FANTASY PREMIER LEAGUE'S PLAYER PREDICTION USING MACHINE LEARNING ALGORITHMS

Dissertation submitted in part fulfilment of the requirements for the degree of Master's in Data Analytics at Dublin Business School

Shreyas Bagur Srinivasan

Student ID: 10383856

Dublin Business School

Supervisor(s): Terri Hoare

Master's in Data Analytics 2018

Declaration

Declaration: I, Shreyas Srinivasan, declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this work and this this work is fully compliant with the Dublin Business School's academic honesty policy.

Signed: Shreyas Srinivasan
Date: 07-01-2019

Dedication

Firstly I am dedicating this dissertation work to my parents , without their support and motivation this could not be possible.

I also dedicate my dissertation Vaishnavi and Kavitha who was always there to support and motivate me.

Table Of Contents

List of Figures

.

List of Algorithms

Foreword

The major motivation for this research is the game which has always been my area of interest and the game which I played the most. Doing things in which you’re good at was the only well thought approach in conducting this research thus, making it worthy enough in investing time.

Acknowledgements

Firstly, I would like to thank my thesis supervisor Mrs Terri Hoare of Dublin Business School. My supervisor has always been helpful whenever I had any doubts about my dissertation. She constantly supported and believed in me and let me approach the challenges in my own way and at the same time steered me to the right direction when required.

I wish to express my sincere thanks to Prof. John O'sullivan and Prof. Shahram Azizi for the knowledge they shared with me which helped me to understand the concepts of Data Mining and Machine Learning. They also encouraged me and provided insights which were very helpful in my dissertation.

I am also very grateful to my college Dublin Business School for providing me with excellent facilities for my thesis and thank my fellow classmates Abhinov, Akhilesh and Tushar for the stimulating discussions and constant feedback.

Finally, a profound gratitude to my family who have supported and motivated me constantly. This accomplishment would not be possible without their support.

Thank you all for your encouragement!

Shreyas Bagur Srinivasn

Abstract

According to recent studies, sports produce significant statistical information about each player, team, games, seasons and the numbers about the fantasy games . In order to improve the performance of sports, fantasy games provides scientific principles and techniques. Fantasy premier league is one such fantasy game where the data can be analysed and a lot of predictions can be made. This is one major platform for analysis because nearly 6 million people play this game. It is useful for each user to create a best team so that he gets the highest points each. Manager/User having the most points at the end of the season is declared as the winner and gets a free trip to watch his Favourite English Football Club play live. For this to be a success there is a lot of ideas and logic to be used behind the scenes so as to get the highest point each week before becoming the winner. Since it is an herculean task , this prediction model could be the one which would be able to help the users to reduce the complexity for thinking who has to be removed from the team so as to get in a player who can do the best and give the required point to stay on top of the league.

Keywords: FPL, Football, Best Team, Prediction

List of important abbreviations

ANN- Artificial Neural Network

FPL- Fantasy Premier League

BPS- Bonus Point System

SVM- Support Vector Machine

ICT- Influence , Creativity , Threat

1.Introduction

Football as a sport has gained popularity in the past few decades and the numbers of fans following football are increasing rapidly. As almost every nation has its very own league of football, the fan base of the game is increasing day to day. The recent survey tells us, leagues of the European countries are most followed by the fans and they have some of the best teams in the world. From a recent survey in England, reportedly there are around 315m fans around the world supporting different English clubs. With these number of supporters, people tend to play the best fantasy football game, the Fantasy Premier League (FPL). Due to the increasing supporters of the game, the fantasy game has also gained a huge popularity and the people playing this game has been increased from 5 million to nearly 6 million this year from a survey conducted by the fantasy game managers. FPL is like any other fantasy type game where the people must build their own team and the players earn points based on their performance in the live league games. In FPL, we get to choose 15 players from the 20 English clubs with the following constraint that the out of the 550 or more players in the premier league we must choose only 2 goalkeepers, 5 midfielders, 5 defenders and 3 strikers.

So, how does this game work? It is quite simple, and it all starts with choosing a team, once you choose the team satisfying all the constraints then the game moves to the stage where your player earns you points. Before every game week starts, we must submit our team and the points we get are based on how the players play that week and the points are given accordingly. There are certain ways points are given to the players and teams, each position will be awarded points according to the rules that have been set and there are quite a few of them. This game is classic league game where each week the player gets some points and the points are added for the next 37 game weeks and the person owning the team with the most number of points wins the game.

1.1 Background

Rules of the game:

You will be given 100 million at the start of the season and you must choose players accordingly i.e. 2 GK, 5 DEF, 5 MID and 3 FWD. The rules for buying players are, maximum of 3 players can be bought from a single team and the team budget should not cross 100 million and there is no constraint that you must use all the 100 million you have, you can save some too in the bank. Before the start of first game week you can do unlimited changes to your team,

but after game week one you will have only 1 free transfer available to change any one of your players, if you do a second transfer 4 points will be deducted from the total points you get in the next week. If you wish to save the free transfer you have, for next week you will have two free transfers available. After having two free transfers if you wish to save it for the next week again the free transfers will be reset to one.

Each position in the game has different points system i.e. there is different points scale for goalkeepers, defenders, midfielders and strikers and each positions point are assessed on different attributes. All the positions have two attributes points scale in common i.e. the number of minutes played, if a player has played less than 60 minutes, he gets 1 point, if player has played more than 60 minutes, he gets 2 points. Second attribute is "Yellow" and "Red" cards, if a player gets a yellow card one point will be deducted from his overall points and if player gets a red card 3 points will be deducted from his overall points. Except strikers position all the other positions is assessed on 8 attributes but the points scale for different positions are different. Every week there will be a change of players price on the basis of his performance in the last game week. The price changes is main based of the number of transfers in/out of a particular player for that current week. If the player has been transferred in by a lot of players playing his price will be increased by 0.1 million and the price is decreased if the player is taken out from the team by most players or in a simpler way the price change is dependent on the attribute "Total Selected By (TSB)" in the fantasy game. In the above sentence, 'player' refers to the players playing the game and 'players' refer to the people/fans who are playing the fantasy game.

Attributes on which the players are assessed:

- Clean Sheet (Except Strikers)
- Number of goals conceded (Except Strikers)
- Assists made
- Goals scored
- Penalties Saved (Only Goalkeepers)
- Penalties missed (Only Strikers)

Points Scale

Goalkeepers:

- For keeping a clean sheet, they get 4 points

- For conceding 2 goals one point is deducted, and -1 for the next two goals conceded
- For scoring a goal, they get 6 points
- For assisting for a goal, they 5 points

Defenders:

- For keeping a clean sheet, they get 4 points
- For conceding 2 goals one point is deducted, and -1 for the next two goals conceded
- For scoring a goal, they get 5 points
- For assisting for a goal, they 4 points

Midfielders:

- For keeping a clean sheet, they get 1 point
- For scoring a goal, they get 4 points
- For assisting for a goal, they 3 points

Forwards/Strikers:

- For scoring a goal, they get 3 points
- For assisting for a goal, they 2 points
- For missing a penalty 2 points is deducted from their overall points

Every week one player in out team will be named as the captain and his points will be doubled.

Excluding these, the game also has 2 Wildcard options, 1 Bench Boost option, 1 Triple Captain and 1 Free Hit and these are called as chips and these chips can be used anytime according to the rules throughout the season.

Wildcard

This is one chip which can be used before December and one more chip which is available after December which can be used anytime once before the season ends. This wildcard gives the players the option to change the whole team with infinite number of transfers and without deducting the points for transfers after 1/2 free transfers is used up.

Bench Boost

This chip can be used once in the whole season, here all the 15 players points are taken into consideration while counting the points of the player.

Triple Captain

Every team has a captain in their team, when this chip is used his points gets tripled for that particular game

Free Hit

This chip can be used to change the whole team for one particular week, this also works like wildcard having infinite numbers of transfers for that week.

BPS

Other than all these , there is something called as bonus point system(BPS). This gives the players extra 3,2or 1 points after the game. And only 3 players will get this bonus points , it might be all the 3 players from the same team or they might be from the opponents as well. These bps work in a unique way and this has not been used anywhere else except football fantasy or other official fantasy games. Bps is a live score that keeps on changing according to how the player pays the game. This bps system is calculated upon different factors for different positions of players on the pitch.

