DATA ANALYSIS OF AN EUROPEAN SOCCER DATABASE

[IFN701 Project 1- Data Analysis and Research Project]

By: Harmandeep Kaur Bhullar(n9784098)

Academic Supervisor: Dr. Guido Zuccon

Project Coordinator: Dr. Charles Wang

ABSTRACT

Soccer is the most popular and prevalent sport worldwide as this game is the most widely covered and viewed event in the world with the total of 28.8 billion viewers. So, soccer is greatest source of money making in entertainment industry. When we talk about betting, the interesting factor is that almost 70 percent is contributed from soccer betting. Therefore, many data- analysis companies are making lots of money by predicting the results of matches and performance of players by analysing the data which is collected from matches. A lot of data acquired regarding the matches. Available data can be used to perform different type of analysis tasks.

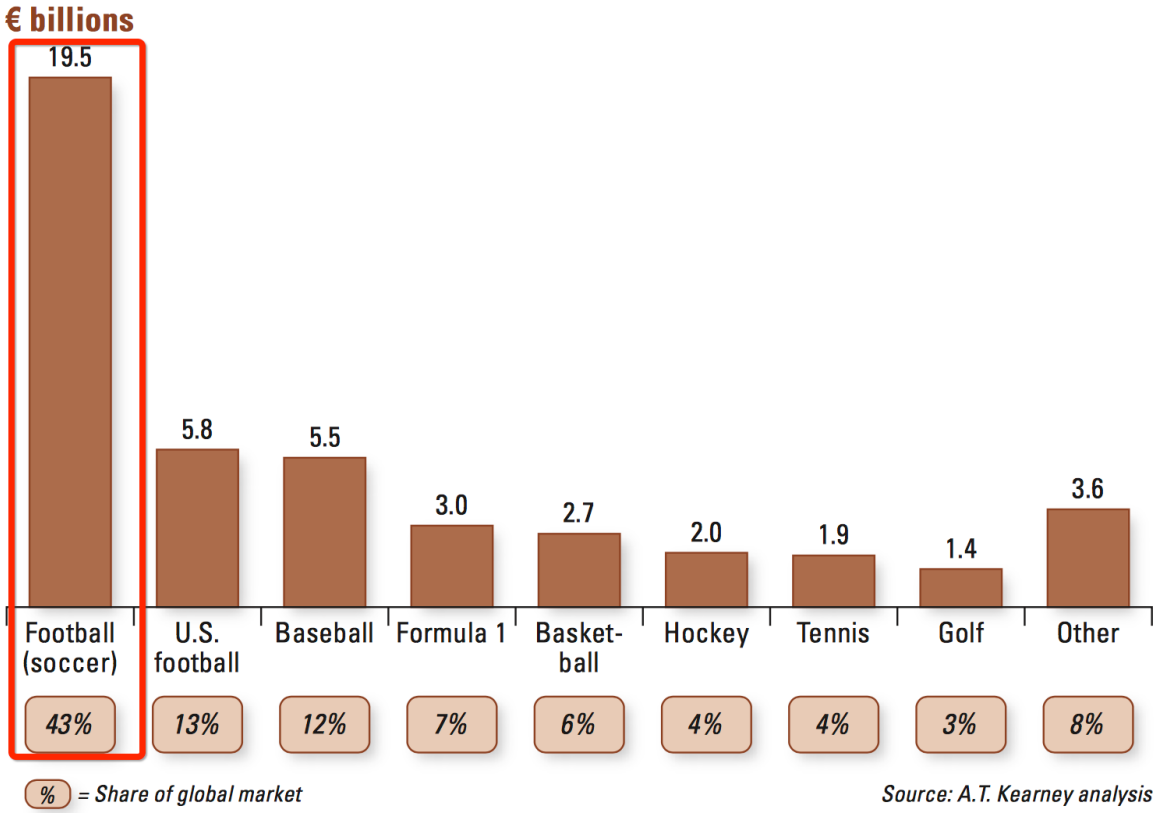
It is certain that the results of matches are very difficult to predict accurately. Sometimes, things went in unexpected way as we expect or predict like it is not obvious that best team always wins the match and weak team always loss the match. But it does not mean that there is no possible way to predict the results of match accurately. To do so there is need of deep analysis of factors that can affect the outcome of match results and analyse the positive and negative effects that are directly affecting the outcome of results of matches. Therefore, there are some strong factors that should be taken into account while making the predictions about the results of matches. This project main objective is to find out those factors that affects results of matches and implement the prediction model by using those factors to predict the results of matches.

1.INTRODUCTION

This report is about the data analysis and research project to implement predictive modelling of European soccer matches database. The main objective of this project is to analyze data to find out interesting facts about this game results, teams and players by understanding the characteristics of teams and players. These analyses are done by using the dataset which is acquired from the Kaggle and coding is in R studio software.

1.1 Context of the Project

Soccer is the most popular and prevalent sport worldwide. According to a recent A.T. The global sports events market (includes all ticketing, marketing and media revenues) for all sports was worth €45 billion in 2009. Soccer remains king as always with global worth of €20 billion every year. Particularly in Europe, it is €16 billion which is almost half of the market. Moreover, this game is the most widely covered and viewed event in the world with the total of 28.8 billion viewers. So, soccer is greatest source of money making in entertainment industry. (Collignon, H.& Sultan, N.)



On the flip side, According to Sportradar director Darren Small, the worldwide betting industry for sports is worth about $700 billion to $1 trillion every year for both legal and illegal betting markets. The interesting factor is that almost 70 percent is contributed from soccer betting. Therefore, many data- analysis companies are making lots of money by predicting the results of matches and performance of players by analysing the data which is collected from matches. It can be expected that this market will continue to rise to trillions in next years which will make the betting industry major part of data companies. (Mr Neutral (January 21, 2014))

1.2 Research gap and related work

A lot of data acquired regarding the matches. Available data can be used to perform different type of analysis tasks. However, to make the sense of using the data is to find out the facts to answer the question **why (i.e.** understanding the behavior and characteristics of teams and players by analyzing the factors behind actions features w.r.t results of matches) instead of **what (**who will win or loss)**.** Because by understanding the factors of performance of teams in manner of technical, physical and tactical strengths and weakness can help to reach the correct prediction about the results of matches and performance of teams and players. Moreover, by tracking the past performance over the time period of players can be used to analyze the trends of performance of players. This information can be stored and sell on demand according to time period, player, match, speed etc. Also, these analyses are displayed graphically to track the performance of team or player over the certain time period.

Moreover, the knowledge is hidden within the soccer clubs/management level, it's not disclose to the general public. The project aimed to provide insights to team's characteristics / strength/weakness to general public. Feature analysis based on win, loss and draw class for home and away teams not done yet.

