National College of Ireland

Artificial Neural Network for Betting Rate In Football

MSc Research Project

Data Analytics

Sumeet Kumar

Student ID: X18188231

School of Computing

National College of Ireland

Supervisor: Mr. Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sumeet Kumar……
Student ID:	X18188231……
Programme:	MSc Data Analytics…… **Year:** 2019-2020
Module:	Research Project……
Supervisor:	Mr. Vladimir Milosavljevic ……..…
Submission Due Date:	27 September 2020……..……
Project Title:	Artificial Neural Network for Betting Rate In Football……
Word Count:	9200…… **Page Count** 25…….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sumeet Kumar……
Date:	27 September 2020……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Artificial Neural Network for Betting Rate In Football

Sumeet Kumar

X18188231

Abstract

Prediction has been an integral part of human lives, no matter it is weather, flipping a coin, or a match between opponent. Introduction of machine learning has brought about a massive change to how people can do prediction using different attribute which they never thought of. Today prediction is almost a part of everything, be it election to president or a game between opponents. Football has massive following and during a match a lot of data is generated. Betting prediction in football is very interesting but also a very challenging task. It's challenging because a lot of factors need to be identified before using machine learning models. This challenge has been taken up for this thesis and we would be predicting betting rates for a team at home and away. For this, two different datasets from opensource have been identified for this: one – European Soccer Database and second- Complete Football Dataset. A lot of research has been done in the domain of football, but the use of neural network in predicting betting hate has not been seen. Neural Network is a developing field of Artificial Intelligence, in this research it will be put to test with Ridge Regression, Lasso Regression, Random Forest Regression and XGBoost for the prediction of Betting Rates in football. Also, after doing a thorough research it was found certain footballing parameters are vital for any prediction in football. Since these parameters were not available in the data taken, they had to be built for which Ms Excel was used. It would be interesting to see how the machine learning models and neural network model perform when these parameters are used in our research.

1 Introduction

1.2 Overview & Background of Sports

Sports have been a long part of human civilization. If we go back to 70,000 BC spear throwing was considered not only as a mechanism to improve on the hunting skill but also to challenge each other to game of throwing. With the arise of modern technology there has been a rise in the number of sports, and these are responsible for generating large amount of employment for people. It has

also been observed that sports have acted as a medium to disengage tensions between two countries, be it between the Republic of India & Pakistan or be it France and Germany during the world war. With the current scenarios there are many sports out there all having importance in one or another. Football overpowers everything out there. Having over 3.6 Billion people following it out of a total population of 7.8 Billion say's the atmosphere the sports have created. During a Football world cup which happens after every four years there has been over 3.7 Billion people watching it be it on television or live. Considering the massive number of people from all races and sexes, the amount of investment and generation of funds from the sports is massive. Brazil is considered as of the finest country in the world for producing amazing talents throughout and has won the competition five times, which is the highest. England on other hand has always been considered as the home for Football, however has won the competition just once, which in itself is a dream for many. With these massive number of followers and the investments, a lot of jobs are created which is a source of income for many, in Europe there are many nations which have their entire economy dependent on it. Amid the recent COVID-19 these countries had to or ease the lockdown and start the games because the nation's economy was going under high damage.

1.3 Motivation

Every country has its own like England has Premier League, France has French League and so on where in each team will show 24 players that they will or can use during a season. A season is a whole year calendar of football match, these matches are between the teams with each one playing one another at home (their own stadium) and away (at opponent's stadium). The playing team would be of 11 players with a total of 5/7 players on bench. Among these 5/7 players 3 players can be used as substitute during a play which is of 90 minutes. There are many positions for players to play in from ST which is striker to GK who's goalkeeper. A manager is responsible for selecting these players and also for buying player if a case arises and he feels his team needs more strength in certain areas, to buy a player a transaction occurs and is just similar to buying something from Amazon, every player has a value which depends on number of attributes like skills, years left on his contract with the current club and demand of that particular player. For any team to buy the player, a minimum amount set by the selling club has to be paid, important thing in this is that the player can be brought during a transfer period, a transfer period is the one in which these transactions are proceeded and held valid.

This whole set up about the game is very interesting to people, a lot can be done when it comes to machine learning in relation to football, from trying and predicting the match results, identifying the next best "Talent" in the game, Figuring out who is the best player for the season, Predicting the highest overall of a player(which tells who's upgraded most), Predicting the best goalkeeper and so on. The motivation of the thesis came from these discussed factors, using machine learning to predict them. It's very interesting for an individual who is or isn't a football enthusiastic person to predict different things, try and build a model which gives less error and identify the variable one which would support the built model. Many industries have their entire business dependent on betting, as these industries have their business completely on betting have their own model and

algorithms to give the betting rate. In a game these rates are needed to update, in this thesis we would focus on building a neural network model along with traditional machine learning algorithms which were taught during the course of Masters which can identify these betting rates.

1.4 Research Objective

The game of football can have three possible outcomes: win, draw or loss. For this thesis we will be building and comparing Neural Network with four other machine learning models for prediction of betting odds. Like discussed no such study has been done when neural network is being used for betting odds.

The research question:

"How *better can we build a neural network to find the betting odds, then compare it with Ridge Regression, LASSO Regression, Random Forest, XGBoost Regression and see if the results are better than neural network.* "

We would be building these mentioned models and will be using MAE, RMSE and MSE evaluation metrics and concluding which model best suits the research question for this thesis. Also, with EDA we would be elaborating on the results.

As interesting as it seems to an individual, there is however a limitation to this task, the data taken had some missing matches between opponent also the data was highly unclean and lot of data pre-processing was required for this task. Pre-processing was done through Ms Excel, but it can be very interesting yet challenging to perform it through python or any other programming language. This can be a big project and in many companies like Paddy Power™, Betfair™, bet365™ etc it is. This paper will be divided into sections, with Section 2 explaining the previous researches in this domain, Section 3 will talk about Methodology, Section 4 has the design which is implemented, Section 4 covers the implementation of design, Section 5 has implementation of ML models, Section 6 contains the evaluation done and results and then last is Section 7 which will conclude the study.

2 Literature Review

2.1 Introduction

With the sports being watched and followed in over 200 countries according to Richard et al; 2018, the amount of data being generated through the game is massive. Also, very interesting for enthusiastic individual who'd like to exploit and discover something through this data. Even though a lot of research has been done in the domain of football, there is still a lot to be achieved. Betting system is very famous in many countries, and a lot of individuals from these countries are involved in it. It's always interesting if one can figure out and predict the bets that they are placing even before the game or turn of events. In this research five different machine learning models have been built, three evaluation metrics have also been identified and noted. In this section a

thorough discussion would be done regarding the modelling techniques used before in the domain of football.