Goalkeepers rating are mainly done on the basis of their saving skills and how many saves do they on an average in any match and if the keeper saves a penalty there are chances for him to get all the three points but the keeper getting all three points is a very rare case. This mostly happens when the game is a draw or the keeper has exceptionally played well than the rest of the team.

Defenders rating are mainly based on their skills in the defence like , how good they are tackling , how good are the interceptions , how is their duel wins ratio. They get higher score usually when they assist a goal or score themselves which is a likely thing in modern football. When compared to GK's they have a higher chance of getting all the three points.

Midfielders rating are mainly based on their game play because midfielders are the heart of a football team as most of the things are dependent on them. They re the playmakers of the game and some of the best players in the work dare playmakers. Their ratings are based on how well they play around with the ball and help in the team doing well and also scoring goals. They have the highest chances of getting all the 3 points along with strikers.

Forwards/Strikers ratings are assed on the basis how they play in front of opponents goal. Their ratings are mainly given on the basis of goals scored or assists given. By chance if they miss a

penalty given to them their rating will be decreased but that is rare case. They also along with midfielders have a good chance of getting all the 3 points from the bps.

1.2 Business Background

Seeing this in terms of business perspective, improving the accuracy of the prediction using different technologies would be helpful for the users to make a best team which would in turn help the user to win a fully paid trip to see their favourite team play live, which is still dream to a many football fan.

2. Objective

The main aim of this research is to make a model which will predict the " player who has to be removed from the team so as the get the best player for the next game week so that we will have the best playing XI and will get the most number of points. Therefore, to do this, understanding the different influencing factors and getting to know the previous study conducted in the field remains as a vital aspect in the overall research.

2.1 Research Question

Can a better prediction system help users in getting to the know who is the player that can be replaced in order to bring in a best player so that the current team gets the most number of points in the that game week

3. Literature Review

3.1 Scope

This research generally considers in referring to the findings of previous researchers for a decade. By doing this, it provides an overall scope and a niche in understanding the importance of making the predictions. The reason of absolutely considering most of the papers from 2010 to 2018 is because from past few years the game has been at its peak spreading its craze across the globe where millions of people have started playing the game. So, getting to know the

importance of predictions during this span makes the subject worthy enough in considering some of these research papers.

3.2 Existing Methods

In 2016, Sidhartan Selvaraj (2016) focused on what are the factors that are needed for a player to be in the playing eleven when the team is being selected and also the players performance matters for selecting him into the team. Some of the factors that the author has discussed are the skill set of the player, previous performances and the results of the player during the medical and training before the games begin. This is one of the base model that can be used to select the player who can be bought in for that particular game week. Author has used various machine learning algorithms to predict the player rating on the factors that are required for the player to be selected to play game on that particular day. Author has used Random Forest and ANN to predict the player ratings , he also integrated H2O with these two algorithms and has compared how the model works when it is not integrated with H2O and how does it work when it is integrated with H2O.He also concluded that the computational time when H2O was integrated was less when compared to the algorithms not integrated with H2O From the results author was able conclude that the algorithms when integrated with H2O gave good performance and the RMSE values were comparatively less when compared to the algorithms not integrated with H2O

Dr. Manjula Sanjay et. al (2016) had predicted how to pick the best team from a set of given players. Here they had the data of one whole team with all their attributes and the respective rating for the attributes, since we had to choose players for each position, each position had different attributes to choose from. Data Mining techniques were used here to select the best for each position and to select him for the playing XI only if he had no injuries at that particular time. This would help me a lot in this research as we would have chosen fifteen players from all the twenty clubs that are playing in the league. Here Tangera tool was used to choose the best player from the given lot in the dataset according to his ratings on the attribute on which he is chosen into the team.

S Drawer et.al (2002) focused on the injuries of the English professional football players using a risk based assessment process. The author in this paper is mainly focused on the levels of injury that a player can get and how much risky is that injury for the player to play again in the near future. Drawer has differentiated the injury levels in to 5 different categories namely , Slight(just a knock or the smallest of injuries which can be cured within few days), Minor(

this might something slightly more than a knock and this might take maximum of a weeks time to recover), Moderate(these injuries mainly might be some small problems related to body of the player like wrist injury , back pain , lowest level hamstring , twisted ankle etc., this usually takes 2-3 weeks to get cured) , Major (these injuries are big problems to the players where they will be side lined for weeks together, these injuries mainly are : higher levels of hamstring , knee pain / injury , shoulder injury etc..) and the last and the major thing Fatality (these are the injuries which nearly decide the careers of players telling them whether they can play again or not. These injuries require major surgeries and might require months for it to cure. Broken bones, shoulder dislocations are the main injuries in this case.). From the injury type we would get to know how long would a player be side lined and we can also see that how prone the player is for injuries, this helps the users a lot in selecting a player into the team. In simpler ways if the player is injured and is side lined for more number days, we would not pick him in the team and if the player is available for selection see all the other constraints, we would pick him in the team.

Jhawar et al (2016) focused on predicting the results of a match by gauging the strengths of the 2 teams. For this, the performances of the individual players were calculated of both the teams. For modelling they developed algorithms for the performances of different players at different positions of play , they confirm the potential of a player by examining his career performance and so his recent performances. (Jhawar and Pudi, 2016)

Singh et al (2017) used the increasing range of matches day by day, to extract all the information and manage it very tough because of the number of matches will be increasing every week. They present an information mental image Associate in Nursing prediction tool within which an ASCII text file, distributed, and non-relational information, HBase is used to stay the information associated with matches and players. This information is then used for visualizing the past performance of players' performance. In addition, the information is employed to predict the end result of a match through varied machine learning approaches. The planned tool will prove helpful for the team managements within the player auctions for choosing the correct team. They address the matter of predicting the result of associate match and conjointly the player identification system which might be a good facilitate for the team leaders on the auction day. The statistics of various matches are employed in the experiments. Factors like luck and player strength are used as key options in predicting the winner of a match. This can be mainly used to predict the winner of the match and trying out that since that

team is going to win, what are the chances that if we have a player from that winning team he will give us points to the dream team.

Ahmad, H. et al (2017) used online social databases square measure wealthy sources to retrieve acceptable data that's after analysed for forthcoming trends prediction. During this work, they tend to determine rising stars in cricket domain by using machine learning techniques. Additional exactly, they tend to predict rising stars from batting additionally as from bowling realms. For this intent, the ideas of Co-players, Team and Opposite groups square measure incorporated and distinct options at the side of their mathematical formulations square measure conferred. For classification purpose, generative and discriminative machine learning algorithms square measure used, and 2 models from every class square measure evaluated. As a symbol of pertinence, the projected approach is valid through an experiment, whereas analysing the impact of individual options. Besides, model and class wise assessment is additionally performed. Using cross-validation, they tend to demonstrate high accuracy for rising star prediction that's each sturdy and statistically important. Finally, ranking lists of high 10 rising cricketers supported weighted average, performance evolution and rising star scores square measure compared with the international rankings. Measures are expressly adopted for rising star prediction in batting and bowling domains. A lot of exactly, 3 classes (Co-players, Team and Opposite teams) are incorporated, within which nine and eleven options are outlined for the prediction of batting and bowling rising stars, severally. 2 styles of datasets are generated supported weighted average and performance evolution metrics.

Onwuachu et al (2015) was focused on how to select a player in to the starting XI by the help of machine learning algorithms. He used 4 neural networks and their decisions in order to pick player to the starting XI. He took Player’s Resistance, Player’s Speed, Player’s Physical and Player’s Technique as the major factors that will make the final decision whether the player will be in the stating XI or not. For these four major categories he built neural network model respectively and observed the results. In his algorithm he used a formula for finding the average stats of the player i.e. sum of all the four major factors divided by the total number of factors. If the average stat of the player was found to be less than 50 then the player will be rejected and if the rating is above 50, now the manager can use other information and get the player in the starting XI. The simulation for this process was done using MATLAB. Using all the information and Neural networks algorithm ratings for all the 4 major categories was found and the average stats was found. The ratings were from 0-100, 100 being the best which can be achieved by any player but that is impossible practically. He has showed that neural network

is one of the best and proven to give the best performances for the predicting the players whether they are selected in the team or not. The neural network technique can help the football managers in selecting the team. This is one of the main reasons why we are using the neural network technique to predict the player who has to be removed.

