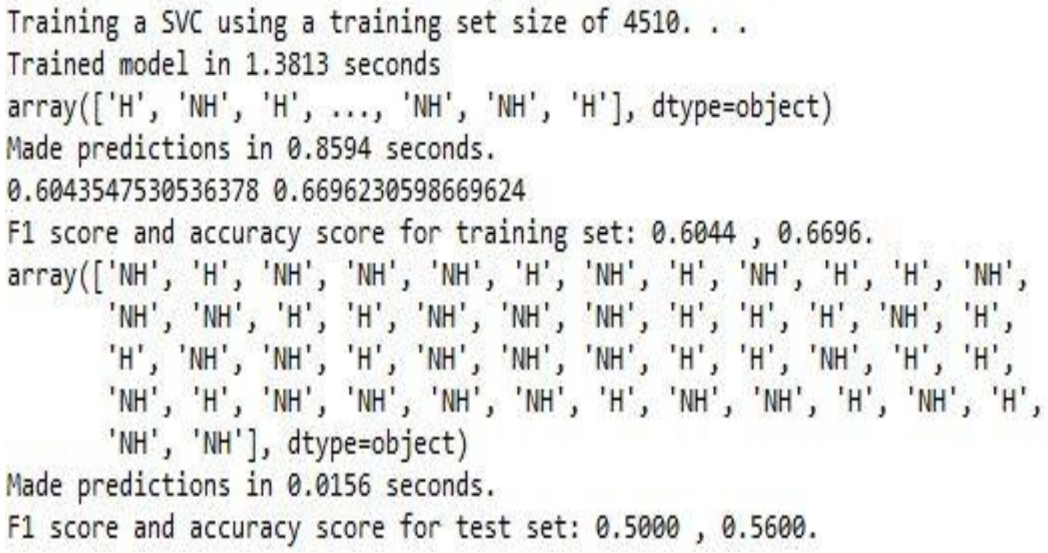
6.5 Performance Comparison

The models were first initialized and then followed by the training. Each and every model was trained separately with the final dataset. Under each model's section a timer was set to record the time it took for training and then these learnt models were tested on the test set and prediction time was noted.