2.2 Understanding LASSO and Ridge Regression

Linear Regression techniques is a common and most used when it comes to prediction. In game of football a player's attribute contribute a lot to the prediction and has been laid along with proved by Richard Pariath et al; 2018 in the study. A player performance prediction has been built which solves the complex problem of figuring out player over all individual performance using the attributes of the player. The different variables taken by them apart from player's attribute are market value & performance value. A simple linear regression model has been built over a previous model increasing the accuracy of the model built to 91% from an initial of 84.34%, this increase was because of the new variables that Richard Pariath et al; 2018 had taken. All the three positions of a player are covered in research. From this research we would be taking player's attribute in this research as that variable has very significant effect on the model. For building any model a set of features are selected prior. Feature selection helps in removing the redundant information or irrelevant features which are strongly correlated and lead to information loss. In this research machine learning models: LASSO and Ridge would be built, so having a prior knowledge about them is necessary. Muthukrishnan R et al; 2016 have focussed on regression technique like LASSO, Ridge and OLS. For the evaluation MSE and Median MSE have been taken and it has been found that LASSO performs better than the other two. CH Raga Madhuri et al; 2019 have used six different machine learning algorithms for the prediction of house price, the taking point from this research would be to see how LASSO, and Ridge perform and using the price as an attribute. Also, MSE and RMSE are two evaluation metrics being used for the same as well. In this research we would be using the two metrics which have been used by the CH Raga Madhuri et al; 2019. After applying the models, it could be seen the OLS outperforms the two and perfectly fits the model. However, both the Ridge and LASSO gave similar value which makes it more interesting to see how the two will turn out in this research which is being conducted. Miao He et al;2015 focussed on building a model in order to find any relationship between the market value of player and performance of player. Market value as a variable is present in the data which is being used in this research. The focus by Miao He et al;2015 was to build a regression model to predict real market value and calculate the performance of a player. As the data contains non-numeric value and trees are trained in order to estimate the real Market Value of player, LASSO regression technique is used for this. LASSO performed variable selection in the linear model and gave better accuracy, Lambda function gives more features to the model. If it is increased then coefficient will be zero meaning a smaller number of variables are selected indicating shrinkage is employed. It will be interesting to see the different lambda values in this research and how LASSO regression performs. Elastic Net has a distinctive feature which is that it uses L1, L2 regularization. L1 is the Lasso Regression and is used to select the parameters, L2 is Ridge Regression and performs overfitting control at the learning process. Sergi Anfilets et al: have built a system which is based on Deep Elastic Net which would predict the winner of the English

Premier League football matches. A number of parameters are selected and Team Goals is one which is to our interest apart from how Elastic Net performs. The activation function ReLU was used for calculating the pattern of inputs. And Softmax activation function is used for output, ReLU will be used in the Neural Network built in this research. The model was successful in predicting to 64.14%, this could be used by individual for placing bets in future.

2.3 Understanding Neural Network

With the recent increase in the gambling industry, there is also increase in the set of people who'd be interested in prediction regarding specific games, Tetyana Koroteyeyeva et al; 2018 talk about the professional other than these set of people who are into the industry and can benefit from the neural network algorithm given by them. In the research Tetyana Koroteyeyeva et al; 2018 have used inverse error propagation algorithm to train the neural network. Inside which the transmission of error takes place from output to input. They have used variables like statistics of home team, statistics of at away stadium, Last five matches history, current team statistics. The two variables of home team and home team at away stadium are to be considered in this research as these too will be one of the variables being constructed for the research. Apart from the Neural Network research Tetyana Koroteyeyeva et al; 2018 used a) Naïve Bayes classifier and b) k-medium (cluster analysis) as alternative algorithm for analysis. Among the two K-medium performed better. The research was concluded by putting a platform that in order to improve the research a greater number of hidden layers can be used in the neural network which would increase the training time. S. Mohammad et al; 2014 used a previous record of seven matches for predicting matches results in future. For this research S. Mohammad et al; 2014 worked on a 3-layer back propagation technique where after receiving information at each neuron, calculation is performed and signals are forwarded to connected neurons. For determining the calculation which are performed activation function is used, in this research logsig is used on the initial layer, Tansig is used for the second layer and Poslin for the output, In the input the last seven matches criterion is used and previous results of leagues are taken as output. In order to determine and analyse the result ANOVA was used. Through the review it is understood that no prediction was done on the draw of a match. But the difference of this research to others was that no features previously discussed like players, average of teams was not discussed which however would be important part of this research conducted by us. Miao He et al;2015 conducted the research to evaluate a player's performance and the market value simultaneously. Martinjn Wagenaar et al; 2017 used Deep Convolution Neural Network to predict the opportunity a player would have to score in the game. Images of 256 x 256 pixels were created from the dataset. The images which were developed were focussing on the movement of the players during a game. Two Deep Convolution Neural Network methods first GoogleNet and second a 3-layered CNN are used; both the models are trained with use of Nesterov's accelerated gradient solver. Apart from these another technique K-Nearest neighbour was used. It acted as a baseline experiment, also to measure the distance between the ball position Euclidean distance was used. When concluded it was found that GoogleNet performed the best.

Md. Ashiqur Rahman;2020 worked on developing a framework by deep neural networks and artificial neural network in order to predict the result of football match. The highlighting point from his research was the variables taken for building the models. Apart from variables like the ranking of team, team performance, previous football matches focus was on seeing how performance of one player effects the team, player attributes among others were considered. The variables used by Md. Ashiqur Rahman ;2020 are important as relatable variables will be used in this research by us. Another important fact was that all three-probable result of a match that is win, draw and loss were considered. The test data is divided into classification and predict. The 3D facial scans are converted into metadata for classification. To identify the sat images, BoVW which is Bag of Visual Word is used. Log Loss is used in order to check the performance of models. The models were good at predicting the match results to a certain stage, but a comparison of the models to other machine learning algorithms could have improved the research. A Multi-layer Perceptron Neural Network was built by Kou- Yuan Huang et al; 2010 for getting the prediction of winning rate in football match. Advantage of MLP NN has more than one linear layer. Apart from this learning algorithm as Back Propagation Algo is used. A BP – algo calculates the loss which has been made on individual nodes and adjusts the weights so as to decrease the loss by nodes. For the research the tournament was divided into stages with 48 matches in first stage, 2 X 8 matches in next stage, and then four matches and a grand finale at last. A total of 8 variables were chosen, and important from them being goals scored, ball possession which will be part of the research conducted by us. Both the training and prediction samples are normalized before training the model. A cycle of results is used meaning results of first stage are used as input for second and then result of second as input for third and so on till final stage. The prediction accuracy decreased overall but removing the drawn games made to get good accuracy, this would be a testing challenge for our research as we will be considering all games result. All games result is important however we will be skipping the betting odds for games which are drawn but considering the drawn result of a match. Elnaz Davoodi et al; 2010 applied artificial neural network to horse racing prediction. A back propagation with momentum, conjugate gradient descent, Quasi-Newton, and Levenberg-Marquardt learning algorithm were used for this prediction. The data was of 100 actual horse races. Race time of every horse was predicted along with building architecture, for the error rate in built model mean squared error was taken as metric for evaluation. Apart from these a Multi layered Feedforward neural network is also used this along with network growing method helps in minimizing the MSE value. The MFFNNs consists of one input signal, two hidden and one output layer. When concluding it was found that the CGD or conjugate gradient descent algorithm was good at predicting which horse is at last, BPM and BP on other hand were good at predicting the first horse, also the chosen variables like weight of horse, height of horse, sprint speed of horse are quite similar to a football player. This research opens up the possibility that different models can perform differently for predicting different independent variables which are possibility of a same result like there can be a model which might be good at predicting betting odds for home and away separately, this would be interesting to see in this research.