How this project addresses this problem

As above discussed s always very hard to predict the soccer match results and it is not obvious like best team always win or team with top/best players is a best team. So, this project will use the dataset from kaggle platform which is about about more than 25,000 matches, 10,000+ players from 11 European Countries with their lead championship from seasons

2008 to 2016. I will use this dataset to answer the following research questions:

- Understand the characteristics of teams to find the strengths and weaknesses of teams.
- Track the performance of players over time and understand the trends of performance.
- Analyze the co-relation between the features of teams with respect to win and loss class so that how they affect and contribute towards win class results of matches.
- Analyze and find the factors that affects match results.
- Predict the results of matches.

1.3 Aim and Objective of the project

After reviewing the literature review, I came up with following research questions which I would like to answer by analysing the dataset:

1. Analyse the features of teams and find out which features affects the match results and how they are contributing towards win class?
2. Find the co- relation between the all the features of teams w.r.t. **Win** and **Loss** class like how they their positive relation or negative relation between various features of teams affecting the match results?
3. Weather the home and away feature affects the match results and features of teams for example is there more defence pressure on away teams and result of match?

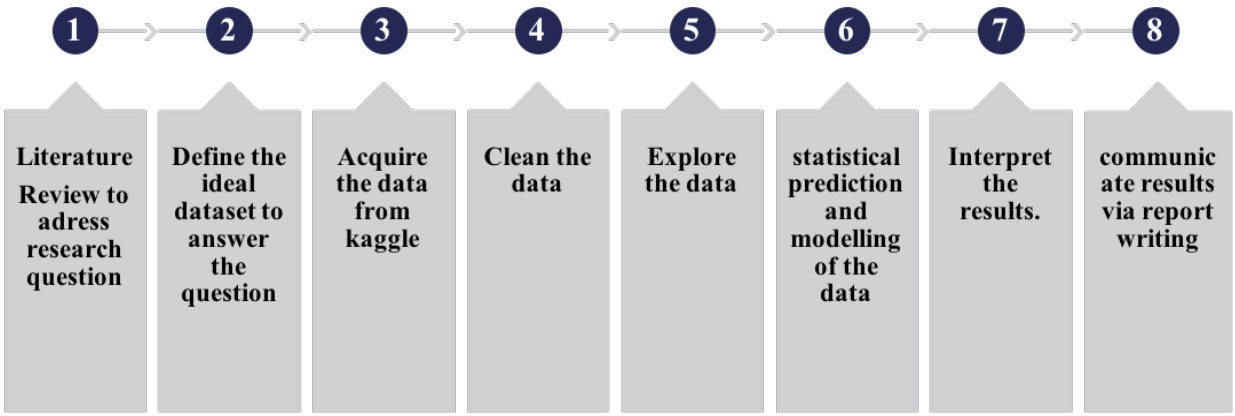
   Therefore, the **main objective** of project is:

   a. Do exploratory analysis to explore and analyse the features of teams and answer the above questions.
   b. Statistically predict the match results and build the model to predict the match results by using answer of above questions
   c. Communicate the results via presentation and written

report.

1.4 Brief overview of methods used in the project

I have implemented this project by using following workflow which includes 8 steps (Byrne, C., 2017)



Step 1: First of all, I did literature review to address the questions of interest to answer by doing analysis of data.

Step 2: Here, I selected the appropriate dataset which is used for data analysis and visualisations in order to answer the questions.

Step 3: Next, I acquired the dataset from the public platform "Kaggle" which is popular for data analysis competitions. This dataset is in form of CSV files. There are total 8 tables about the European soccer matches.

Step 4: Now, as data cleaning is a very crucial step for data analysis project as it is very important to derive accurate and reliable results. Here, I cleaned the dataset data by removing null or NA values or unwanted rows and columns. I selected only 6 tables of interest from the dataset.

Step 5: Here, data exploration is done by visualising the features of teams and players. After visualising the features, I find co- relation between the all features of teams and did feature analysis for home and away teams w.r.t. win and loss class to analyse how these factors affects the match results.

Step 6: After analysis phase, Statistical prediction and modelling of the data is done in R studio to produce the outcome to understand

the characteristics of players and uncover the factors and trends of match results. Here, I implemented three models are decision tree, SVM (support vector machine) and random forest to predict the results of matches.

Step 7 and 8: Finally, the outputs and observations from Analysis with R markdown file is interpreted as written report to explain and communicate the results and answer the targeted questions.

1.5 Outcome of this project

The tangible results of this data analysis project maare as listed below:

- **Data analysis report:** This data analysis report gives the detailed theory about the outcomes derived from the projects which includes analysis of features of teams and visualisations which indicates which factors affects the match results and prediction models which can predict match results and done in R studio.
- **R markdown file:** Analysis and visualisations are done by programming in R Studio software. The R markdown file includes all the source coding with all visualisations and analysis.

The outcome of this project can be used by football clubs, scouting, coaches and betting agencies to determine the team's performance, to improve training techniques, to find trends of performance of players over the certain time period and to predict the results of matches.

2. LITERATURE REVIEW OF PREVIOUS WORK

There is no doubt, many football games are played during every year season and there are so many people who bet on games and teams that who is going to win a game to make money. Soccer betting has grown very fast over the last decades. Moreover, with the advent of Internet betting exchanges increases very quickly. Therefore, the

soccer sports-betting industry remains healthy as well as it will keep on expanding in next coming years. (Mr. Neutral, 2014)
It is certain that the results of matches are very difficult to predict accurately. Sometimes, things went in unexpected way as we expect or predict like it is not obvious that best team always wins the match and weak team always loss the match. But it does not mean that there is no possible way to predict the results of match accurately. To do so there is need of deep analysis of factors that can affect the outcome of match results and analyse the positive and negative effects that are directly affecting the outcome of results of matches. Therefore, there are some strong factors that should be taken into account while making the predictions about the results of matches. (GamblingSites.com (n.d.)).
This motivates to first find out the factors that affects the match results by analysing the features of teams. This is my first research question which can be answered by analysing and visualising the dataset given on “kaggle”.
The home advantage can be considered as good factor that can contribute towards results of match. According to some authors the home advantage may also be considered as fact that indicated that there are more chances of home team to be a winner in their home ground (Irving and Goldstein, 1990), and this may be due to travel fatigue. Moreover, the author argues that the social crowd would be positive advantage for home country and can be reason for more pressure on defence team (away team) (Smith et al., 2003).
Therefore, above argument motivates me to analyse whether the home and away team is good feature to consider or not. How it affects the match results and also the features of defence team like pressure, aggression and team width of opposite team (away team). So, I wanted to find out the co-relation between features and feature analysis for home and away teams and results of matches.

According to the Oxford dictionary, tactics is defined as “an action or strategy carefully planned to achieve a specific end”. In context of soccer game competitions, the main objective is to win the game. Therefore, there is need of selecting an appropriate tactic is very crucial for game preparation before match. (Carling et al. 2005; Sampaio & Macas 2012; Yiannakos & Armatas 2006). Hence, the coach of team should consider the factors like status of the team, the status of the opposite teams and also the external factors like playing at home or away team to ensure the successful implementation of tactical levels (Gréhaigne & Godbout 1995; Mackenzie & Cushion 2013).