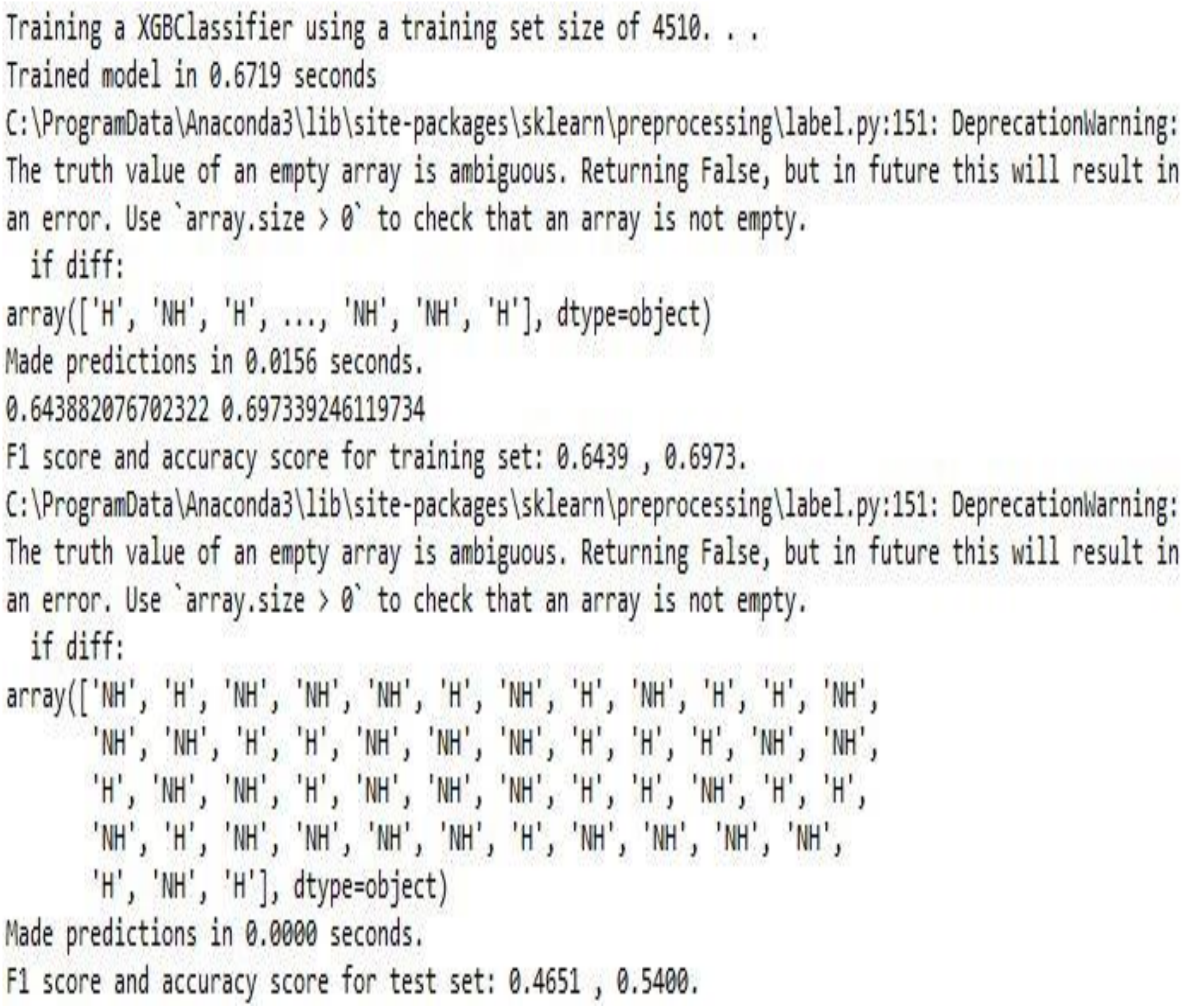
```
Training a LogisticRegression using a training set size of 4510. . .
Trained model in 0.0156 seconds
array(['H', 'NH', 'H', ..., 'NH', 'NH', 'H'], dtype=object)
Made predictions in 0.0000 seconds.
0.6119178768152228 0.656319290465632
F1 score and accuracy score for training set: 0.6119 , 0.6563.
array(['NH', 'H', 'NH', 'NH', 'NH', 'H', 'H', 'H', 'NH', 'H', 'H', 'NH',
       'NH', 'NH', 'H', 'H', 'NH', 'NH', 'NH', 'H', 'H', 'H', 'NH', 'H',
       'H', 'NH', 'NH', 'H', 'NH', 'NH', 'NH', 'H', 'H', 'NH', 'H', 'H',
       'H', 'H', 'NH', 'NH', 'NH', 'NH', 'H', 'NH', 'NH', 'H', 'NH', 'H',
       'NH', 'NH'], dtype=object)
Made predictions in 0.0000 seconds.
F1 score and accuracy score for test set: 0.5652 , 0.6000.
```

Figure 6.11, shows the results of Logistic Regression



```
Training a SVC using a training set size of 4510. . .
Trained model in 1.3813 seconds
array(['H', 'NH', 'H', ..., 'NH', 'NH', 'H'], dtype=object)
Made predictions in 0.8594 seconds.
0.6043547530536378 0.6696230598669624
F1 score and accuracy score for training set: 0.6044 , 0.6696.
array(['NH', 'H', 'NH', 'NH', 'NH', 'H', 'NH', 'H', 'NH', 'H', 'H', 'NH',
       'NH', 'NH', 'H', 'H', 'NH', 'NH', 'NH', 'H', 'H', 'H', 'NH', 'H',
       'H', 'NH', 'NH', 'H', 'NH', 'NH', 'NH', 'H', 'H', 'NH', 'H', 'H',
       'NH', 'H', 'NH', 'NH', 'NH', 'NH', 'H', 'NH', 'NH', 'H', 'NH', 'H',
       'NH', 'NH'], dtype=object)
Made predictions in 0.0156 seconds.
F1 score and accuracy score for test set: 0.5000 , 0.5600.
```