2.4 Understanding other Machine Learning Techniques

Champions League is a competition which is played by top teams from all the different leagues of Europe. Josip Hucaljuk et al; 2011 has built different machine learning models for the prediction of outcomes of the matches in champions league. In total there are five rounds with 32 teams participating in the competition. A framework was developed for classification, feature selection and identifying the ideal algorithm among Naïve Bayes, Bayesian Network, LogitBoost, K-Nearest Neighbour, Random Forest, ANN. Java was used as a programming language. Important point from this conducted research is impact of an injured player on the result. Random Forest outperformed among all; it will be interesting to see how random forest perform in our research as it is one of machine learning algorithm being used. Apart from this Naïve Bayes was worst. LogitBoost took ahead the research conducted by Bayesian network by improving the difference found between the sets. However, the prediction by ANN was way better than LogitBoost. Limitation to this study was that variables like attributes of team, player, overall, of player were not used. This would be covered in our research also it would be an interesting to see who among the two performs better between the ANN or Random forest. Stefan Dobravec; 2015 modelled a prediction algorithm that has used latent features which were obtained from matrix factorization. For the prediction Stefan Dobravec; 2015 has used FIFA World Cup 2014 data. For the prediction of outcomes of match Naïve Bayes has been used. Cross fold technique was used and roc curve was taken to explain the results as the data was less. Naïve Bayes however improved the accuracy of the model. The research could have been improved if the available data had been more which would have put the Naïve Bayes to test as well. Ben Ulmer et al; predicted the result of English premier league with the help of artificial intelligence, machine learning algorithms. The variables taken for the study are game day data and current team performance. A total five different classifier models were built with linear from stochastic gradient descent, naïve Bayes, hidden markov model, SVM and random forest. The model was gradually reduced from 3-class classification problem to 2-class classification problem and predicted if a team would win, lose or draw a match. SVM, Random forest performed equally well and achieved compatible error rates. However, the research could have been better if more data would have been used, to which we would be taking 7 different leagues not just one which has been taken in this research. Also covering more variables which are used in player analysis.

Two season data of English Premier League team Tottenham FC of 1996/95-97 was used by N.E. Fenton et al;2005. This research uses expert knowledge about the building of machine learning models that just creating one which on other case can backfire if the knowledge is not correct. The feature selection technique has both filter and wrapper in order to choose the best supported variables. The variables taken are player's overall, his position, home or away match, and opponent teams attributes these will be used in our research apart from others. For Bayesian Network, Hugin tool was used to run and construct the model. The focus of this research was on early prediction of result also treating them qualitative and quantitatively. This would help in foretelling the novel outcome of result. Other machine learning technique used in the study were MC4, a decision tree learner, Naïve Bayes, Data Driven Bayesian and K-Nearest Bayesian Neighbour learner. The study was concluded by BN showing far better result than the others. The

research would have been more interesting had more team's data been used and then decision tree algorithms which have been performing well in other research put to test. In this research we would be using data from one season of 7 different leagues and over 100 teams and it will be interesting to see how the models perform there.

3 Methodology

For the implementation of the research the study has gone through a number of stages following a CRISM-DM methodology. Below we have an illustration of FIFA-Betting Odds prediction methodology in Figure 1 which has been divided into 6 stage namely Data Understanding, Data Acquisition, Pre-Processing, Modelling and Evaluation. Then trying and getting on with the decision. The results after getting through this can help in further building and modifying algorithms used in Multi-National companies, improving on their business.

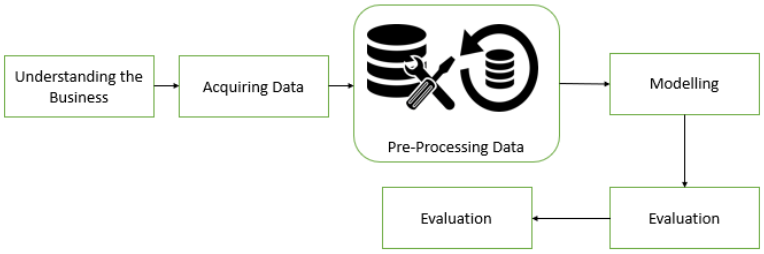

Figure 1: Design for Methodology

3.1 Data Understanding

There are numerous factors which can lead to betting odds during a match or even before a match is played like playing eleven, opponent team, if match is in home or away stadium, weather, fans in the stadium, team's average, previous scores and many more. However, some of the factors like weather, fans, and emotional state of a team are out of hands and might be important to effect on betting odds but we'll consider them as exception to this thesis.

Moreover, to understand and challenge the model we have built variables like Away Team Potential Average, Home Team Potential Average, Away Team Overall Playing Average, Home Team Overall Playing Average, Away Team Average, Home Team Average. These variables are built by taking average of attributes which were available to us from the data, this whole process is done in MS Excel.

3.2 Data Acquisitions

Data is acquired from a single data source but two different data files. First one being European Soccer Database, the file is in SQLite format and taken from Kaggle, which is an open source for data analyst or enthusiastic who wants to build on their own model and predict something of their own. Second file is FIFA 20 Complete Dataset, single file of player data 2015 is taken. The available data in the two downloaded file is shown in Table 1.