Therefore, above discussion gives me an idea that there is need to deep analysis of all the features and factors of teams that affects the results of matches. Because these analyses help the coach how to train the teams before matches. The coach would able to get an idea that on what area weakness and strengths of teams should be taken into account while giving training prior to match.
Moreover, if I talk about “kaggle” work done on this dataset, there was feature analysis w.r.t win and loss class for away and home is not done yet. Therefore, by doing feature analysis of all teams for home and away teams w.r.t. will be beneficial in finding the factors which affects the results of match results.
Also, co-relation between all the features for win and loss class has not been done. Co-relation of all features of teams w.r.t win and loss can be very helpful to analyse that how and what kind of positive or negative relation between various features of teams contribute towards results of matches. Hence, by doing so can give and idea about how these features combination should be to achieve win.
The comparison between the prediction models can give an idea about which predict model can be best suited to predict the results of matches.
By reviewing the literature and work done in this context helped me to come up with following research questions:

1. Understand the characteristics of teams to find the strengths and weaknesses of teams.
2. Track the performance of players over time and understand the trends of performance.
3. Analyze the co-relation between the features of teams with respect to win and loss class so that how they affect and contribute towards win class results of matches.
4. Analyze and find the factors that affects match results.
5. Predict the results of matches.
6. Comparison between the prediction models

3. PROJECT METHODOLOGY

This project was managed by DSDM approach (Dynamic Systems Development Method) as this approach supports the iterative and incremental development by keeping the time, quality and cost fixed but features are variable for optimal and flexible solution which

helped me to control the risk. Moreover, an active communication and involvement of stakeholders throughout the development period to give feedback and review the deliverable. I have selected this approach because of following reasons:

- **Delivery on time:** This project was needed to complete on time to meet the academic requirements. So, time boxing helped me to deliver on time.
- **Never compromise quality**: With the help of this approach, I able to achieve the quality by MOSCOWS rule and regular meetings with supervisors and project coordinates to get regular feedbacks through reviewing the deliverables which led me to deliver expected results and quality.
- B**uild incrementally from firm foundations:** This approach supported the concept of developing the firm foundation for this project before committing to any significant development. It advocated understanding the scope problem first which is required to be solved with proposal of solution with not deep details so that the project becomes complex by overdetailed plan.

3.1 MoSCoW Prioritization for Scope:

Prioritization	deliverable
Must have (60%)	• data analysis and visualization to understand the characteristics of teams • feature analysis of teams for home and away teams w.r.t. results of matches • find co-relation between all features of teams w.r.t win and loss class • report writing
Should have (20%)	• Uncover the factors which affects the results of matches and performance of teams and players
Could have (20%)	• track the performance of players over time period to figure out trends • analyze the features of players

Would not have	• build the predictive model which predict the results of matches with high accuracy rate (99%).

3.2 Tasks breakdown

The following table gives the detailed breakdown of tasks according to the methodology steps mentioned in “section 1.4” which I have completed during project implementation.

Phase	Tasks	Outcome	Duration
Phase 1: initialization (step1, 2 & 3)	• finalized Project management team • reviewed literature to understand the problem • after literature reviewed, I finalized research question • Then, I defined the project scope and objective • Joined slack group for continuous communication • Made a repository in GitHub • Project Proposal • Download R • Acquired dataset	• Project agreement • Repository in GitHub • Project proposal presentation • Project proposal plan report	5 weeks (week 1 – week 5)
Phase 2: Analysis and visualization of data (step 4, 5 & 6)	• Cleaned the data • Explored the data • Analysis and visualization • Statistical modelling and prediction model	R markdown file Analysis report	6 weeks (week 6-week 11)
Phase 3: Consultation	Reviewed and feedback from supervisor and co-ordinate via project presentation	Feedback and suggestions on project presentation	1 week (week 12)
Phase 4: Final	Final Data analysis report	Final report	Throughout from week 6-week 13) and submit on week 14

3.3 Time frame

This project took 13 weeks to complete and delivered in three increments. Each increment took 4 weeks to be done and each increment divided into two timeboxes. All details are given in below table:

Total timeframe taken by the project	**13 weeks (almost 3 months)**
Number of timeboxes	6
Time period of each timebox	2 weeks
Number of increments	3
Time period of each increment	1 month (4 weeks)

3.4 Detailed Weekly Plan

Increment 1: Increment 1 took 4 weeks (academic study period week1-week4). Here, in this period I finalised the project, project team(supervisor), presented project proposal presentation and delivered detailed project proposal report to project supervisor and project co-ordinator.

Increment 2: It also took 4 weeks (academic study period week5-week8) to be delivered. In this period, I acquired the data, started working on that and done with visualisations and analysis of team and players features and delivered the analysis report to supervisor (Dr. Guido) to get feedback.

Increment 3: It delivered in almost five weeks (academic study period week9-week13). Prediction models for prediction of match results was implemented during this last increment and results was communicated via presentation in week 12 and final report will be delivered in week 14 to supervisor and co-ordinator.

Duration	Period	Planned tasks	Outcome
Increment 1			
Timebox 1			
Week 1		• meeting with supervisor • created repository on gitHub	• Installed Rstudio • Repository on GitHub
Week 2		• Project selected	• Project allocated
Timebox 2			
Week 3		• Sign off the project agreement • Prepared project presentation	• Agreement Submission • Acquired the Dataset • Joined Slack group • Literature reviewed • Project proposal pitch
Week 4		**Started working on project proposal**	**Note: got extension for project proposal as special reason late project agreement was done**
Increment 2			
Timebox 3			
Week 5		• Project proposal report • Acquired dataset • Cleaned the data	• Submitted project proposal report • Downloaded dataset from "Kaggle"
Week 6		• Data Analysis	• Data cleaned by coding in R
Timebox 4			
Week 7		• Analysed and Explored the data	• Data explored by coding in R
Week 8		• Visualized the data	• Data visualised via R coding
Increment 3			
Timebox 5			
Week 9		• Prediction modelling • Registration for week 12 presentation • Worked on presentation	• Data modelling and statistical prediction by coding in R • uploaded slides on GitHub to seek feedback
Week 10		• Project presentation • R markdown and updated on GitHub to seek feedback from supervisor	• Improvements in R markdown coding according to received feedback
Timebox 6			

Week 11		• **Documentation of data analysis in report** • **Submitted the draft on GitHub for review and feedback**	• **Data Analysis report** • **Project plan presentation**
Week 12	**week 13**	• **Final report**	• **Final report submission**