Goddard, J. and Asimakopoulos, I., (2004) talks about forecasting football results of fixed-odds betting estimated based on 10 years data using a probit regression model. This is one of the first papers to quantify the predictive quality for not only the past matches but also for the other numerous explanatory variables. The model was used to test the efficiency of prices of fixed-odds in the betting market and the structure of which is on the top 20 teams from the Premier League (PL) along with the next 72 teams competing in 3 league football league divisions of 24 teams per division (Goddard, J. and Asimakopoulos, I., 2004). The paper also looked at the home win, draw or away win mainly using regression model that was created as the probit model having the main performance indicators as recent match results and other explanatory variables significant for championship, promotion and regulation also checking on the efficiency of the prices in the fixed-odd betting among which the regression-based and economy based weak form efficiency tests were also conducted. The results compared were on seasonality basis for the years. Many factors contribute to the performance of the forecasting model such as match for championship, promotion or regulation issues, involvement of teams, geographical distance of home towns of the teams, etc (Goddard, J. and Asimakopoulos, I., 2004). Goddard, J. and Asimakopoulos, I., (2004) take a statistical approach and their probit model is easier to implement than other forecasting models and their model seems to achieve a comparable forecasting performance and results and model based on regression also has been used to test the weak form which also indicates the forecasting model contain information about match outcomes later looking at efficiency for final few seasons where using the models probability based on the exante expected return is generated with a positive result of around +8% for the matches plated in that season around the April and May months for both the seasons of years 1999 and 2000 (Goddard, J. and Asimakopoulos, I., 2004). The prediction of team winning a game is helpful for the FPL players to choose the player to buy for a player who is transferred out satisfying all the constraints. The user will be able to choose the best player according to his constraints from the team which is winning giving him points where his team could end up with the most number of points for that game week.

Chenjie Cao el at (2012) focused on the match prediction using data mining techniques, in order to complete his task author used 4 algorithms to predict the outcome of the games and

used previous 5 season information as the data for completing his task. The four algorithms he used is Simple Logistic Classifier, SVM, ANN, and Naïve Bayes. Here in this he is mainly focusing on how the teams play and line up in order to predict the outcome of the game. It depending up on who are the players in the position assigned in the game and how strong they are at their position and how can they be able to their job in the opponent half and your half as well when defending. Like the other games dependent on he home away , even in this the similar type of factors are used in prediction like , where are they playing the next game and also all the statistics of the teams that are playing is required such as how are the performances of both the teams in their previous matches, win/ loss ratio of the two teams is taken in two categories, one is the teams overall win/loss ratio and the other home teams win/loss ratio at home ground. He has taken one more major factor to determine the winner is that , the number of days between the matches played by the teams , from this it is concluded that if there is no gap or one day gap between the matches the win ratio decreases and if the gap between the matches is 2 days or more the chances of winning are high. The same can be applied to football when coming to prediction, when there are continuous matches with gap of day or two , there are chances of the team winning is less and indirectly telling that a particular player has to be removed from the team.

Gary, G., (2007) focused on the strengths between the national teams, how the factors influence the team from being the best/strongest when compared to another team. Firstly, the OLS Regression model was used to predict performance of the 201 international teams and from these performance ratings the rankings to the teams were given. From these Regression result he got to a conclusion that the strength of a team is dependent upon number of men who play football regularly in the nation. The results also showed that not only the number men play is responsible for the strength of the team depends upon other factors of the nation too. There are some dependent as well as independent variable on which the strength of the team depends on. Some of the variable are Size of the pool, this is one of the main factor which tells how strong the team is because , more people playing and training football , there are chances more getting more talent when there are more people , since there are more talented people in the pool the team gets stronger, when the model is run for the size of the pool if the values of R^2 and RMSE values are high and less respectively indicate that the model is fit. Next factor is a major factor because the strength of the team is dependent on how football has evolved in a particular country, meaning if a country has rich history of football and they have been playing from years, they would be having a lot of talent in them as they are playing the game from a long

time. Economic resource is another factor where the strength of the team is dependent on, if a county is less wealthy or is not able have good resources or they are economically facing problems , they would have a team which is less stronger to the teams where their county can afford the resources to play football and helping build the more stronger so the GDP of the country is the factor deciding the strength. Author has also spoke about the players playing abroad and gaining experience, lower ranked teams or less strength teams lack in quality of play as their players mainly play in and around themselves other than going broad where they can learn skill which might help in strengthening their national team , from the results we can see that 40.9% of the players have played football abroad , in there 40.9% of the players 76.9% of the players have played in a wealthier than their or 86.2% of the players have played in a country which is ranked above them in the FIFA rankings. Factors here also help is in the prediction of FPL like how good the teams pool is i.e. The depth of the squad and their academy players who can be future stars, the clubs financial strength is also one of the main factor as we can buy hugely talented players if the club has good economy and that player's inclusion would bring strength to the squad.

Doanna, O.,(2016) focused on the talent identification and selection of players in the team. Talent identification is mainly based on the managers and scouting people who rate the players according to the skills and performances of players shown in the matches or during training. The author is predicting the team selection by conducting tests on young talented professional players who play for their nation. Out of the 127 youth talents who played in the tournament 22 players got a full time recruitment for the clubs to play for the senior teams. So he conducted the tests on the players who got selected and few other volunteers to get to know the difference between them and the others, from the tests he made he got to know that there some factors which influenced the players to play well and get the selection for the senior teams. The factors that affected players gameplay/performance were, the region whey they were trained on the game, how often they used to play, from when did they start to get interest on the game, how often they trained and how did their performance affect the outcome of the game. Once the managers/scouting people analysed the result they trained the same 127 players according to the factors affecting the game and they were re tested again , from the new analysis they got a higher accuracy where instead of 27 players , 50 players got selected and from the results of the ANOVA test made we could conclude that some attributes had a lesser effect of the selection of the players and some had more effect like decision making , match play and cumulative performance. These factors/attributes that are affecting the players selection plays

an important role in the FPL team to make the decision whether to keep they player in the team or transfer him out for another player who is more talented than him or a better player who can give more points in the game week.

Rabah, A., (2017) is focusing of the automated team prediction using competitive Neural networks. He has majorly use two factors to predict the best team that is selected. Firstly, he has calculated the player rating and then he has done the selection of the team according to the player ratings. The predictive model generates the contribution of player in the game individually relative to his contribution to the team. Since the naïve approach of the win/loss might be similar for many players, so he used a neural network in order to find the player importance. Author assigned the input links of each player to the model so that he could get n evaluation for each player in the process. For the team selection process, the result of the player rating evaluation is one of the most important one as the maximum player rating is taken from all the positions of play. After selecting the best players from their respective positions we sum up the ratings value and then see how good the team is from the summed up value we get. This approach is done for both the home team and the opponent team. By doing this we will get to know the combination for the home team which should be stronger than the opponent and also giving the home more combinations to play. A neural architecture was built for the players rating and the input was the players and matches and the target variable was the outcome of the match for the home team with set of playing XI. For the given set of data in the first trail he achieved 54% accuracy, and in the second run of the model he only reduced the model to 22 players, 11 of each team, players selected from the output of player ratings, he achieved 60% accuracy. From the results he concluded that this model can give the best team for selection in the 4-4-2 formation which is 4 defence,4 midfield and 2 strikers. This model is one of major factor influencing my model as the data is taken from English Premier League players and the best team is selected from the basis of best player rating. In FPL that is how the logic runs keeping the best rated player and remove the least good player and get in a better player than the transferred one so as the team will be strong compared to others.