Dataset	Record Counts	Attributes Count
European Soccer Database		
Country	11	2
League	11	3
Match	25,979	135
Player	11,060	8
Player Attributes	1,83,978	50
Team	299	5
Team Attributes	1458	26
FIFA 20 Complete Dataset		
FIFA 15 –> Player_15 Dataset	15,458	104

Table 1: Description of Data

The data present in European Soccer Database is in SQLite format and DB Browser is used to extract this SQLite file data, a representation of which is shown in Figure 2, the DB Browser converts it into CSV and save as individual files. Also, the data from FIFA 20 complete Dataset is available in CSV format. From both the files we would take data for one season, we have chosen 2015-2016 season for the research.

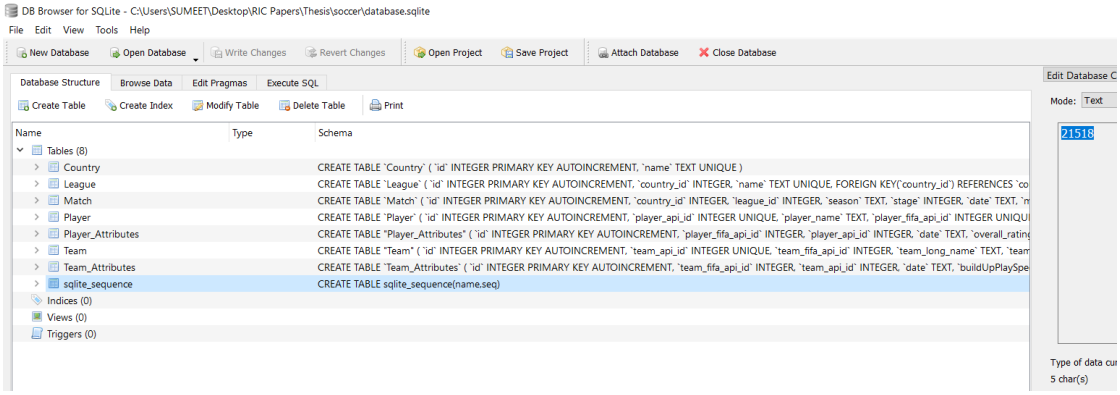

Figure 2: DB Browser SQLite to CSV

Figure 1 would help us to know the number of variables present in the file, these are converted into CSV and saved so in total we have 7 CSV file in this data. Let's discuss what each file holds:

- Country: - Has the number of countries taking part in the sports, though there are many in the thesis we would focus on the France, Germany, Italy, Netherlands, Portugal, Scotland, Spain.
- League: Every country has its own league so 7 country would have 7 different leagues, namely League 1 from France, Bundesliga, Serie A, Eredivisie, Liga ZON Sagres, Scotland Premier League, and LIGA BBVA.

- Match: This is most important file for the thesis as it contains the matches between two teams and also the betting odds, which help us in building the models. It contains over 100 Columns and 25400 rows. It also has valuable information about the scores each team has against each other as per their individual league.
- Player: This file is equally important but the data taken from FIFA 20 Player Data would be taken as that has more updated data of players playing in the leagues.
- Player Attribute: This file is responsible for generating the playing 11 along with the 6-substitute player who are involved in the team, the skills and average of these player is taken together and then average is taken which is a player's average.
- Team: It contains all the different team which are there in the different leagues. For

3.3 Data Pre-Processing

As for any research related with data, data pre-processing step is very important and crucial task which was also the case in our research. Pre-processing is done by handling null values, missing values in each considered file and merged the two files into one after which Feature scaling is done.

3.4 Missing Values

As all the files had some form of missing values it was very clear that they will not be part of the research, with the help of Excel these attributes were removed. After treating the data of the NAN values and missing data the initial data was reduced.

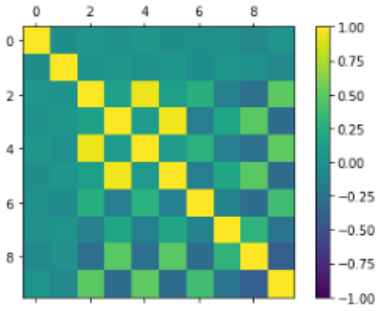

Figure 3: Correlation Matrix

- The correlation matrix below in Figure 3 shows that the values are not correlated to one another. It can be clearly seen that majority of the values lie 0.50-0.25. Also, the figure 4 illustrates that the values as in the region on 0.50-0.25 and are not co-related this provide as support to the correlation matrix which is seen in the figure 3. So, all the considered variables are retained.

	Home Team Average	Away Team Average	Home Team Overall Playing Average	Away Team Overall Playing Average	Home Team Potential Average	Away Team Potential Average	home_team_goal	away_team_goal	B365H	B365A
Home Team Average	1.000000	-0.004239	0.049525	0.008507	0.054799	-0.004971	0.007456	0.008841	-0.059192	0.032863
Away Team Average	-0.004239	1.000000	0.026220	0.060958	0.007695	0.051908	-0.031223	0.038049	0.013842	-0.057128
Home Team Overall Playing Average	0.049525	0.026220	1.000000	0.145230	0.943404	0.110591	0.259589	-0.130553	-0.259033	0.495633
Away Team Overall Playing Average	0.008507	0.060958	0.145230	1.000000	0.105385	0.956717	-0.145425	0.179709	0.489065	-0.288686
Home Team Potential Average	0.054799	0.007695	0.943404	0.105385	1.000000	0.078408	0.273506	-0.134117	-0.270251	0.500413
Away Team Potential Average	-0.004971	0.051908	0.110591	0.956717	0.078408	1.000000	-0.134039	0.188570	0.488257	-0.284933
home_team_goal	0.007456	-0.031223	0.259589	-0.145425	0.273506	-0.134039	1.000000	-0.107721	-0.278016	0.385120
away_team_goal	0.008841	0.038049	-0.130553	0.179709	-0.134117	0.188570	-0.107721	1.000000	0.298618	-0.233660
B365H	-0.059192	0.013842	-0.259033	0.489065	-0.270251	0.488257	-0.278016	0.298618	1.000000	-0.403778
B365A	0.032863	-0.057128	0.495633	-0.288686	0.500413	-0.284933	0.385120	-0.233660	-0.403778	1.000000

Figure 4: Correlations

3.5 Merging

Data from two different files from Kaggle has been taken. One named European Soccer Data base contains 7 different csv files and other one FIFA 20 Player Dataset from which 2015 – Player Data is taken. Both the files after removal of null values and the attributes which are not impacting on our research are saved. The merging of the different files can be seen in Figure 5.