3.5 Project Methodology

1. **Phase 1: Initialisation phase:** This phase includes the step 1(finalise question of interest),2(ideal dataset) and 3(download the data). This phase was delivered as increment 1. Here, first I finalised the project, questions of interest and acquired the dataset from “Kaggle” and load into R studio to perform analysis, visualisations and predictions.
   Here, I loading data in R:

```
#load soccer dataset
soccerData <- dbConnect(SQLite(), dbname="/Users/harmanbhullar/Documents/database.sqlite")
#list all tables in dataset
dbListTables(soccerData)
```

here is list of all 8 tables which are available in the soccer dataset:

```
> dbListTables(soccerData)
[1] "Country"            "League"             "Match"
[4] "Player"             "Player_Attributes"  "Team"
[7] "Team_Attributes"    "sqlite_sequence"
```

2. **Phase 2: Analysis and visualization of data:** This phase includes the step 4(clean data), 5(explore data) and 6(prediction and modelling).
   First of all cleaning the data: here I selected only required tables of interest which I used for further analysis.

```r
#loading all tables of interest
#player table contains data about 11060 players
player <- tbl_df(dbGetQuery(soccerData,"SELECT * FROM player"))
#player_Attributes consists of all the features of players
player_Attributes <- tbl_df(dbGetQuery(soccerData,"SELECT * FROM player_Attributes"))
#team table consists of information about the 299 teams
team <- tbl_df(dbGetQuery(soccerData,"SELECT * FROM Team"))
#team_Attributes consists of all the features of all teams
team_Attributes <- tbl_df(dbGetQuery(soccerData,"SELECT * FROM Team_Attributes"))
# Match table contains the information about the all matches from 2008 to 2016
Match <- tbl_df(dbGetQuery(soccerData,"SELECT * FROM Match"))
```

Since soccer database has 8 tables. Here, I am interested in only 5 (team, team_Attributes, player, player_Attribute and Match) tables. Therefore, I loaded only five tables of interest as I am interested to analyse the teams level and players level features and to predict the results I need match table which gives the results of matches w.r.t. teams api id, date of match and match api id.

Thereafter, I did analysis on team's level and player's level. But unfortunately, there data about players was missing so I could able to analyse the teams features and predict results based on the factors extract from analysis of team's features.

To analyse the features of teams I extract the results for all teams from match table so that I will connect the teams features with team results to analyse the factors affects the match results and came up with table team_A as with help of following coding in R:

First of all, I am going to extract the results for home from match table as it has values for number of goals for home and away country.

```
#subsetting the table by selecting required rows and storing into
datafram a
a <- Match[,6:11]
# stats tables to check what (contents of table) is available inside
table Match
str(a)
#adding new column results which gives the match results of home
team by comparing home_team_goals with away_team goals like if if
home team score more goals than away then it win and if less then
loss and if equal the draw
a$results <- a$home_team_goal
a$results[a$home_team_goal>a$away_team_goal] <- "Win"
a$results[a$home_team_goal<a$away_team_goal] <- "Loss"
a$results[a$home_team_goal==a$away_team_goal] <- "Draw"
#subsetting the table by selecting only required columns and storing
in team_A for home team
team_A<-a[c(1,2,3,7)]
library(dplyr)
#rename the column name home_team_api_id to team_api_id
team_A<-team_A %>% dplyr::rename(team_api_id = home_team_api_id)
str(team_A)
```

Here, I am going to extract the results for home from match table as it has values for number of goals for away and away country:

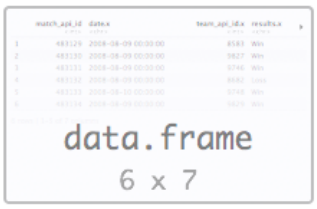
```r
#subsetting the table by selecting required rows and storing into
datafram b
b <- Match[,6:11]
#adding new column results which gives the match results of away
team by comparing away_team_goals with home_team goals like if away
team score more goals than home team then it win and if less then
loss and if equal the draw
b$results <- b$away_team_goal
b$results[b$home_team_goal<b$away_team_goal] <- "Win"
b$results[b$home_team_goal>b$away_team_goal] <- "Loss"
b$results[b$home_team_goal==b$away_team_goal] <- "Draw"
#subsetting the table by selecting only required columns and storing
in team_B for away team
team_B<-b[c(1,2,4,7)]
#rename the column name away_team_api_id to team_api_id
team_B<-team_B %>% dplyr::rename(team_api_id = away_team_api_id)
#team_B<-rename(team_B, "team_api_id"="away_team_api_id")
str(team_B)
```

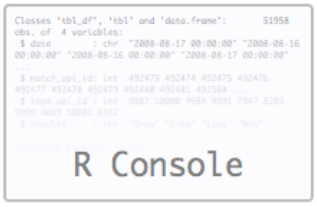
```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       25979 obs. of  4 variabl
 $ date        : chr  "2008-08-17 00:00:00" "2008-08-16 00:00:00" "2008-
6 00:00:00" "2008-08-17 00:00:00" ...
 $ match_api_id: int  492473 492474 492475 492476 492477 492478 492479 4
0 492481 492564 ...
 $ team_api_id : int  9987 10000 9984 9991 7947 8203 9999 4049 10001 834
.
 $ results     : chr  "Draw" "Draw" "Loss" "Win" ...
```

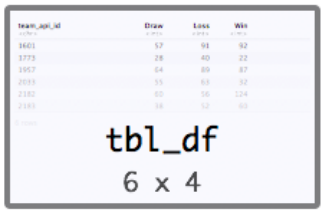
Here, team A table is derived from Match table which has information about the date of match result of match for home team with api id for home team and match api id. team B table is derived from Match table which has information about the date of match, result of match for away team with api id for away team and match api id. Then I joint the table team_A and team_B into single table so that I will able to do next analysis.

Next, I calculated the number of draws, wins and loss for each team for further analysis:

```
#getting the number of wins loss and draws of each team
z1 <- with(z,table(team_api_id, results))
z1<-as.data.frame.matrix(z1)
z1 <- add_rownames(z1, "team_api_id")
head(z1)
```







team_api_id <chr>	Draw <int>	Loss <int>	Win <int>
1601	57	91	92
1773	28	40	22
1957	64	89	87
2033	55	63	32
2182	60	56	124

Next, I combined the tables of results of teams and features of teams into single table so that I can do feature analysis and co-relation of all features w.r.t results of all teams.

```
> str(vis_features)
'data.frame':   50 obs. of  14 variables:
 $ team_api_id           : int  10000 10215 10251 10264 10267 10268 10268 2182 4087 6403 ...
 $ date                  : chr  "2014-09-19 00:00:00" "2010-02-22 00:00:00" "2012-02-22 00:00:00" "2013-
00:00:00" ...
 $ match_api_id          : int  1717876 686171 1024435 1498038 684955 1505856 1778070 1467907 1709747 17
...
 $ results               : Factor w/ 3 levels "Draw","Loss",..: 2 3 2 3 3 2 2 2 2 1 ...
 $ team_type             : Factor w/ 2 levels "away","home": 1 1 1 1 2 1 2 2 1 1 ...
 $ buildUpPlaySpeed      : int  54 30 55 66 30 55 55 67 61 38 ...
 $ buildUpPlayDribbling  : int  42 NA NA NA NA NA 47 NA 51 40 ...
 $ buildUpPlayPassing    : int  51 30 44 54 30 73 69 54 44 39 ...
 $ chanceCreationPassing : int  47 50 46 53 55 53 53 56 52 52 ...
 $ chanceCreationCrossing: int  52 60 48 53 60 54 54 62 61 53 ...
 $ chanceCreationShooting: int  32 55 44 52 70 57 57 67 48 52 ...
 $ defencePressure       : int  44 30 42 25 55 49 49 47 47 36 ...
 $ defenceAggression     : int  58 30 40 56 60 51 51 47 56 39 ...
 $ defenceTeamWidth      : int  37 30 40 39 60 58 58 63 50 64 ...
```