Figure 6.12, shows the results of SVM



```
Training a XGBClassifier using a training set size of 4510. . .
Trained model in 0.6719 seconds
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\preprocessing\label.py:151: DeprecationWarning:
The truth value of an empty array is ambiguous. Returning False, but in future this will result in
an error. Use `array.size > 0` to check that an array is not empty.
  if diff:
array(['H', 'NH', 'H', ..., 'NH', 'NH', 'H'], dtype=object)
Made predictions in 0.0156 seconds.
0.643882076702322 0.697339246119734
F1 score and accuracy score for training set: 0.6439 , 0.6973.
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\preprocessing\label.py:151: DeprecationWarning:
The truth value of an empty array is ambiguous. Returning False, but in future this will result in
an error. Use `array.size > 0` to check that an array is not empty.
  if diff:
array(['NH', 'H', 'NH', 'NH', 'NH', 'H', 'NH', 'H', 'NH', 'H', 'H', 'NH',
       'NH', 'NH', 'H', 'H', 'NH', 'NH', 'NH', 'H', 'H', 'H', 'NH', 'NH',
       'H', 'NH', 'NH', 'H', 'NH', 'NH', 'NH', 'H', 'H', 'NH', 'H', 'H',
       'NH', 'H', 'NH', 'NH', 'NH', 'NH', 'H', 'NH', 'NH', 'NH', 'NH',
       'H', 'NH', 'H'], dtype=object)
Made predictions in 0.0000 seconds.
F1 score and accuracy score for test set: 0.4651 , 0.5400.
```

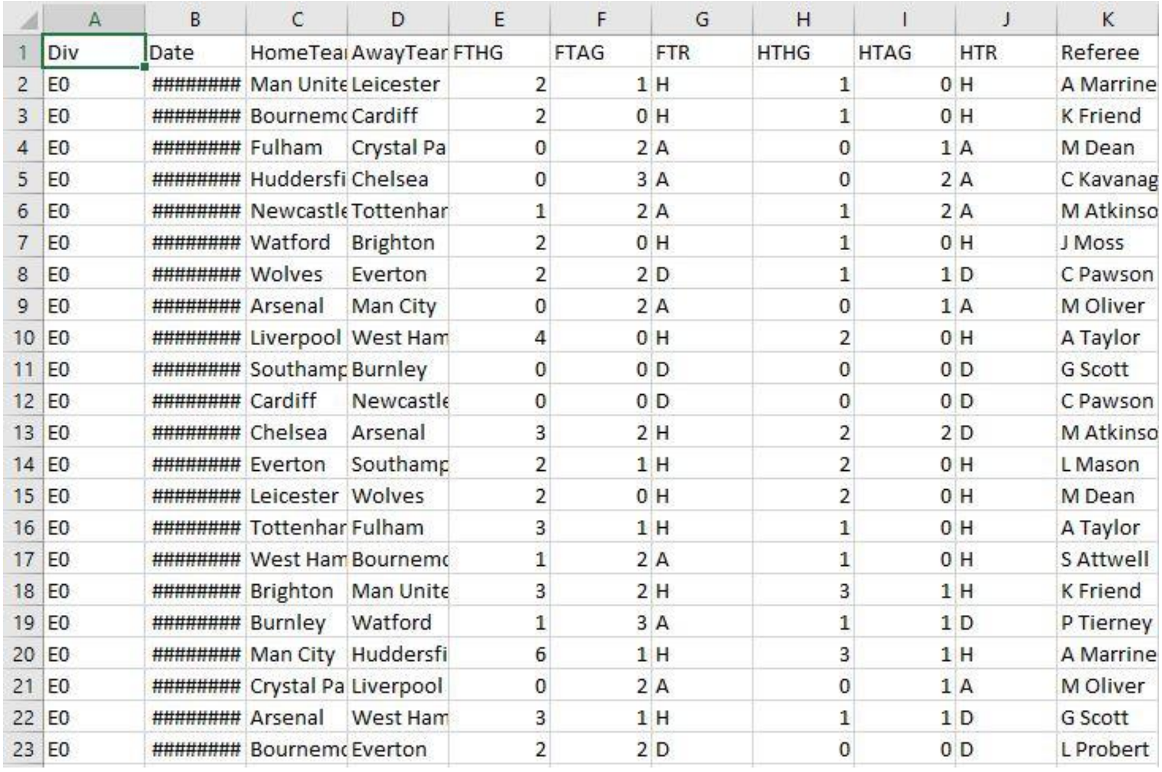
Figure 6.13, shows the results of XGBoost