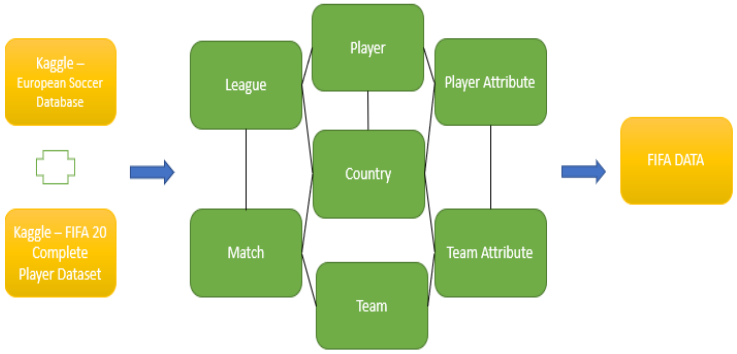

Figure 5: Merging of data file

A new csv file is created which has all the important attributes that will impact the conducting research, to test these attributes correlation matrix is built. Also, before building models MSNO Matrix is performed to check for any null values.

The chosen attributes for our research are following:

- Home team goal: This has the number of goals a team playing has scored at home.
- Away team goal: This contains the number of goals a team playing away from home has scored.
- Home Team Average: This has been created after selecting the top 18 players from the players file and then an average has been taken of those selected, building a team and further an average of team.

- Away Team Average: Similar task has been done here for calculating the away team average.
- Home Team Overall Playing Average: This attribute has been created by choosing the top 11 players who would participate in the game.
- Away Team Overall Playing Average: Similar task has been performed for this attribute like the playing home team average.
- B365H: This is betting rate of Bet365 which is a company which lets people place bets. For the data we are going to use betting rates provided by Bet365. B365H is the betting rate for a team at Home.
- B365A: This column contains the betting odds from bet365for a team at Away. This away signifies that the team playing at away has a certain weightage of winning or losing the match.
- Home Team Potential Average: This column has the values of teams playing at home who have player with potential, many a times one player outperforms other and wins the game, this can thus be considered as a variable for model building.
- Away Team Potential Average: Similar to the above column this one has values of teams playing away games.

3.6 Feature Scaling

The formulated data contains numeric values in all the attributes build after the merging of different files into one. The values in score which is goals scored and the values in team's or player's average have different scales and would affect the machine learning model which are built.

In order to lower that, scaling is done so as to bring all the values to a common scale which not only would improve the performance of model but improve the speed as well. If this isn't done then model would give higher weight to higher values and lower weight to the lower values.

4 Design Specification

For understanding the research through a better way, a three-way design approach has been used which can be understood with the help of Figure 6. In the Design phase we simply Preparing the Data set first, then Modelling and applying the machine learning models on it and then performing EDA.

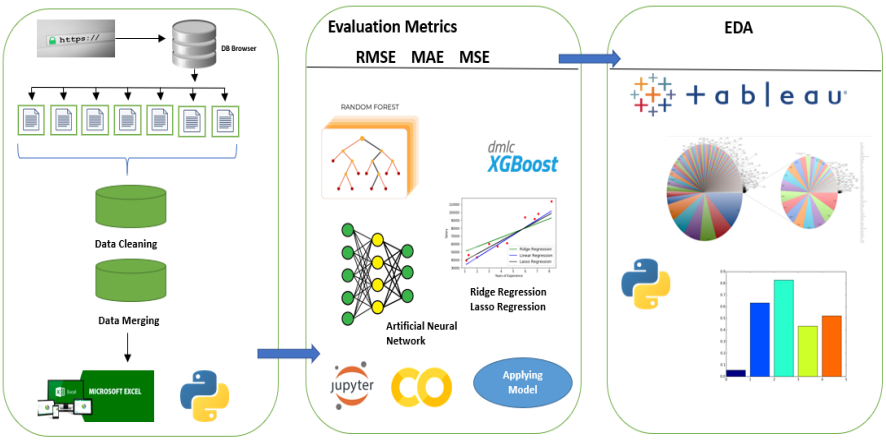

Figure 6: Design

- The first part consists of collecting the data then with help of DB browser converting into CSV file, pre-processing the data through the help of Excel. Then Merging of data is done, after which exploratory analysis is done. The selected file is processed for modelling and mounted on the Google Colab for applying models.

- The second stage consist of applying the models - Artificial neural network, XGBoost, Ridge Regression, LASSO Regression, Random Forest and Multiple Linear Regression. For evaluating these model's metrics like MAE (Mean Absolute Error), MSE (Mean Squared Error) and RMSE (Root Mean Square Error) are used. Google Colab is used as a platform to build the models and for evaluation.

- The Final results which are obtained are shown with help of plots and graphs and meaningful insights can be achieved through them.

5 Implementation

5.1 Preparing Data for models

As we are focusing on data from 7 leagues the large data which was initially was 222,796 rows and 119 columns in 7 different files and 15, 458 rows and 104 columns in another eventually comes down to 1712 rows and 10 columns. This can be explained as there are limited number of games in a league. The calculations of these column are based on previous studies along with testing from correlation matrix. Also, before dividing the data scaling was done so as to bring all the values to a common platform.

The data is divided into test and train. But as there are two different variables to be predicted the project is divided into two parts:

- First Case – "B365H" is taken as independent variable. Models will be built for the prediction of this variable.

- Second Case – “B365A” this is taken as the independent variable, and models are built for prediction of this column.

The results of the models are evaluated on the three chosen metrics. And the best fit model which gives least error figured.

5.2 Artificial Neural Networks

Artificial Neural Network are computational algorithms. Just like the human neurons behave, they are intended to do the same. These are capable of performing both the machine learning and pattern recognition. The neurons can take value from the inputs provided. It consists of three layers, and keras library in python is used to implement them. Dense constructor is used for defining the neurons. For the testing the multiple layers, epochs and batch size are tweaked. Different Epochs, batch size is used in order to reduce the error.