By using these table I calculated feature analysis of teams for home and

away teams w.r.t. results of matches and co-relation between features towards the results of matches. Now, I would like to use these factors that affects match results (which I extract from above analysis) to implement results of matches. For prediction model, I create new table, which gives information about the results of teams with date, type of team (home or away) and difference of all features for team A – team B where I will consider type of team for team A. this table looks like as below which I used to build prediction models:

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       51258 obs. of  10 variables:
 $ results               : Factor w/ 3 levels "Draw","Loss",..: 1 1 2 3 1 1 3 3 2 2 ...
 $ team_type             : Factor w/ 2 levels "home","away": 1 1 1 1 1 1 1 1 1 1 ...
 $ speed_diff            : num  10.33 -8.5 2.33 2.5 7.5 ...
 $ playPassing_diff      : num  2.833 -0.833 -1 -5.333 0.833 ...
 $ creationPassing_diff  : num  9.67 6 -4.5 -10.5 -1.33 ...
 $ crossing_diff         : num  -1.17 1.67 5.33 4 -6 ...
 $ shooting_diff         : num  2.08 -12.83 -3.33 -1.83 2.67 ...
 $ defencePressure_diff  : num  0.917 -0.167 0.5 5.333 -6 ...
 $ defenceAggression_diff: num  1.67 1 2 5 5 ...
 $ defenceTeamWidth_diff : num  1.083 -7.667 0.833 -1.333 -2.667 ...
```

3. **Phase 3: Consultation: In** this phase, I presented the final presentation (interpret results: step 7) of my project and got feedback from my project supervisor and project coordinator.

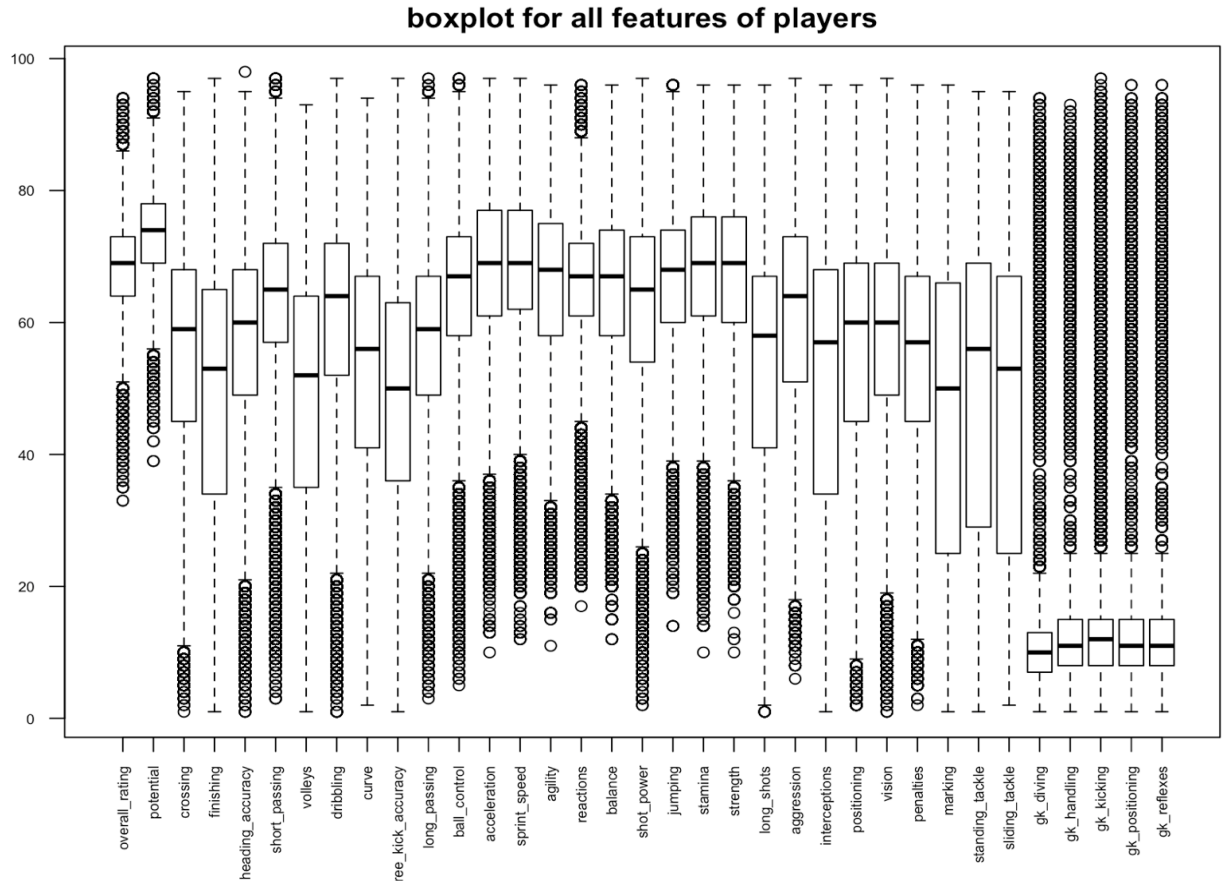
4. **Phase 4: Final:** Completed the data analysis report and research project (communicate results via report: step 8) and submission will be in week 14.

4. FINDINGS OF THE PROJECT

This section gives the details about all the outcomes of this project which are going to answer all the above said research questions with evidence built in R studio. First, I am going to start with the feature analysis and visualisations of players and teams. Thereafter, I will be going to give explanation about all the prediction models which I have done in R and comparison between them.

4.1 Visualisations and feature analysis at players level:

Here I used box plot to visualise the features of players:

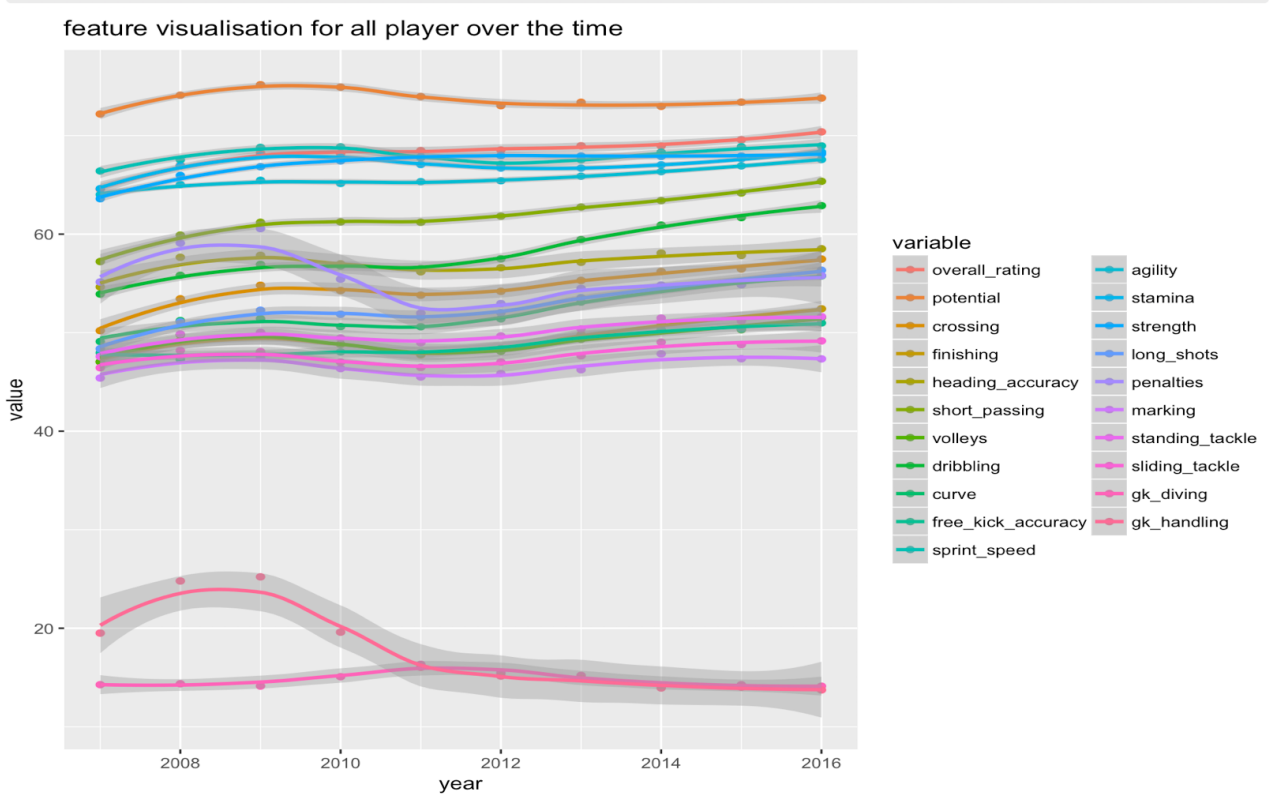


In above boxplot, there are so many variables who has a lot of outliers in both direction minimum and maximum. which means that this dataset has so many variations which could be possible as this data is quite big and has features for almost 11k players.