The three models made their predictions accordingly and then the models with the least time of training and prediction making was selected. Logistic Regression came out as the best of all three and it was then used to predict fixtures that haven’t yet been played.

6.6 Application

After the training and testing of the three models, Logistic Regression gave the most optimum results. Therefore, Logistic Regression is further used for predicting the outcomes of those matches which haven't been played yet i.e. the matches under the match week 38 of Premier League 2018/19 season.

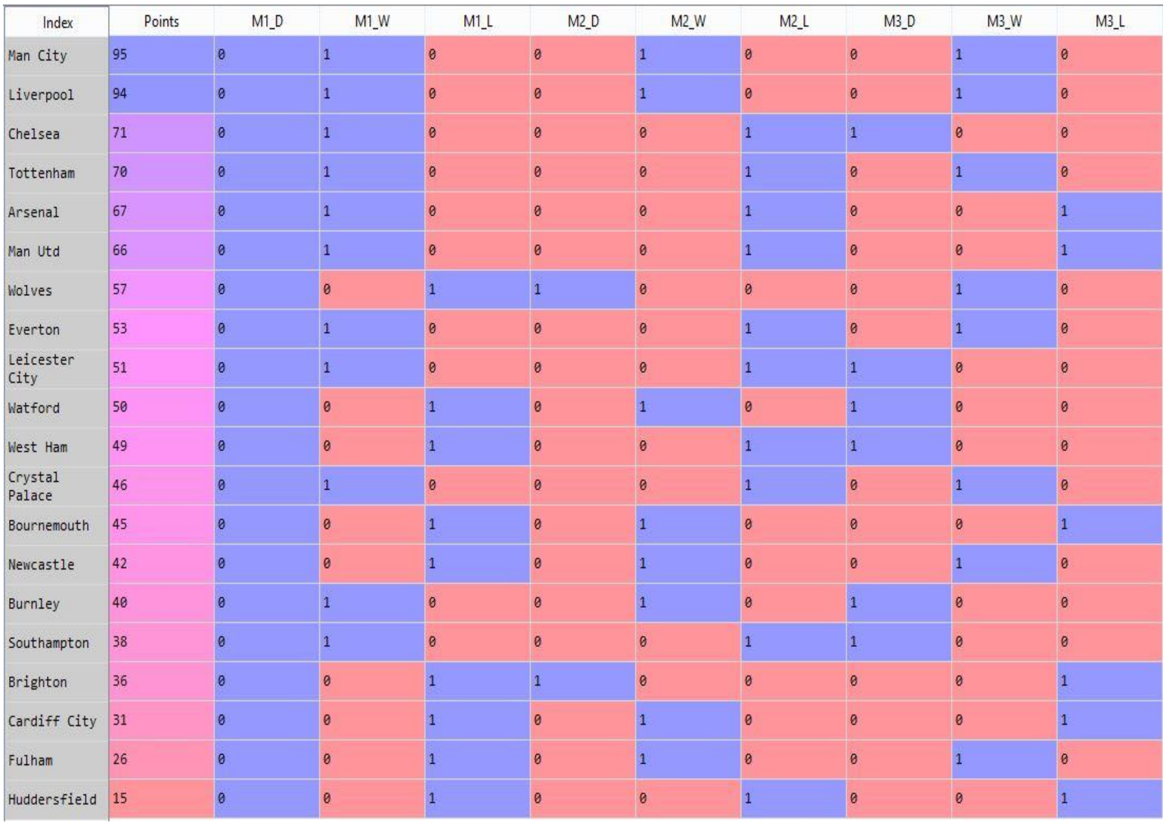
Firstly the dataset of the season 2018/19 was obtained which is shown in the figure 6.14