For improving results, the two output have been calculated separately. First case is where ‘B365H’ which means the betting rate for home team is taken and model is built and this is taken as output. Figure 7 shows that 81 total trainable parameters are given by the neural network for different dense layers, each dense layer has a set of neurons. The number of neurons in dense layer is set to 8. And the output is 1.

```
Model: "sequential_11"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_22 (Dense)             (None, 8)                 72
_________________________________________________________________
dense_23 (Dense)             (None, 1)                 9
=================================================================
Total params: 81
Trainable params: 81
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 7: ANN architecture showing the input, hidden and output layer.

For the evaluation part error minimization is used, after increasing number of epochs to 100 it can be seen the error rate is continuously decreasing with minor fluctuation, Figure 8 gives us complete knowledge about this.

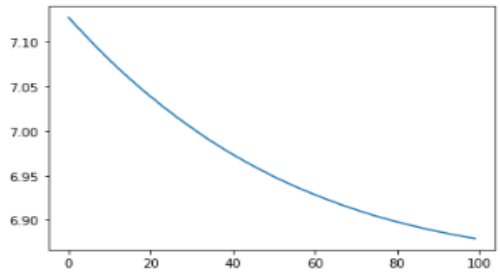

Figure 8: Error Minimization

Case two where ‘B365A’ which means odds of team playing away is taken as output and put to model. Figure 9 shows the base line model with 81 trainable parameters having dense layer and set of neurons. Number of neurons in the dense layer is 9.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 8)                 72
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 9
=================================================================
Total params: 81
Trainable params: 81
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 9: Neural Network Layers

Through the Error minimization graph, it can be seen that error is not getting down even when the number of epochs has been increased to 200 which in case-1 were 100. Figure 10 shows how the error has lot of fluctuation.

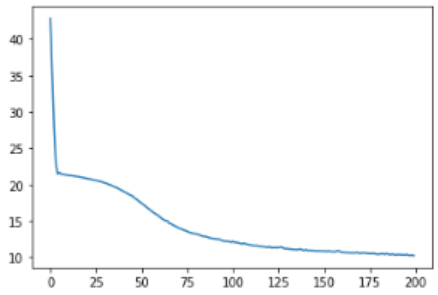

Figure 10: Error Minimization Visualization

The table 2 provides with details of the loss which would occur at different epochs for the neural network. This experiment was performed for the case 2. The epochs are measure of number of times all the training vector are used to update the weights, the batch size tells us about the number of samples processed before the entire model is updated. Through the table it can be understood that even when loss is going less but very marginal change in loss. Therefore, we have taken 200 epochs for the second case through this table.

Model	Epochs	Batch size	Loss
ANN	50	128	9.9
ANN	100	128	9.43
ANN	150	128	9.19
ANN	200	128	9.03
ANN	250	128	8.9
ANN	400	128	8.8

Table 2: Case 2

5.3 XGBoost Regression

XGBoost or Extreme Gradient Boosting regressor is type of decision tree-based algorithm and uses gradient boosting framework. It is a very usable algorithm when it comes to regression problem.

After the parameters were scaled then best parameters available were used for applying to the model. Below the Figure 11 gives us understanding of different setting to the XGBoost Regressor.

Best parameters were selected using the gamma value as 0, the n_estimator tells us about the number of trees taken before any prediction, for the experiment this was set to 100. Learning rate which explains the shrinkage which happens at each level is set to 0.1, the learning rate usually is kept low to get better result. Verbosity helps in viewing the result which take place in the experiment, if this is kept to 0 no result would be viewed and 1 makes one print the result. And then sub sample which tells about the ratio of sub sample at each node in the tree is set to 1.

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0,
             importance_type='gain', learning_rate=0.1, max_delta_step=0,
             max_depth=3, min_child_weight=1, missing=None, n_estimators=100,
             n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
             silent=None, subsample=1, verbosity=1)
```

Figure 11

5.4 Random Forest Regressor

This is a supervised learning algorithm and uses the ensemble method for the regression technique. Random forest in itself is a bagging technique, the formed multiple decision trees run side by side. In this research scaling was performed and then model was applied.

After scaling the model was built and parameters selected are defined below in Figure 12: Random Forest builds multiple decision trees normally using bootstrap and the output to be predicted is determined across the tress mainly by taking majority of votes and aggregating the prediction. For the research parameters were selected through the previous researches conducted and domain knowledge, when optimization was done the parameters selected were taken with minimum sample split as 2, n_estimator which tells about the number of trees is taken as 20. For the evaluation of the parameters taken we have used MSE which is minimum squared error calculated on the split in the model built, minimum sampled leaf is 1.

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=None, max_features='auto', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=20, n_jobs=None, oob_score=False,
                      random_state=0, verbose=0, warm_start=False)
```

Figure 12

5.5 Ridge Regression & LASSO Regression

Both Ridge and LASSO are types of Regularization techniques. In case of Ridge the coefficients are shrink and complexity along with multi-collinearity are reduced. Lamda the penalty term regularizes the selected coefficient that whenever they take large value the optimizer function is penalized; this makes ridge regression shrink the coefficient and reduce model complexity. LASSO Regression help in feature selection, as the cost function Lambda can be controlled. Muthukrishnan R et al; 2016 elaborated that both the models shrink the estimate regression coefficients which are approaching to zero, Ridge on one hand minimizes the squared sum of

coefficient which is L2 regression and the LASSO minimizes the absolute sum of coefficient which is L1 regularization. Any value of Alpha which is non zero would give value less than that of simple linear regression. We have taken value of Alpha as one in both the models so as to not make models more conservative. Figure 13 and Figure 14 shows model fitting in Ridge and LASSO Regression.

```
Ridge(alpha=1, copy_X=True, fit_intercept=True, max_iter=None, normalize=False,
      random_state=None, solver='auto', tol=0.001)
```

Figure 13: Ridge Regression

```
Lasso(alpha=1, copy_X=True, fit_intercept=True, max_iter=1000, normalize=False,
      positive=False, precompute=False, random_state=None, selection='cyclic',
      tol=0.0001, warm_start=False)
```

Figure 14: LASSO Regression

6 Evaluation

For the evaluation of the models which have been used in the research three different statistics have been used. These help in determining whether they best fit the model or not. MAE value, RMSE value, MSE value are the three parameters which have been used.

6.1 Experiment with Neural Network

Neural Network is used as the main model for the research. The model is divided into two cases as per requirement. Individually for both the cases evaluation has been done and then compared with other machine learning models built. The MAE, MSE and RMSE value have been calculated for both the cases.

When taking case – 1 which is taking the "B365H" which is betting rate for home team as an output and building model for that. The data has been split into X and Y, where X is the input and Y is the output value. The activation function ReLU or Rectified Linear Unit is used, in which the argument either doesn't go through or just lets it pass. The disadvantage being it doesn't retain negative value but the data that is used in research has no negative value. Optimizer Adam which is stochastic gradient is used. This is adaptive estimation of first order and second order moments. The MSE value is calculated and then verified with K-fold Estimator with 10 splits, here the data is randomly chosen and verified. Figure 15 shows the loss error and the MSE value 1.235 for the evaluation of model.

```
43/43 [==============================] - 0s 693us/step - loss: 3.3442
11/11 [==============================] - 0s 910us/step - loss: 1.2357
1.2357306480407715
```

Figure: 15 Evaluation of model for Case - 1

Talking about Case – 2 where the "B365A" is taken which is betting rate for away team as an output. Again, the same steps are followed and same optimizer and activation function are used. But when evaluated the model the MSE value is high also the error rate is high as well. Figure 16 shows the MSE value as 5.4151 and the loss as well.