Kou-Yuan, H., (2010) focused on neural network architecture of predict the winners of the game in FIFA 2006 world cup. He predicted the games based on the strength of the team and 8 features that the match is significantly dependent on. He got the values of these 8 factors and normalized them into values that ANN can process on and then built neural network architecture for the predicting the winning rate. He tested the dataset 3 times and then took the average of them for the final ANN process to predict the winner. He built the model for each

stages of the world cup , and he concluded that the stages having more number of teams has the highest accuracy and as the number of teams decreases the accuracy also decreases as there will be less data and as the number of teams decreases the strength of teams will also be similar which will be tough to predict and also the game is played on foot and some of the factors and the constraints tend to change in some situation , due to these changes we cannot be able to predict the games so accurately, For example , In a football match between a strong side and less strong side if the less strong side score a goal in the staring seconds of the game and then defend the whole match it is difficult to predict this as the teams scoring in the starting is very less or negligible. So in general prediction of the winner of football match is tough, but through the machine learning algorithms and from the dataset we get , we can predict this to slightly bigger extent having 60-75% accuracy which is a good number compared to number of matches player all over and the upsets made by the less stronger teams on the stronger teams compared to them. Similar to the other research topic this helps in getting the best player from a winning team so that he can give me more points so that I have the most points every week.

Torben, T., (2010) focused on the assessment of the players performance in German Bundesliga. He used two performance processing methods DEA and SFA respectively to assess the players performance. He has taken 5 main factors that help in predicting the performance. The main factors that affect the performance are playing time , goals , assists , tackle ratio and passing accuracy. From these factors we get to know how good the player performs in the match. So from this research I can get to know the player performs who is in my team and will get to know who can I keep and who can I transfer out.

Hence, we can predict that each model has its own importance in inferencing us that which player plays well and what does his performances mean to the team. Thus, when going through all these studies, we can understand that how vital it is to consider each, and every factor associated with historical data to have a successful prediction accuracy. Thus, making an impact on people with accuracies worthy enough to be considered.

4.Methodology

There is a lack of specific and detailed framework for conducting data mining analysis. Cross Industry Standard Process for Data Mining (CRISP-DM) presents a hierarchical and iterative process model and provides an extendable framework with generic-to-specific approach, starting from six phases, which are further detailed by generic and then specialized tasks. Cross Industry Standard Process for Data Mining (CRISP-DM) defines following data mining context dimensions: application domain, problem type, technical aspect, and tools & techniques. Data mining processes such as the CRISP-DM have been developed as a guideline for data mining projects (Shearer, 2000). However, such processes have been developed prior to the 'data boom' age are without due cognizance to the amount and multi-structured nature of data generated by modern information systems. Methodological processes such as CRISP-DM could benefit from additional guidelines for generating insights from large volumes of electronic social network (SN) data.

The CRISP–DM is used in this article, which consists mainly of the following tasks:

• Business Understanding

The first stage of the CRISP-DM process is to understand what you want to accomplish from a business perspective. Your institution may have challenging objectives and limitations that must be properly balanced. The goal of this stage of the process is to discover important factors that could impact the consequence of the project. Overlooking this step can mean that a great deal of effort is put into making the right answers to the wrong questions.

• Data Understanding

The second stage of the CRISP-DM process requires you to obtain the data listed in the project. This initial collection includes data loading. For example, if you use a specific tool for data understanding, it makes the right sense to load data into this tool. If you get multiple data sources then you need to consider how and when you are going to integrate these.

• Data Preparation

This is the stage of the project where you decide on the data that you are going to use for the analysis. The conditions you might use to make this decision include the importance of the data to your data mining goals and the outcome u wanted to get, the quality of the data, and

also technical conditions such as limits on data size or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

• Modelling

As the first step in modelling, you'll select the real modelling technique that you'll be using. Although you may have already selected a tool during the business understanding phase, at this stage you'll be selecting the specific modelling technique e.g. decision-tree building with C5.0, or neural network generation with back propagation. If multiple techniques are applied, perform this task separately for each technique.

• Evaluation

Past assessment steps managed factors, for example, the precision and all inclusive statement of the model. Amid this progression you'll surveys how much the model meets your business goals and try to decide whether there is some business motivation behind why this model is lacking. Another choice is to test the model(s) on test applications in the genuine application, if time and spending limitations allow. The assessment stage additionally includes surveying some other information mining results you've created. Information mining results include models that are essentially identified with the first business destinations and every single other finding that are not really identified with the first business targets, but rather may likewise reveal extra difficulties, data, or clues for future bearings.

• Deployment

In the organization arrange you'll take your assessment results and decide a methodology for their sending. On the off chance that a general technique has been recognized to make the significant model(s), this strategy is archived here for later arrangement. It bodes well to consider the available resources of sending amid the business understanding stage also, in light of the fact that arrangement is totally significant to the accomplishment of the undertaking. This is the place prescient examination truly enhances the operational side of your business.

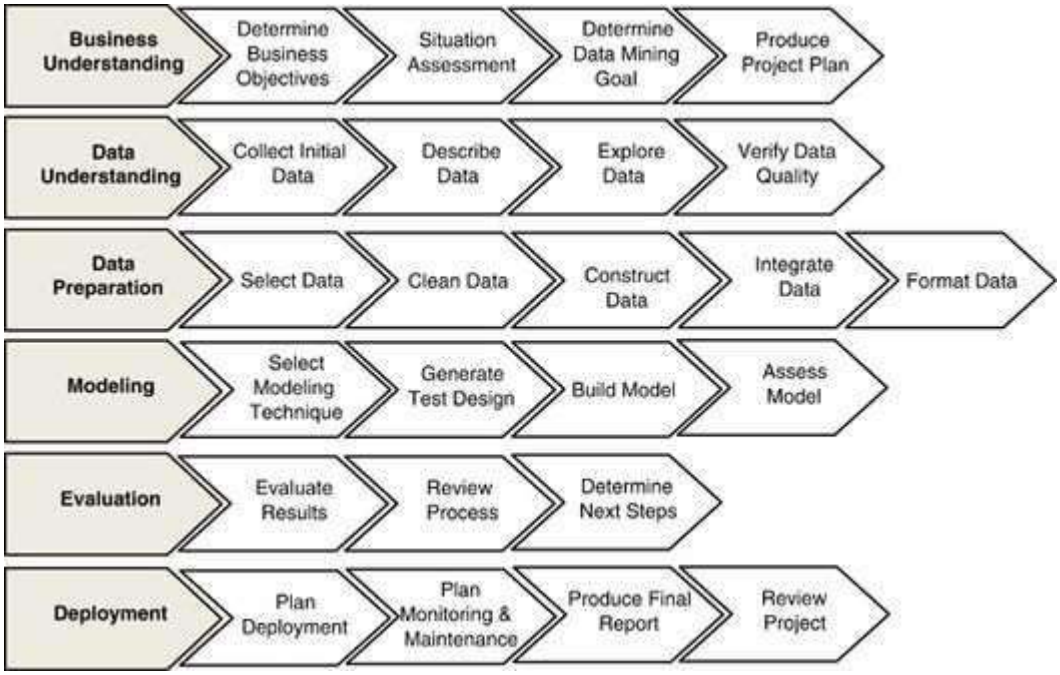
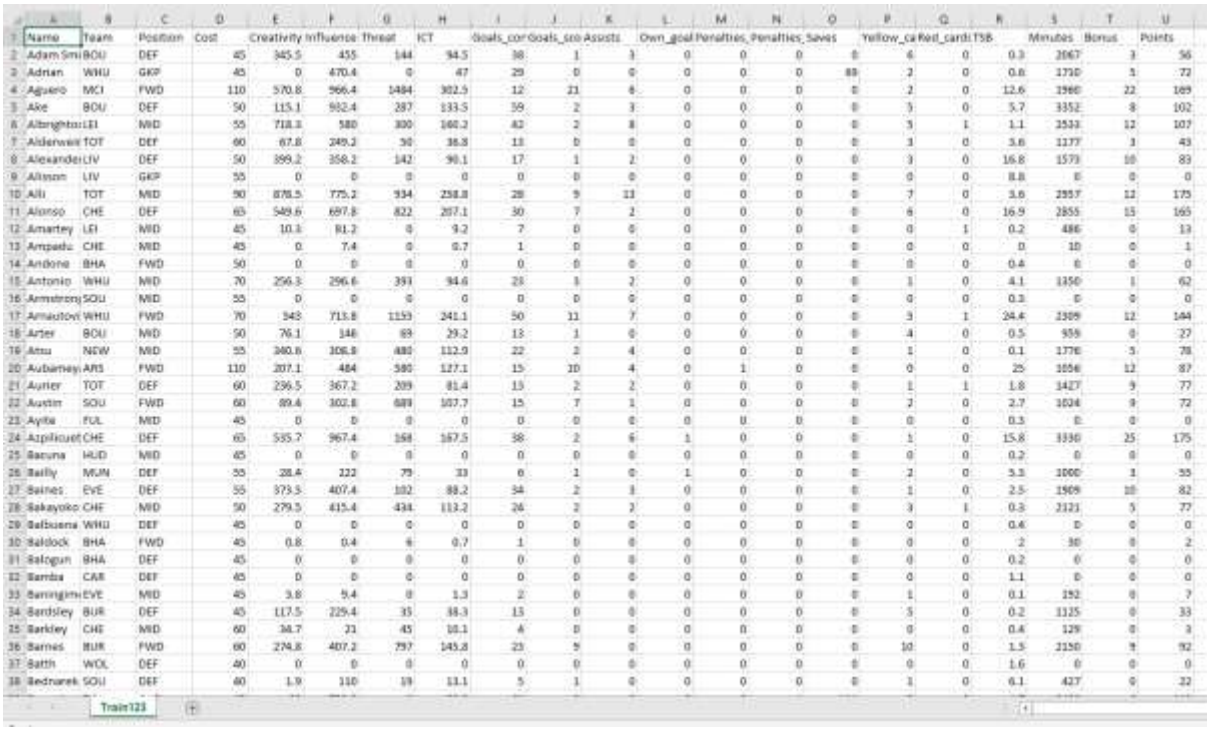