	A	B	C	D	E	F	G	H	I	J	K
1	Div	Date	HomeTeai	AwayTear	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee
2	E0	########	Man Unite	Leicester	2	1	H	1	0	H	A Marrine
3	E0	########	Bournemc	Cardiff	2	0	H	1	0	H	K Friend
4	E0	########	Fulham	Crystal Pa	0	2	A	0	1	A	M Dean
5	E0	########	Huddersfi	Chelsea	0	3	A	0	2	A	C Kavanag
6	E0	########	Newcastle	Tottenhar	1	2	A	1	2	A	M Atkinso
7	E0	########	Watford	Brighton	2	0	H	1	0	H	J Moss
8	E0	########	Wolves	Everton	2	2	D	1	1	D	C Pawson
9	E0	########	Arsenal	Man City	0	2	A	0	1	A	M Oliver
10	E0	########	Liverpool	West Ham	4	0	H	2	0	H	A Taylor
11	E0	########	Southamp	Burnley	0	0	D	0	0	D	G Scott
12	E0	########	Cardiff	Newcastle	0	0	D	0	0	D	C Pawson
13	E0	########	Chelsea	Arsenal	3	2	H	2	2	D	M Atkinso
14	E0	########	Everton	Southamp	2	1	H	2	0	H	L Mason
15	E0	########	Leicester	Wolves	2	0	H	2	0	H	M Dean
16	E0	########	Tottenhar	Fulham	3	1	H	1	0	H	A Taylor
17	E0	########	West Ham	Bournemc	1	2	A	1	0	H	S Attwell
18	E0	########	Brighton	Man Unite	3	2	H	3	1	H	K Friend
19	E0	########	Burnley	Watford	1	3	A	1	1	D	P Tierney
20	E0	########	Man City	Huddersfi	6	1	H	3	1	H	A Marrine
21	E0	########	Crystal Pa	Liverpool	0	2	A	0	1	A	M Oliver
22	E0	########	Arsenal	West Ham	3	1	H	1	1	D	G Scott
23	E0	########	Bournemc	Everton	2	2	D	0	0	D	L Probert

Figure 6.14, shows the dataset of season 2018/19

Now the above shown dataset is used to create a data frame that contains the information about each teams' statistics up to the 37th match week. It is shown in the figure 6.15.



Index	Points	M1_D	M1_W	M1_L	M2_D	M2_W	M2_L	M3_D	M3_W	M3_L
Man City	95	0	1	0	0	1	0	0	1	0
Liverpool	94	0	1	0	0	1	0	0	1	0
Chelsea	71	0	1	0	0	0	1	1	0	0
Tottenham	70	0	1	0	0	0	1	0	1	0
Arsenal	67	0	1	0	0	0	1	0	0	1
Man Utd	66	0	1	0	0	0	1	0	0	1
Wolves	57	0	0	1	1	0	0	0	1	0
Everton	53	0	1	0	0	0	1	0	1	0
Leicester City	51	0	1	0	0	0	1	1	0	0
Watford	50	0	0	1	0	1	0	1	0	0
West Ham	49	0	0	1	0	0	1	1	0	0
Crystal Palace	46	0	1	0	0	0	1	0	1	0
Bournemouth	45	0	0	1	0	1	0	0	0	1
Newcastle	42	0	0	1	0	1	0	0	1	0
Burnley	40	0	1	0	0	1	0	1	0	0
Southampton	38	0	1	0	0	0	1	1	0	0
Brighton	36	0	0	1	1	0	0	0	0	1
Cardiff City	31	0	0	1	0	1	0	0	0	1
Fulham	26	0	0	1	0	1	0	0	1	0
Huddersfield	15	0	0	1	0	0	1	0	0	1