```
43/43 [==============================] - 0s 772us/step - loss: 8.9331
11/11 [==============================] - 0s 2ms/step - loss: 5.4152
5.415185928344727
```

Figure: 16 Evaluation model for Case – 2

The table 3 contains the value of three different metrics that is MSE, MAE, RMSE for the two cases for which research is done. Table 3 is the evaluation results of Neural Network model.

Metric	Case-1	Case-2
MSE	1.228	1.555
MAE	0.640	5.818
RMSE	1.059	2.412

Table 3

6.2 Experiment with Ridge Regression

Among the four other models used, Ridge Regression is first model. Before Executing these models, the data is completely scaled by using feature scaling. This scaling brings all the value to a common level. The data is divided into test and train and model is applied to both the cases. With "B365H" being considered as Case - 1. And Case – 2 for "B365A". The table 4 contains the evaluation metrics of experiment of Ridge regression. The three metrics MSE, MAE, RMSE have been calculated for the evaluation of the model. Values of the metrics for both the cases of research have been illustrated in table 4.

Metrics	Case- 1	Case- 2
MSE	0.006	11.757
MAE	0.047	2.229
RMSE	0.076	3.428

Table 4

6.3 Experiment with LASSO Regression

LASSO Regression shrinks the data, this shrinkage is toward the central point which is the mean. The divided data is used for the model building. Both the cases have different values exploiting the model. Even after the Alpha value is increased from 1 to 10 no good results or values of the metrics was observed. At alpha value of 0 this model acts like linear regression so the value is kept 1. Table 5 has the values of MSE, MAE, RMSE which are the three different metrics used to evaluate the model. The experiment is associated with Lasso Regression, and the three different result have been defined in table 5.

Metrics	Case- 1	Case- 2
MSE	0.009	12.519
MAE	0.057	2.213
RMSE	0.098	3.538

Table 5

6.4 Experiment with Random Forest Regression

Random forest is applied to both the cases and metrics are used for the evaluation; Table 6 contains the values of different metrics for the experiment performed for Random Forest Regression. The three different metrics MAE, MSE, RMSE have their respective results for both the cases which is betting odds for home team and betting odds for away team.

Metrics	Case- 1	Case- 2
MSE	0.664	6.711
MAE	0.315	1.342
RMSE	0.815	2.591

Table 6

6.5 Experiment with XGBoost

Experiments were performed on XGBoost and then evaluation was done using the metrics MAE, RMSE and MSE. The data was divided into Test and Train and efficiency of model was recorded. Table 7 provides us with the results which came for the evaluation metrics used in the XGBoost regressor. The result is for both the cases which are there that is Case 1 and Case 2.

Metrics	Case- 1	Case- 2
MSE	0.576	6.603
MAE	0.301	1.364
RMSE	0.759	2.569

Table 7

Comparing the models in both the cases: -

Figure 17 and Figure 18 provides vital visual information about how good a model fit. The error also helps in interpreting the result of the models as well.

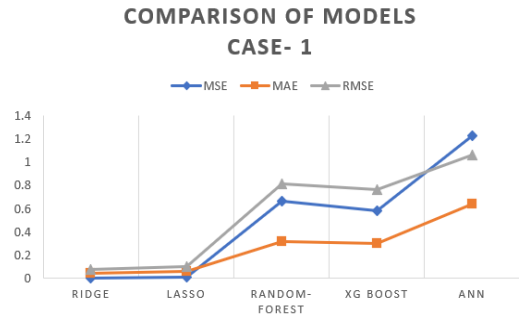

Figure 17

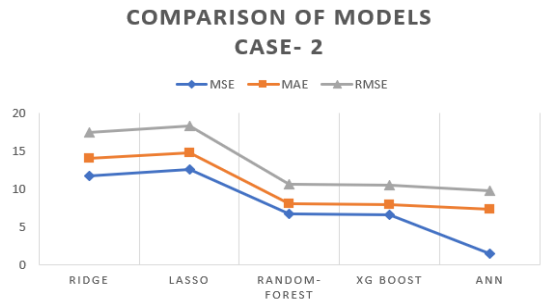

Figure 18

From the figure it can be easily interpreted that Ridge regression outperformed all other models for all the three metrics in the Case-1 where "B365H" bets were predicted, and on the other hand Case-2 had ANN outperforming the rest of the models for all the three metrics. We will discuss more along with the tables in next section.

6.6 Discussion

The research is done with prior knowledge about football and using the help of previous researches done in the domain. A lot of researchers have taken variable based on prior knowledge, among them is N.E. Fenton et al;2005. The variables chosen for this research have been developed as per the research done in the domain. The selection of four models used apart from neural network was taken due to the research problem being classification and also in the research done it was noticed that random forest, ridge regression

is tested models. Lasso which is a bit like ridge was taken due to the fact it performed better in Muthukrishnan R et al; 2016 and this would have been a testing phase for this model against ridge.

In total five different machine learning models were implemented and evaluated in the research. Different techniques were used to improve the performance of the models for example merging the dataset on basis of the requirements, feature scaling, and using MSNO matrix for finding null values. In both the cases Case-1 for "B365H" and Case- 2 for "B365A" different models have outperformed the others. The evaluations metrics used helped to reach this conclusion. Table 6 gives a proper idea about which metrics or model is favouring which Case. Ridge Regression gives most significant values of the metrics used in the first scenario that is Case- 1 and then Artificial neural network for the Case-2. In the neural network the error rate is coming down to 100 epochs in Case- 1 and is very discontinuous in Case-2 but has still performed better than the different models used. Random forest performed when compared to ANN according to Josip Hucaljuk et al; 2011, This proves that different models have to be used for the identifying the betting rate. The XGBoost and Random Forest Regression are the two ensemble learners but the way of building the trees is different in both. Also, Ridge and Lasso previously have given similar values as per Muthukrishnan R et al; 2016 and CH Raga Madhuri et al; 2019 it can be seen through the Table 6 and Table 7 that they have very marginal change and performed as they did in previous researches.

Models	MSE	MAE	RMSE
Ridge	0501	0.047	0.076
Lasso	0.009	0.057	0.098
Random-F	0.664	0.314	0.815
XGBoost	0.577	0.302	0.759
ANN	1.229	0.641	1.059

Table 6: Case1 ("B365H")

Models	MSE	MAE	RMSE
Ridge	11.756	2.229	3.428
Lasso	12.528	2.2131	3.538
Random-F	6.711	1.343	2.591
XGBoost	6.603	1.365	2.569
ANN	1.555	5.818	2.413

Table 7: Case 2 ("B365A")

7 Conclusion and Future Work

In the study, focus was on identifying the betting rates of a game. Different attributes were build using the help of MS Excel and Python. Two different data sources were used and merged. The build attributes favoured different models. It was found that Ridge Regression suits best for finding out the betting rate of a team playing at home using the taken variables in research. Similarly, Neural Networks model can be for finding betting rate for the team playing matches away from home.