It can be clearly seen that finishing, volleys, curve, gk_accuracy, long-shots, interception, marking, standing_tackle, sliding_tacle has no outlier at all which means all values are near to mean value. These features could be considered as good to anlayze the features and performance of players. howeever, there are some features like gk_diving, gk_handling, gk_positioning and gk_reflexes have least variability but has a lot of outlier above the mean value which are scatter above the whole axis.

On the other side, marking, standing_tackle, sliding_tacle has more variability with no outlier at all. Crossing can be also good feature to be considered as has very small number of outliers below the mean.

Most of features has outliers below the mean value.
Hence, I have shortlisted important features overall rating, potentail, crossing, finishing, heading_accurancy, short_passing, volleys, dribbling, curve, free_kicking_accuracy, sprint_speed, agility, stamina, long_shorts, penalities, marking, standing_tackle, sliding_tackle, gk_diving, gk_handling for further visualisations.
Here, I visualised all the features of players over the time from 2008 to 2016:



Here, as in above figure of visualisation for all player over the time, interesting fact is there was major drop can be seen in the penalties over time which means that players improve their mistakes over the time as penalties decreases over the time. Moreover, gk_diving also decreases over the time period. on the other side, short_passing, dribbling, free_kick_accuracy and strength improves over the time as there was increases can be observed in the plot over the time.
However, rest of features remains almost stable over the time period.
Next, I was interested in match table as I wanted to connect the features of all players with match to extract the results of matches with players details so that I would able to analyse the features which

affects the results of matches. To do so I checked what is given inside the match table.

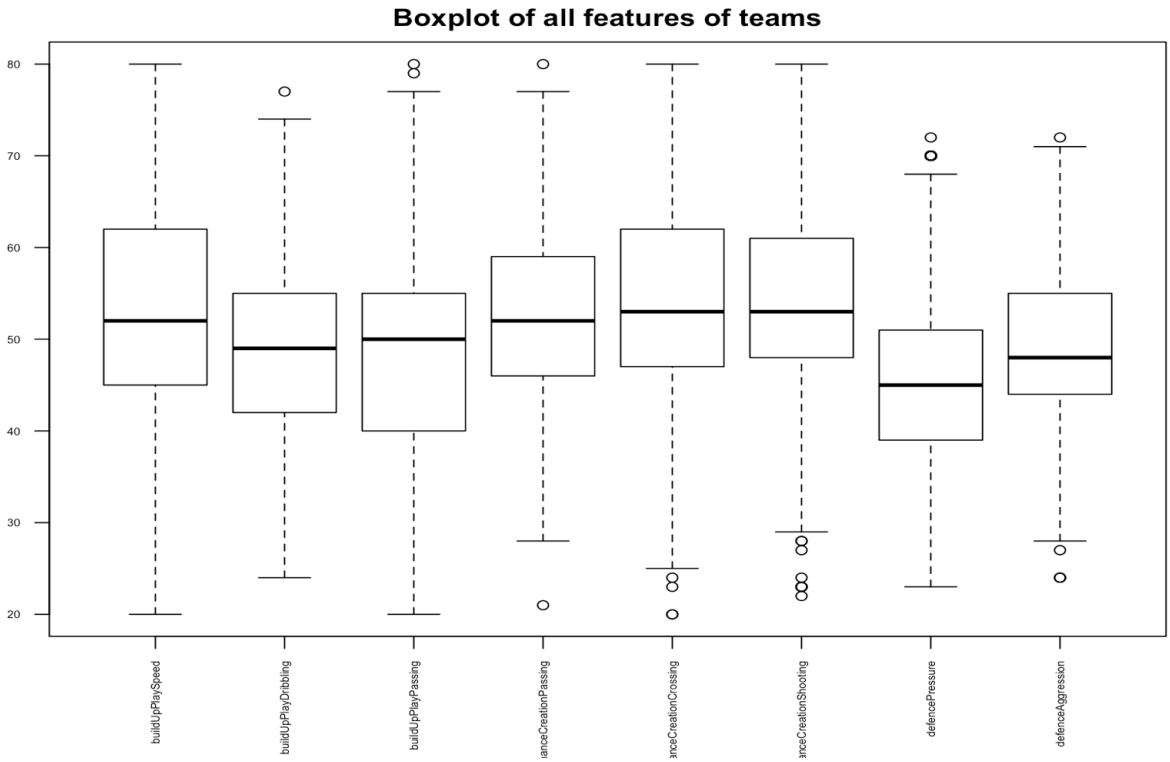
```
#checking what is available inside the match table for players and to observe how to connect match table and player
str(Match)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':       25979 obs. of  115 variables:
 $ id              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ country_id      : int  1 1 1 1 1 1 1 1 1 ...
 $ league_id       : int  1 1 1 1 1 1 1 1 1 ...
 $ season          : chr  "2008/2009" "2008/2009" "2008/2009" "2008/2009" ...
 $ stage           : int  1 1 1 1 1 1 1 1 10 ...