Figure 6.15, shows the dataframe created using dataset of season 2018/19

Basic functionality of this part is to take the names of two teams as input i.e. one team will be the home team and the other one will act as the away team and result will be the predicted winner of the fixture with the corresponding probability.

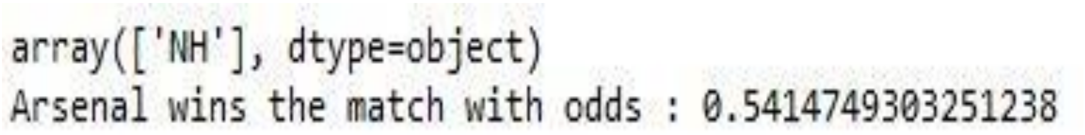
Given below is the code:

```
loc = "C:/Users/Desktop/4th Year Project/Code/"

data = pd.read_csv(loc + 'Information.csv', index_col = "Team")

str1 = "Man Utd"

str2 = "Arsenal"

team1 = data.loc[str1]
```

```
team2 = data.loc[str2]
match = pd.DataFrame(columns=X_test.columns)

match = match.append({'HTP':team1.Points, 'ATP':team2.Points,
'HM1_D':team1.M1_D, 'HM1_W':team1.M1_W, 'HM1_L':team1.M1_L,

'HM2_D':team1.M2_D, 'HM2_W':team1.M2_W, 'HM2_L':team1.M2_L,

'HM3_D':team1.M3_D, 'HM3_W':team1.M3_W, 'HM3_L':team1.M3_L,

'AM1_D':team2.M1_D, 'AM1_W':team2.M1_W, 'AM1_L':team2.M1_L, '

AM2_D':team2.M2_D, 'AM2_W':team2.M2_W, 'AM2_L':team2.M2_L,

'AM3_D':team2.M3_D, 'AM3_W':team2.M1_W, 'AM3_L':team2.M3_L, '

HTGD':team1.GD, 'ATGD':team2.GD, 'DiffPts':(team1.Points - team2.Points),

'DiffFormPts':(team1.FormPoints - team2.FormPoints), 'DiffLP':(team1.LP –

team2.LP),},

 ignore_index=True)

cols = ['HTGD','ATGD','DiffPts','DiffFormPts','HTP','ATP', 'DiffLP']

for col in cols:

    match[col] = match[col] / 38

pred = clf_A.predict(match)

display(pred)
```

```
clf_A.predict_proba(match)

prob = pd.DataFrame(clf_A.predict_proba(match))

prob = prob.astype(float)

n1 = prob.iloc[0][0]
n2 = prob.iloc[0][1]

n3 = float(n1)/float(n2)
n4 = float(n2)/float(n1)

if(n1 > n2):

    print("{} wins the match with odds : {}".format(str1,n1))

else:
    print("{} wins the match with odds : {}".format(str2,n2))
```

Now, the input given in the above code was Man Utd as the home team and Arsenal as the away team therefore, the output is shown in the figure 6.16.



```
array(['NH'], dtype=object)
Arsenal wins the match with odds : 0.5414749303251238
```

Figure 6.16, shows the final prediction with the corresponding probability

7. CONCLUSION

Under the given set of attributes the model that gave the best results and made the predictions in the lowest time was Logistic Regression. For classification purposes, Logistic Regression is the best algorithm. Hence to predict the results of upcoming matches Logistic Regression was chosen. Logistic Regression learned with a training dataset of 4561 already played fixtures. It took only 0.0156 seconds to get trained. It made the predictions in only 0.0000 seconds and the related f1 score for this model was 0.6119 and the accuracy got out to be 0.6563. Whereas, when the models was tested on the test dataset, the predictions were made only in 0.0000 seconds and the f1 score was 0.5652 and the accuracy was 0.6000.