Figure 4.1 : Explanation about Crisp-DM process

Whereas the approach is generalizable for most forms of data analytics, it does not adequtely respond to more recent research hurdles posed by large volumes of data. There are some recent studies such as the HACE theorem to model big data characteristics that were developed and there have also been studies that have been produced which provide a generalizable and yet parsimonious approach to managing complex data patterns as generated on electronic SNs (Wu, X., Zhu, X., Wu, G.-Q., and Ding, W., 2014). (Tinati, R., Halford, S., Carr, L., and Pope, C., 2014) notes that in social research, methodological hurdles to manage huge amount of data puts a cap on the extent to which ontological and epistemological questions can be asked. In this study we recognize the lack of a sound methodology that drives the conceptual and analytical questions posed to large volumes of data extracted from electronic SNs. This lets us ask the question "what is the process or method that will leverage large volumes of data in a social science research?" Adapting CRISP-DM Process for Social Network Analytics Twenty-first Americas Conference on Information Systems (Asamoah, D. and Sharda, R., 2015). Because of this we propose CRISP-eSNeP, an extension to the CRISP-DM methodology, as a guideline to manage, analyse and generate insights from large volumes of data acquired from electronic SNs.

4.1 Data Understanding

For the purpose of this research we have the data of all the 528 players who are in the 20 premier league clubs. The dataset was obtained from the open source domain Kaggle.com. we have used feature selection to determine the factors that affect the points of the player every week. In the dataset we have 528 rows mainly names of the players and 20 columns. Out of these 20 columns by conducting feature selection we came to know that there 7 attributes that are the factors affecting the points of the player which is the target variable.



Name	Team	Position	Cost	Creativity	Influence	Threat	ICT	Goals_cor	Goals_sco	Assists	Own_goal	Penalties_	Penalties_	Saves	Yellow_ca	Red_cards	TSB	Minutes	Bonus	Points
Adam Smi	BOU	DEF	45	345.5	455	144	94.5	38	1	3	0	0	0	0	6	0	0.3	2067	3	56
Adrian	WHU	GKP	45	0	470.4	0	47	29	0	0	0	0	0	88	2	0	0.8	1710	5	72
Aguero	MCI	FWD	110	570.8	966.4	1484	302.5	12	21	6	0	0	0	0	2	0	12.6	1960	22	169
Ake	BOU	DEF	50	115.1	932.4	287	133.5	59	2	3	0	0	0	0	5	0	5.7	3352	8	102
Albrighton	LEI	MID	55	718.3	580	300	160.2	42	2	8	0	0	0	0	5	1	1.1	2533	12	107
Alderweir	TOT	DEF	60	67.8	249.2	50	36.8	13	0	0	0	0	0	0	3	0	3.6	1177	3	43
Alexander	LIV	DEF	50	899.2	358.2	142	90.1	17	1	2	0	0	0	0	3	0	16.8	1573	10	83
Alisson	LIV	GKP	55	0	0	0	0	0	0	0	0	0	0	0	0	0	8.8	0	0	0
Alli	TOT	MID	90	878.5	775.2	934	258.8	28	9	13	0	0	0	0	7	0	3.6	2557	12	175
Alonso	CHE	DEF	65	549.6	697.8	822	207.1	30	7	2	0	0	0	0	6	0	16.9	2855	15	165
Amartey	LEI	MID	45	10.3	81.2	0	9.2	7	0	0	0	0	0	0	0	1	0.2	486	0	13
Ampadu	CHE	MID	45	0	7.4	0	0.7	1	0	0	0	0	0	0	0	0	0	10	0	1
Andone	BHA	FWD	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0
Antonio	WHU	MID	70	256.3	296.6	393	94.6	23	1	2	0	0	0	0	1	0	4.1	1350	1	62
Armstrong	SOU	MID	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0	0	0
Arnautovi	WHU	FWD	70	343	713.8	1153	241.1	50	11	7	0	0	0	0	3	1	24.4	2309	12	144
Arter	BOU	MID	50	76.1	148	69	29.2	13	1	0	0	0	0	0	4	0	0.5	959	0	27
Atsu	NEW	MID	55	240.6	306.8	480	112.9	22	2	4	0	0	0	0	1	0	0.1	1776	5	78
Aubameyo	ARS	FWD	110	207.1	484	580	127.1	15	10	4	0	1	0	0	0	0	25	1056	12	87
Aurier	TOT	DEF	60	236.5	367.2	209	81.4	13	2	2	0	0	0	0	1	1	1.8	1427	9	77
Austin	SOU	FWD	60	89.4	302.8	689	107.7	15	7	1	0	0	0	0	2	0	2.7	1024	9	72
Ayite	FUL	MID	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0	0	0
Azpilicuet	CHE	DEF	65	535.7	967.4	168	167.5	38	2	6	1	0	0	0	1	0	15.8	3330	25	175
Baruna	HUD	MID	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0
Bailly	MUN	DEF	55	28.4	222	79	33	6	1	0	1	0	0	0	2	0	5.3	1060	3	55
Baines	EVE	DEF	55	373.5	407.4	102	88.2	34	2	3	0	0	0	0	1	0	2.5	1909	10	82
Bakayoko	CHE	MID	50	279.5	415.4	434	113.2	26	2	2	0	0	0	0	3	1	0.3	2121	5	77
Balbuena	WHU	DEF	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0
Baldock	BHA	FWD	45	0.8	0.4	6	0.7	1	0	0	0	0	0	0	0	0	2	30	0	2
Balogun	BHA	DEF	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0
Bamba	CAR	DEF	45	0	0	0	0	0	0	0	0	0	0	0	0	0	1.1	0	0	0
Bannigim	EVE	MID	45	3.8	9.4	0	1.3	2	0	0	0	0	0	0	1	0	0.1	192	0	7
Bardsley	BUR	DEF	45	117.5	229.4	35	38.3	13	0	0	0	0	0	0	5	0	0.2	1125	0	33
Barkley	CHE	MID	60	34.7	21	45	10.1	4	0	0	0	0	0	0	0	0	0.4	129	0	3
Barnes	BUR	FWD	60	274.8	407.2	797	145.8	23	9	0	0	0	0	0	10	0	1.5	2150	9	92
Batth	WOL	DEF	40	0	0	0	0	0	0	0	0	0	0	0	0	0	1.6	0	0	0
Bednarek	SOU	DEF	40	1.9	110	19	13.1	5	1	0	0	0	0	0	1	0	6.1	427	0	22

Figure 4.2 : Screenshot of the dataset

4.2 Implementation

Here we are using machine learning algorithms to do the prediction of the player who has to be removed/transferred out so as to get in the best player. Regression technique was used in this process to predict the player because the target variable was in numeric so we had to choose regression as the technique to proceed with the analysis, I also trained my dataset in Rapidminer the auto model results showed that the RMSE of these technique were less, so we chose to use these algorithms.