 $ date            : chr  "2008-08-17 00:00:00" "2008-08-16 00:00:00" "2008-08-16 00:00:00" "2008-08-17 00:00:00"
 $ match_api_id    : int  492473 492474 492475 492476 492477 492478 492479 492480 492481 492564 ...
 $ home_team_api_id: int  9987 10000 9984 9991 7947 8203 9999 4049 10001 8342 ...
 $ away_team_api_id: int  9993 9994 8635 9998 9985 8342 8571 9996 9986 8571 ...
 $ home_team_goal  : int  1 0 0 5 1 1 2 1 1 4 ...
 $ away_team_goal  : int  1 0 3 0 3 1 2 2 0 1 ...
 $ home_player_X1  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X2  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X3  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X4  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X5  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X6  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X7  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X8  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X9  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X10 : int  NA NA NA NA NA NA NA NA NA NA ...
 $ home_player_X11 : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X1  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X2  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X3  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X4  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X5  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X6  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X7  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X8  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X9  : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X10 : int  NA NA NA NA NA NA NA NA NA NA ...
 $ away_player_X11 : int  NA NA NA NA NA NA NA NA NA NA ...
```

Unfortunately, in match table there is no information is available for players in the match table. Here all the columns for home and away players are empty. so, there is no scope to connect the player table with match to analyse the characteristics of player through the performance in matches.

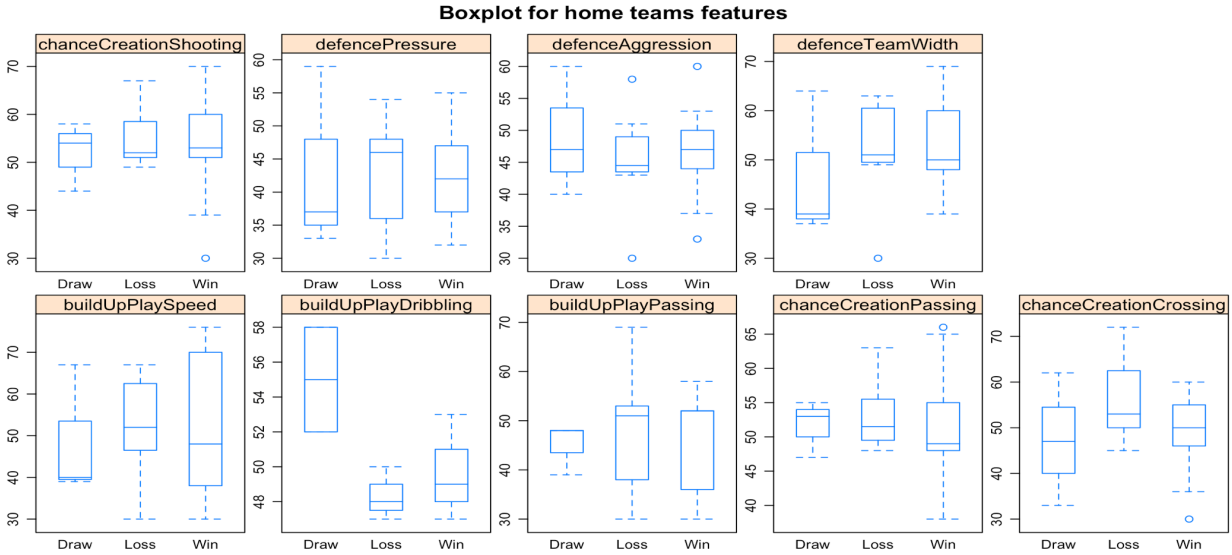
4.2 Visualisations and feature analysis at teams levels:

Next, I jumped to team's level visualisation. First of all, I checked the variation of different features with the help of box plot



Here all features of all teams are stable. if i compare the all features with each other, then most variant features are play speed, play passing and crossing. also, there are some outlier exists in all cases except play speed. Moreover, I am going to exclude dribbling as there are so many NA/null values are in dribbling.