This project is very flexible in nature. Up until now the models that got trained are able to make predictions. Logistic Regression that came out to be the best model for making predictions is also able to provide the probability for that particular prediction. Therefore, the calculated probability could be converted into odds and hence, further could be used in the process of making bets.

BIBLIOGRAPHY

1. Rory P. Bunker, Fadi Fayez”A MACHINE LEARNING FRAMEWORK FOR SPORTS BETTING PREDICTION”, September 2017

2. Fabián Enqrique Moya, “STATISTICAL METHODOLOGY FOR PROFITABLE SPORTS GAMBLING”, 2001

3. Dr. Ross Gordon and Michael Chapman, “BRAND COMMUNITY AND SPORTS BETTING IN AUSTRALIA”, 2014

4. Lisandro Kaunitz , Shenjun Zhong and Javier Kreiner, “BEATING THE BOOKIES WITH THEIR OWN NUMBERS – AND HOW THE ONLINE SPORTS BETTING MARKET IS RIGGED”, 2012

5. Stefan Luckner, Jan Schröder and Christian Slamka, “ON THE FOREST ACCURACY OF SPORTS PREDICTION MARKETS”, 2007

6. Emmanuel Olosemeka Esumeh, “USING MACHINE LEARNING TO PREDICT WINNERS OF FOOTVALL LEAGUE FOR BOOKIES”, June 2015

7. Daniel Pettersson, Robert Nyquist, “FOOTBALL MATCH PREDICTION USING DEEP LEARNING”, 2017

8. Amani Mwadime, “IMPLICATIONS OF SPORTS BETTING IN KENYA: IMPACT OF ROBUST GROWTH OF THE SPORTS BETTING INDUSTRY”, 2017

9. Thedoros Evegeniou, Massimiliano Pontil, “SUPPORT VECTOR MACHINE: THEORY AND APPLICATION”, Jan 2001

10. Tianqi Chen, Carlos Guestrin, “XGBOOST: A SCALABLE TREE BOOSTING SYSTEM”, 2016

11. Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll, “AN INTRODUCTION TO LOGISTIC REGRESSION ANALYSIS AND REPORTING”, September 2002

12. Stylianos Kampakis, “USING MACHINE LEARNING TO PREDICT THE OUTCOME OF ENGLISH COUNTRY TWENTY OVER CRICKET MATCHES”, 2015

13. Ram Raj S., Nishant Uzir, Shatadeep Banerjee, “EXPERIMENTING XGBOOST ALGORITHM FOR PREDICTION AND CLASSIFICATION OF DIFFERENT DATASETS”,2016

14. Robin Praet , “PREDICTING SPORTS RESULT BY USING RECOMMENDATION TECHNIQUES”, 2016

15. Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll, “AN INTRODUCTION TO LOGISTIC REGRESSION ANALYSIS AND REPORTING”, September 2002

16. Sungbin Lim, Hdoo Kim, Taesup Kim, Sungwoog Kim,”FAST AUTOAUGMENT”,May 2019

17. Tianqi Chen, Carlos Guestrin,”XGBOOST: A SCALABLE TREE BOOSTING SYSTEM”, 2016

18. L. Breiman, [Random Forests], “MACHINE LEARNING”, Oct. 2001

19. O. Chapelle and Y.Chang,”JOURNAL TO MACHINE LEARNING RESEARCH”, 2011