Before starting the whole process, we are taking a survey to understand the mindset of the majority of the players who play FPL to get to know on what basis they select the player who

has to be transferred out. From this survey we will get to know what the players mindset are, are they safe players or aggressive like, taking a hit (-4) and get in the best player.

The constrains that are to be considered while choosing a player and while transferring out the player are:

- Which team does he play for?
- How good is he as a player on paper?
- What are the next fixtures of that particular team?
- How much budget is left in the bank?
- Players form in the previous matches.
- Is the team playing away or home?

The different parameters used in the FPL data are:

- Price: The price of the player
- Creativity: The tells us how creative he is in the game
- Influence: This tells us how much impact does the player do in a winning cause
- Threat: This tells us how much threat he imposes on the opponent
- Goals Conceded: The number of goals let in by the subject team
- Goals Scored: The number of goals scored by player of subject team
- Assist: How many times player has assisted in scoring goals
- TSB: How many FPL players have chosen a particular player in their team
- ICT: This is the combined score of Influence, Creativity and Threat
- Minutes: This tells how many minutes has the played in the premier league till date from the current season started
- Bonus: These are points given to the player depending upon the bonus point system score
- Points: This is the target variable in the following research and this tells us how many points a player has scored till date in the premier league since the start of the season

The machine learning algorithms that have been used are:

- Linear Regression
- Artificial Neural Network
- Support Vector Machine

We have used “R” for the coding of these algorithm and also visualizing the outputs of these techniques.

4.2.1 Linear Regression

Simple Regression is the most usually utilized calculation utilized for prescient investigation. The primary concern of relapse is to look at two things

(1) independent variable's effect on the objective variable

(2) identifying the critical free factor for the objective variable and the manner in which it may affect

The condition for Simple straight relapse with one target and one free factor is appeared by recipe y = c + b*x, where y signifies the catch of target variable, c is the steady, b is the coefficient of relapse and x is capture of autonomous variable. In various straight relapse there would be one target variable and more than one autonomous variable.

4.2.2 Support Vector Machine

The support vector machine searches for the closest points, which it calls the "support vectors" (the name "support vector machine" is due to the fact that points are like vectors and that the best line "depends on" or is "supported by" the closest points). Once it has found the closest points, the SVM draws a line connecting them (Mustafa et al., 2017). It draws this connecting line by doing vector subtraction (point A - point B). The support vector machine then declares the best separating line to be the line that bisects and is perpendicular to the connecting line.

SVMs (Supper Vector Machines) (Vapnik, 2013) exhibit unrivalled execution increases and strength in numerous applications over conventional methods1. One striking property of SVMs is the capacity to deliver the remarkable worldwide least of the mistake work (Burges, 1998). Another striking property is that its capacity to learn is free of the dimensionality of the element space (Joachims, 1998) in light of the fact that SVMs measure the multifaceted nature of speculations dependent on the edge with which they separate the information, not the quantity of highlights. These properties make the SVM show a perfect contender for tending to the prerequisites on high exactness and high dimensionality. Be that as it may, the SVM show does not address the interpretability prerequisite: it includes a large number of highlights in a solitary portion work, making it difficult to see a basic connection between the expectation and highlights that trigger it. A standard based model, for example, ID3 and C4.5, then again,

presents the rationale of forecast in the easy to understand whether at that point rule organize, however is second rate in execution for high dimensional issues.

To address all the above prerequisites, we incorporate the SVM display with the standard based model. The thought is to segment the SVM grouping with the goal that few principles catch "straightforward, real structures" and the SVM demonstrate catches "perplexing, inconspicuous structures". The coordinated model, called rule-SVM (rSVM), places the principles at the best and the SVM at the base: to arrange a protein, the SVM classifier is connected just if there is no coordinating standard. Accordingly, the principles take grouping from the SVM.

The rSVM is to uncover the method of reasoning of the SVM order through justifiable on the off chance that rules while safeguarding the exactness (i.e., review and accuracy) of the SVM classifier. Vitally, our emphasis isn't on enhancing the exactness of SVM; we just safeguard it. Or maybe, our emphasis is on showing the justification of SVM through extricating in the event that rules and safeguarding the precision of SVM. To save the exactness, it is vital not to supplant the SVM classifier completely with tenets, yet to supplant just the piece of order that can be communicated by basic guidelines precisely; the SVM still covers the rest of the grouping that is unreasonably unpredictable for standards. In this sense, our methodology joins SVM with tenets. The above centre separates our work from those in the writing. In (Yu et al., 2002) utilizes a sequential association of a choice rundown and a straight separator to enhance precision, where a choice rundown is a positioned rundown of highlights (not rules). In our terms, such highlights are length-1 rules. In our encounters, such highlights are too broad to be in any way precise, particularly in a high dimensional space. We saw that a mix of a few lower positioned highlights, i.e., a length-k rule for k>1 frequently has a higher exactness than a solitary very positioned highlight, as shown by the way that most quality standards separated in our investigations have length longer than 1. In (She et al., 2003) utilizes a 2-level standard based classifier to enhance review, to the detriment of decreased accuracy. Notwithstanding, the execution is substandard compared to SVM. Neural system has the comparative "discovery" issue as SVM, and there has been take a shot at separating in the event that rules from neural system. Those works endeavoured to supplant a neural system classifier totally with principles (Andrews et al., 1995), which needs to pay the expense of unmatched execution.

4.2.3 Artificial Neural Networks

An artificial neural network is an artificial portrayal of the human cerebrum that attempts to mimic its learning procedure. It makes utilization of accumulations of scientific models that copies the watched structure and elements of the cerebrum so as to mimic its learning procedure). The human cerebrum comprises of billions of neuron cells, each having restricted capacities, yet when associated together, these neurons shapes the most keen framework known.(Saikia et al., 2012) Similarly, artificial neural networks are produced using hundreds or thousands of recreated neurons called Processing Elements (PEs) combined indistinguishable route from the neurons in the human cerebrum. Some have demonstrated that a neural network can rough any given utilitarian shape to any ideal precision level, whenever worked with adequate number of concealed layers. This is because of complex associations that exist between the neurons.

Neural network gains as a matter of fact procured from the past, by searching for examples in the information being displayed or some type of connection between the data sources and the consequence of each record.(Saikia et al., 2012) Artificial neural network has the favourable position over conventional straight models in that it can speak to both direct and non-straight connections and to gain these connections straightforwardly from the information being displayed .On the other hand, customary straight models are just constrained to the demonstrating of direct relationship among information. Neural Networks are being connected to an inexorably quantities of territories. They have been utilized in the errand of planning aircraft flights, characterization of radar mess, programmed target acknowledgment and therapeutic conclusion with significant achievement. They have likewise been utilized in some gauging assignments, for example, climate determining), control stack estimating, bond rating and a large group of different uses in barrier, wellbeing and business. They are known to have delivered the best outcome to date in anticipating optional protein structure and some transient time arrangement forecast assignments. One of the principle issues of ANN is that it requires a significant number of information to gain from (Saikia et al., 2012).