As the average of a team either playing at home or away was calculated by using team attributes it can be said that a team having good players matters but it doesn't give the upper hand in winning the matches. There are different effective attributes which also lead to the identification of the betting rate. Attributes like sentiments of player drive one and the fans present in the stadium build on to this sentiment. In future using these attributes and building the models would be an exciting challenge for anyone, also as a win, draw or loss has been predicted by Sergi Anfilets et al, S. Mohammad et al; 2014, Md. Ashiqur Rahman;2020 this can be use along with this research to predict any future betting rate along with match result. More or so Football has vast area of research which is yet to be covered, any football enthusiastic can work and bring out more insights about the game which can be helpful for the world.

Acknowledgement

First, I would like to thank my research supervisor Mr. Vladimir Milosavljevic for his relentless support and motivation. For 13 weeks, he has assisted me immensely in my queries and guided me. I want to thank my family and friends for their understanding of this crucial time for me in completion of my thesis. At the end I would like to thank the co member students whose questions and queries helped me in understanding more and gaining more knowledge.

References

Anfilets, S. and Bezobrazov, S. (no date) 'Deep Elastic Net for Prediction the Winner of Football Matches', pp. 1–8.

Arabzad, S. M., Araghi, M. E. T. and Soheil, S. (2014) 'Applied Research on Industrial Engineering Football Match Results Prediction Using Artificial Neural Networks ; The', (October).

Davoodi, E. and Khanteymoori, A. R. (2010) 'Horse racing prediction using Artificial Neural Networks', *Proc. of the 11th WSEAS Int. Conf. on Neural Networks, NN '10, Proceedings of the 11th WSEAS Int. Conf. on Evolutionary Computing, EC '10, Proc. of the 11th WSEAS Int. Conf. on Fuzzy Systems, FS '10*, (June 2010), pp. 155–160.

Dobravec, S. (2015) 'Predicting sports results using latent features: A case study', *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings*. MIPRO, (May), pp. 1267–1272. doi: 10.1109/MIPRO.2015.7160470.

Eecs.qmul.ac.uk. 2020. *Spurs And Bits Predicting Football Result Using Bayesian Nets And Other Machine Learning Techniques*. [online] Available at: <https://www.eecs.qmul.ac.uk/~norman/papers/Spurs-2.pdf>

He, M., Cachucho, R. and Knobbe, A. (2015) 'Football player's performance and market value', *CEUR Workshop Proceedings*, 1970(September 2015), pp. 87–95.

Huang, K. Y. and Chang, W. L. (2010) 'A neural network method for prediction of 2006 World Cup Football Game', *Proceedings of the International Joint Conference on Neural Networks*. IEEE. doi: 10.1109/IJCNN.2010.5596458.

Hucaljuk, J. and Rakipović, A. (2011) 'Predicting football scores using machine learning techniques', *MIPRO 2011 - 34th International Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings*. IEEE, 48, pp. 1623–1627.

Korotyeyeva, T., Tushnytskyy, R. and Kulyk, V. (2018) 'Applying neural networks to football matches results forecasting', *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018 - Proceedings*. IEEE, 1, pp. 278–282. doi: 10.1109/STC-CSIT.2018.8526756.

Madhuri, C. H. R., Anuradha, G. and Pujitha, M. V. (2019) 'House Price Prediction Using Regression Techniques: A Comparative Study', *6th IEEE International Conference on "Smart Structures and Systems", ICSSS 2019*. IEEE, pp. 4–8. doi: 10.1109/ICSSS.2019.8882834.

Muthukrishnan, R. and Rohini, R. (2017) 'LASSO: A feature selection technique in predictive modeling for machine learning', *2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016*. IEEE, pp. 18–20. doi: 10.1109/ICACA.2016.7887916.

Pariath, R. *et al.* (2018) 'Player Performance Prediction in Football Game', *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018*. IEEE, (Iceca), pp. 1148–1153. doi: 10.1109/ICECA.2018.8474750.

Rahman, M. A. (2020) 'A deep learning framework for football match prediction', *SN Applied Sciences*. Springer International Publishing, 2(2). doi: 10.1007/s42452-019-1821-5.

Ulmer, B. *et al.* (2014) 'Predicting Soccer Match Results in the English Premier League', *University College London*, 38, pp. 1–17. Available at: http://arxiv.org/abs/1411.1243%0Ahttp://cs229.stanford.edu/proj2014/Ben Ulmer, Matt Fernandez, Predicting Soccer Results in the English Premier League.pdf.

Wagenaar, M. *et al.* (2017) 'Using deep convolutional neural networks to predict goal-scoring opportunities in soccer', *ICPRAM 2017 - Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, 2017-Janua(February), pp. 448–455. doi: 10.5220/0006194804480455.

Bhattacharyya, S., 2020. *Ridge And Lasso Regression: L1 And L2 Regularization*. [online] Medium. Available at: <https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b> [Accessed 2 August 2020].

Malik, U., 2020. *Random Forest Algorithm With Python And Scikit-Learn*. [online] Stack Abuse. Available at: <https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/> [Accessed 14 July 2020].

Glen, S., 2020. *Lasso Regression: Simple Definition - Statistics How To*. [online] Statistics How To. Available at: <https://www.statisticshowto.com/lasso-regression/> [Accessed 14 July 2020].

SHARMA, P., 2020. *Guide To Hyperparameter Tuning, Regularization, Optimization*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2018/11/neural-networks-hyperparameter-tuning-regularization-deeplearning/> [Accessed 14 July 2020].

www.fifa.com. 2020. *Who We Are - News - FIFA Survey: Approximately 250 Million Footballers Worldwide - FIFA.Com*. [online] Available at: <https://www.fifa.com/who-we-are/news/fifa-survey-approximately-250-million-footballers-worldwide-88048> [Accessed 7 August 2020].

Xgboost.readthedocs.io. 2020. *Xgboost Parameters — Xgboost 1.2.0-SNAPSHOT Documentation*. [online] Available at: <https://xgboost.readthedocs.io/en/latest/parameter.html> [Accessed 14 July 2020].

Loy, J., 2020. How To Build Your Own Neural Network From Scratch In Python. [online] Medium. Available at: <https://towardsdatascience.com/how-to-build-your-own-neural-network-from-scratch-in-python-68998a08e4f6> [Accessed 12 August 2020].

Malik, U., 2020. Creating A Neural Network From Scratch In Python. [online] Stack Abuse. Available at: <https://stackabuse.com/creating-a-neural-network-from-scratch-in-python/> [Accessed 5 August 2020].