4.2.4 Prediction

The control of the learning parameters is an unsolved issue in ANN explore and in addition in streamlining hypothesis. The objective is to achieve the ideal execution in little preparing time. The learning parameters of the picked system topology that fit into the examination consider were concentrated to decide the best parameters' setting. The customary methodology, which

was utilized in the examination, is to choose the learning rate and a force term. Force learning is an improvement over the straight inclination plummet seek, by forcing a "memory factor" on the adjustment. This had the advantage of fast adaptation, at the same time reducing the probability of getting hooked at local minima. Thus, we have the learning equation:

$\Delta W_{ij}(K) = \gamma \Delta W_{ij}(K-1) - \mu \partial E(k)$

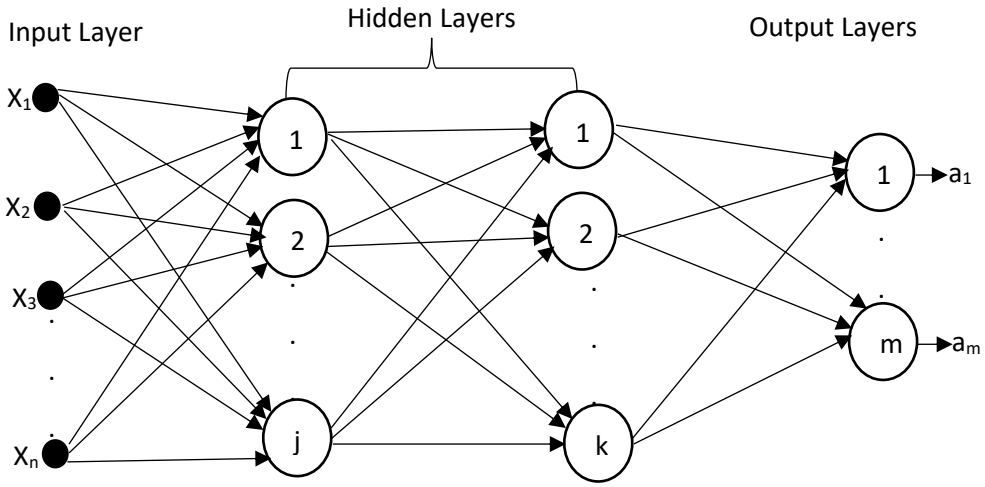


Figure 4.3 Architecture of ANN

Where:

μ = The learning rate

γ = A constant (normally set between 0.5 and 0.9)

Different learning rates ranging between 1 and 0.001 with the momentum term of 0.7 were used for the hidden layers and the output layer.

$F(net) = \tanh(\alpha\ net)$

So as to get a vibe of the exhibitions of the models and how troublesome the issue is, the ANN models were at first prepared without the utilization of the cross-check ceasing basis, until a very much characterized number of preparing ages (set to 1000 in this investigation) has been come to. The adapting at that point continue by changing to cross-check and saw until the Mean Square Error (MSE) for the confirmation set records no enhancement after 100 ages or the quantity of the preset preparing ages has been come to. An expansion in the MSE recommends that the system has started to over trained which can prompts poor speculation.

Execution measures: The general exhibitions of the models were assessed by some determining exactness measures. Five execution measures were utilized in the investigation.

Mean Squared Error (MSE): The mean squared mistake is multiple times the normal cost characterized by the recipe:

$$MSE = \frac{\sum_{j=0}^{P}\sum_{i=0}^{N}(d_{ij} - y_{ij})^2}{NP}$$

Where:

P = Number of output processing elements

N = Number of exemplars in the data set

Yij = Network output for exemplar i at processing element j

dij = Desired output for exemplar i at processing element j

Normalized Mean Squared Error (NMSE): This is defined by the formula:

$$NMSE = \frac{P\,N\,MSE}{\sum_{j=0}^{P}\frac{N\sum_{i-0}^{N}d_{ij}^2 - (\sum_{i=0}^{N}d_{ij})^2}{N}}$$

Where:

P = Number of output processing elements

N = Number of exemplars in the data set, MSE. Mean Squared Error

dij = desired output for exemplar i at processing element j.

The correlation coefficient r: The execution of the system yield to the longing yield can be estimated with Mean Square Error (MSE) esteem, yet it doesn't really reveal to us the bearing of development of the two arrangement of information, subsequently the requirement for the connection coefficient r. The relationship coefficient between a system yield x and an ideal yield d is given by:

$$r = \frac{\frac{\sum_i (x_i - \bar{x})(d_i - \bar{d})}{N}}{\sqrt{\frac{\sum_i (d_i - \bar{d})^2}{N}} \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N}}}$$

It is the proportion of the covariance between the info and wanted information over the result of their standard deviations. The relationship coefficient extends between [-1, 1]. The objective is to have the estimation of r near 1 as could be expected under the circumstances.

Akaike's Information Criterion (AIC): It gauges the trade-off between preparing execution and the span of the system. The point is to decrease this term to get a system having the best speculation. The term is given beneath:

AIC (k) = N ln(MSE) + 2k

Where:

K = Number of network weights

N = Number of exemplars in the training set

MSE = Mean Squared Error

Minimum Description Length (MDL): Rissanen's Minimum Description Length (MDL) paradigm is like the AIC in that attempts to consolidate the model's mistake with the quantity of level of opportunity to decide the dimension of speculation. The objective is to limit this term. It is additionally used to set up the speculation capacity of a model by consolidating the model's blunder with the quantity of degrees of opportunity. The point is to have minimal incentive for the term:

MDL(k) = Nln(MSE)+0.5kln(N)

Where: K = Number of network weights

N = Number of exemplars in the training set

MSE = Mean Squared Error

5. Analysis/Discussion

Firstly, when were told about the thesis, first idea came to mind it that to do research process in an unique manner and do the process in the field where there is no much research done. From the ideology I had made me choose this topic as this was one of the least research topics and it was one of the main problems in the game which had no answers for it except the suggestions given by the official predictors of the game. I took up this model as I was very familiar about the thing that I was doing and it would be easy for me to do it if I am familiar and I was also one of the users of FPL who is facing these problems in the game. By doing this it will help a lot of users where they are facing a problem where they did not know which player to remove from their team and get stuck in a dilemma. I thought to build a model where the data of all the 528 players are fed and by using optimization techniques would build a team that will be the best one for the first gameweek and then for subsequent game weeks choose the best player so that he could get maximum points every week. This proposal had to be done using Python using the Pyomo available for the optimization techniques, even the algorithm was selected and the process was in the prime time when the problems started to occur , the main problem I faced it that the dataset chosen was not giving proper results for the prediction which had to be done, later after analysis of the dataset I had to do slight modification of the dataset and continue with the process , but due to lack of time as discussed with my supervisor it would have taken more time than the given time space so had to drop out on that and kept it for future so that I can complete the research I wanted to do when there no time constraint. Due to these constraints I had to change the dataset and wanted to do research on the topic which can be finished within the time stamp given. Since I did not want to get away from this topic I decided to similar thing where instead of predicting the best player to bring and as we know which player to bring in is definite and the problem was which player to remove from the squad I decided to build a model where I could do predict which player to remove using machine learning algorithms.

5.1 Analysis of Result

According to the dataset we have chosen to use the following machine learning algorithms , Linear Regression , SVM and ANN. Since we are using these algorithms as our target variable is a numeric value we have to look for the RMSE value to find the accuracy. Root Mean Square Error is the value that tells us how good is the model is built and also it is used to assess the fitness of the model. The value of RMSE ranges from 0 to 1 and 0 being the least value and

telling us the model is not good and 1 tells us that model is the best and any value nearing the value 1 is called a good model. In the model which was built we tested it with 3 algorithms and checked out the RMSE value for all the three and then had to conclude about the best model which can be used so that the research process can be done smoothly.

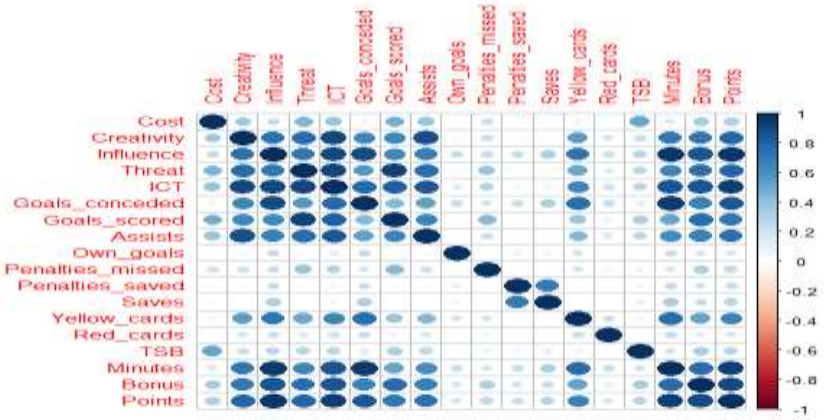


Figure 5.1 : Correlation Matrix

The above correlation matrix tells us that how they attributes are related to each other in the dataset. In the matrix we can see that there are colours showing the relation. Blue and red are the two colours showing the how strongly related , both blue n red mean that they are strongly related but blues tell that are positively related and it is significant and the red ones tell that is negatively related and it not that significant.

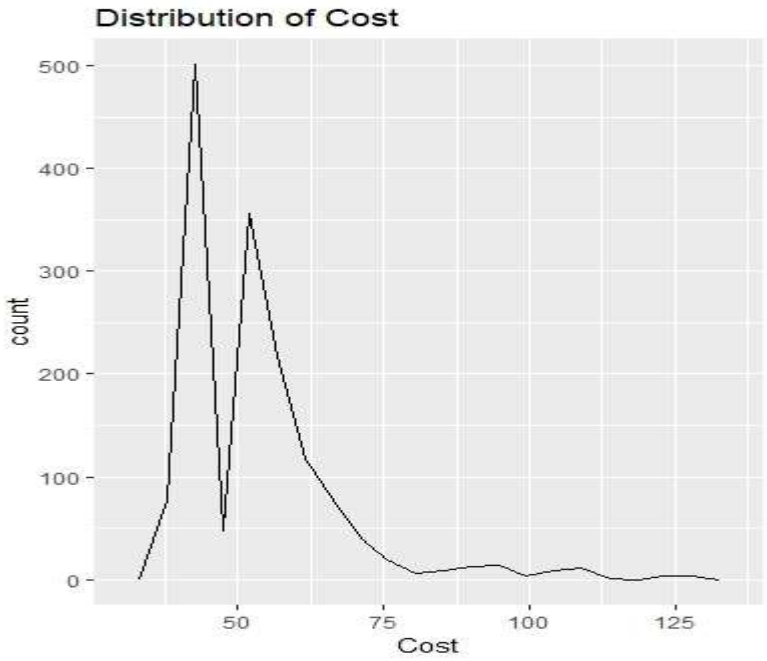


Figure 5.2 : Graph showing the cost distribution

The above graph shows the cost distribution of the players telling that the range of players from the cost 4.0-6.0 is the most.

From the outputs of the three machine learning algorithms we have got the RMSE value of 0.9843 , 0.9805 and 0.9457 for ANN , Liner Regression and SVM respectively.

From these results we can tell that ANN has the highest accuracy and is the best model to use for the prediction in this dataset.

The below graph shows the graph of the original test set and the prepared model using Machine Learning Algorithms.

The black lined graph shows the original values of the and the blue line shows the predicted model using ANN

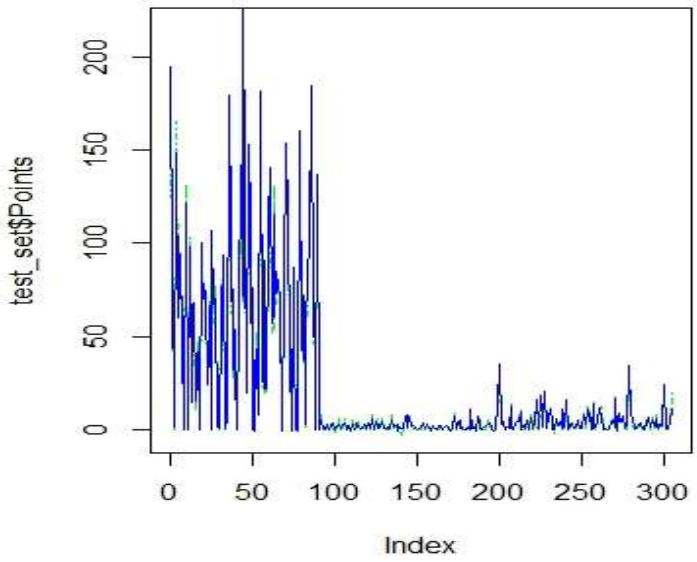


Figure 5.3 : Showing the best Model using ANN

6. Conclusion

In this work, we are predicting which player has to be removed from the team so as to get in a player who can give you points such that you will have the most points that game week using machine learning algorithms. By using the parameters such as goals scored, assists, TSB, yellow cards etc which affects the points of the players we predict which player has to be removed from the team. From all the tests conducted and the simulation results we have compared three algorithms namely ANN, random forest, Linear regression. From results as seen we can conclude that the best prediction algorithm is ANN, which gives the result best compared to state-of-art algorithms. The simulation results proved the accuracy of the system and also show the run time as low for the ANN algorithm.

6.1 Future Work

This research also gives an area of scope as a part of future work in predicting from which position to remove the player rather than just removing a random player who might not be the person to remove but had to be removed according to the algorithm. In simple way ,there are more factors to be taken into account in the future work and data needs to be more precise

which is a tough task to get , this future work will be the updated and the more accurate version of the current algorithm in place and since this one of few model for this game available , the future work also depends upon how the current models work and the flaws of these will be used to build a more accurate model for this game.

References

Al-Shboul, R., Syed, T., Memon, J. and Khan, F., 2017. Automated Player Selection for Sports Team using Competitive Neural Networks. *International Journal of Advanced Computer Science and Applications*, *8*(8), pp.457-460.

Bunker, R.P. and Thabtah, F., 2017. A machine learning framework for sport result prediction. *Applied Computing and Informatics*.

Cao, C., 2012. Sports data mining technology used in basketball outcome prediction. *Masters Dissertation. Dublin Institute of Technology*

Constantinou, A.C., Fenton, N.E. and Neil, M., 2012. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, *36*, pp.322-339.

Constantinou, A.C., Fenton, N.E. and Neil, M., 2013. Profiting from an inefficient Association Football gambling market: Prediction, Risk and Uncertainty using Bayesian networks. *Knowledge-Based Systems*, *50*, pp.60-86.

Drawer, S. and Fuller, C.W., 2002. Evaluating the level of injury in English professional football using a risk based assessment process. *British journal of sports medicine*, *36*(6), pp.446-451.

Gelade, G.A. and Dobson, P., 2007. Predicting the comparative strengths of national football teams. *Social Science Quarterly*, *88*(1), pp.244-258.

Goddard, J. and Asimakopoulos, I., 2004. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, *23*(1), pp.51-66.

Huang, K.Y. and Chang, W.L., 2010, July. A neural network method for prediction of 2006 world cup football game. In *Neural Networks (IJCNN), The 2010 International Joint Conference on* (pp. 1-8). IEEE.

Hucaljuk, J. and Rakipović, A., 2011, May. Predicting football scores using machine learning techniques. In *MIPRO, 2011 Proceedings of the 34th International Convention* (pp. 1623-1627). IEEE.

Joseph, A., Fenton, N.E. and Neil, M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), pp.544-553.

Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zając, A. and Stanula, A., 2014. Application of neural and regression models in sports results prediction. *Procedia-Social and Behavioral Sciences*, *117*, pp.482-487.

O'Connor, D., Larkin, P. and Mark Williams, A., 2016. Talent identification and selection in elite youth football: An Australian context. *European journal of sport science*, *16*(7), pp.837-844.

Razali, N., Mustapha, A., Yatim, F.A. and Ab Aziz, R., 2017, August. Predicting Player Position for Talent Identification in Association Football. In *IOP Conference Series: Materials Science and Engineering* (Vol. 226, No. 1, p. 012087). IOP Publishing.

Sanjay, M., Srinivasan, S., and Kulkarni, K., 2016. Data Mining Technique for Best 11. *International Journal of Conceptions on Information Technology and Computing, 4(2), pp.10-12.*

Srimani, P.K. and Koti, M.S., 2013. Medical diagnosis using ensemble classifiers-a novel machine-learning approach. *Journal of Advanced Computing*, *1*, pp.9-27.

Szczepanski, L., 2015. Assessing the skill of football players using statistical methods, Doctoral dissertation, *University of Salford*

Tiedemann, T., Francksen, T. and Latacz-Lohmann, U., 2011. Assessing the performance of German Bundesliga football players: a non-parametric metafrontier approach. *Central European Journal of Operations Research*, *19*(4), pp.571-587.

Uzochukwu, C. N, and Enyindah, P, 2015, A Machine Learning Application for Football Players' Selection. *International Journal of Engineering Research & Technology (IJERT),* 4(10), pp.459-465.