4.2.1 Visualise all the team A (home team) features with respect to win draw and loss classes:



Note: here above plot gives information about home team so therefore defence team is away team

Above plot gives the information about the features of home team for three classes win, loss and draw.

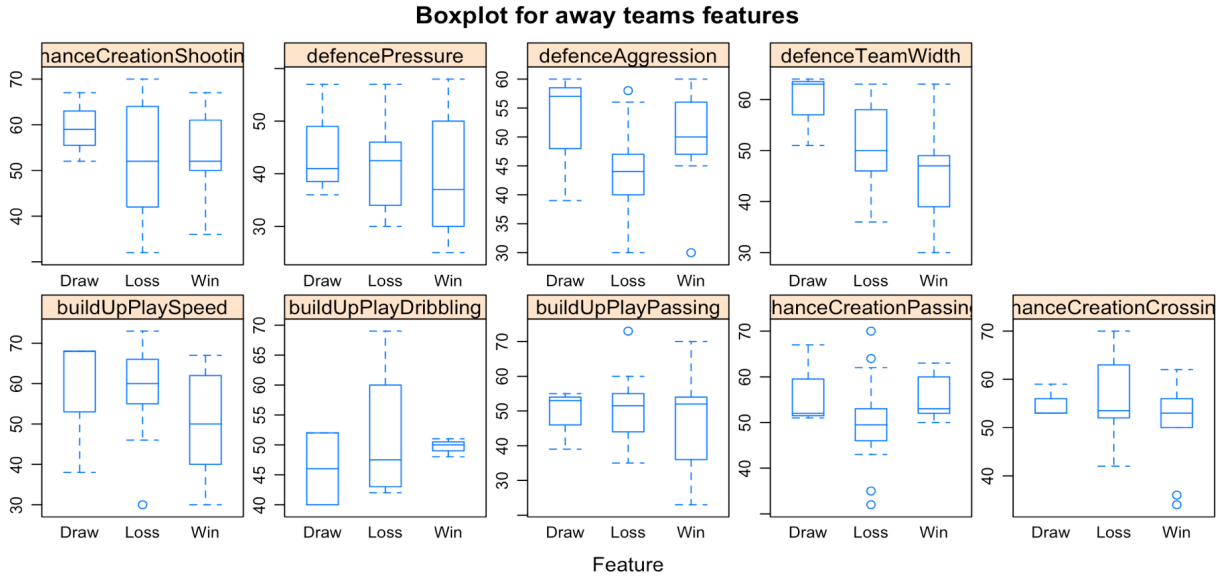
mean of shooting for all of three classes is almost same which means this feature does not contribute towards the match result of home team. But when the defence pressure against home teams is low that time home team wins but when its high against home team that time away team wins and when its very low that team its draw.

The mean of Defence aggression and defence team width against home team is almost same for when home team wins and loss.

when play speed varies and went highest in cases when home team wins.

Hence, defence pressure is very good feature for results of home teams as there was low pressure on home team against away when home team win.

4.2.2 Visualise all the team B (away team) features with respect to win draw and loss classes:

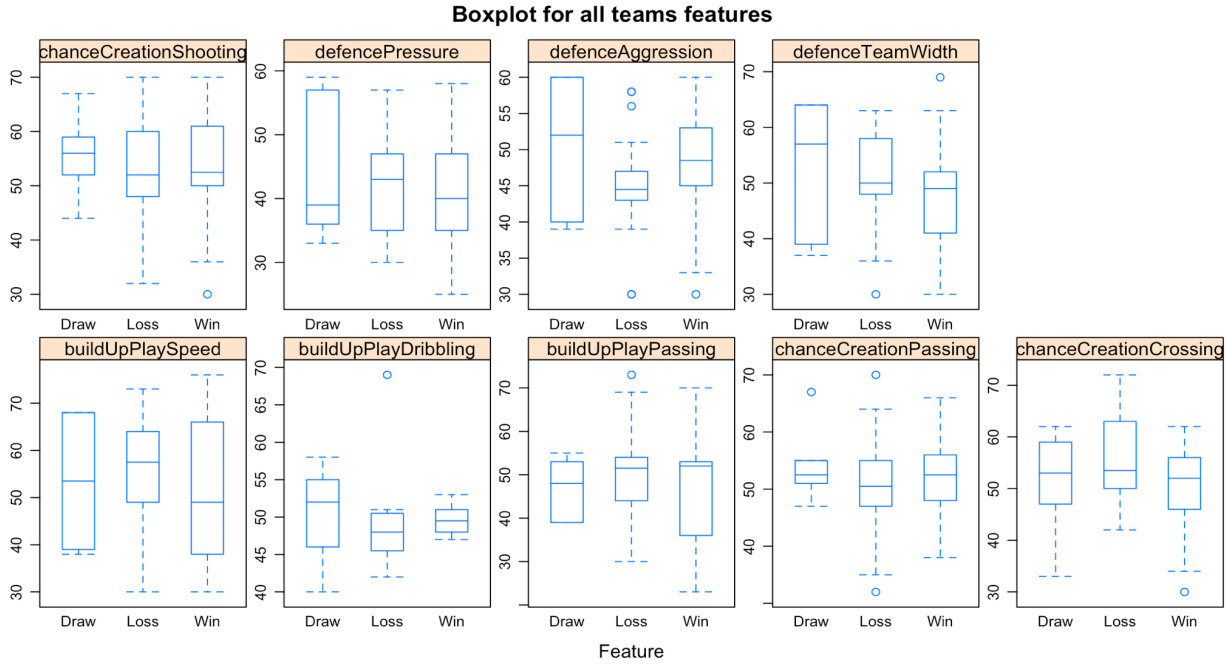


Note: here above plot gives information about away team so therefore defence team is home team

This plot gives the information about the features of away team for

three classes win, loss and draw. if we compare this plot with previous home vs away its clear that defence pressure and aggression of home team is always is very high against away team. now, when defence pressure of home team against away team is high that time away team losses the match when its low that time away team wins the match but opposite trend is observed in case of defence aggression like when home team's aggression is high that time more likely to chance of away team wins and when its low then home team wins the match. also, when home teams defence width id high that time its draw but with high value of home teams defence width it wins and with low it losses and vice versa. Away team wins when creation passing is high and loss when its lower

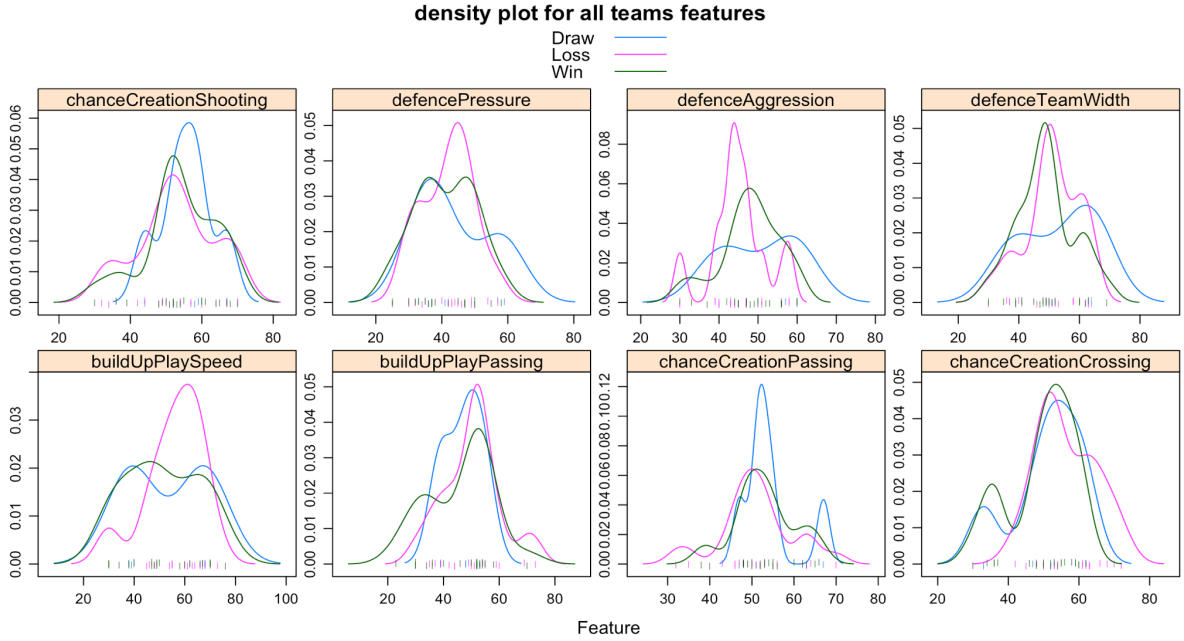
4.2.3 visualise features for all teams w.r.t. W/D/L box plot:



This plot is giving information about all the teams (home and away) for all features against home win and loss class. overall, there is more chances to be winner of match if team has less defence pressure and if its high there is more chance to loss a match. But opposite trend and interesting fact can be seen when it comes to defence aggression like when defence teams show more aggression then team wins against aggression but losses when its

low.

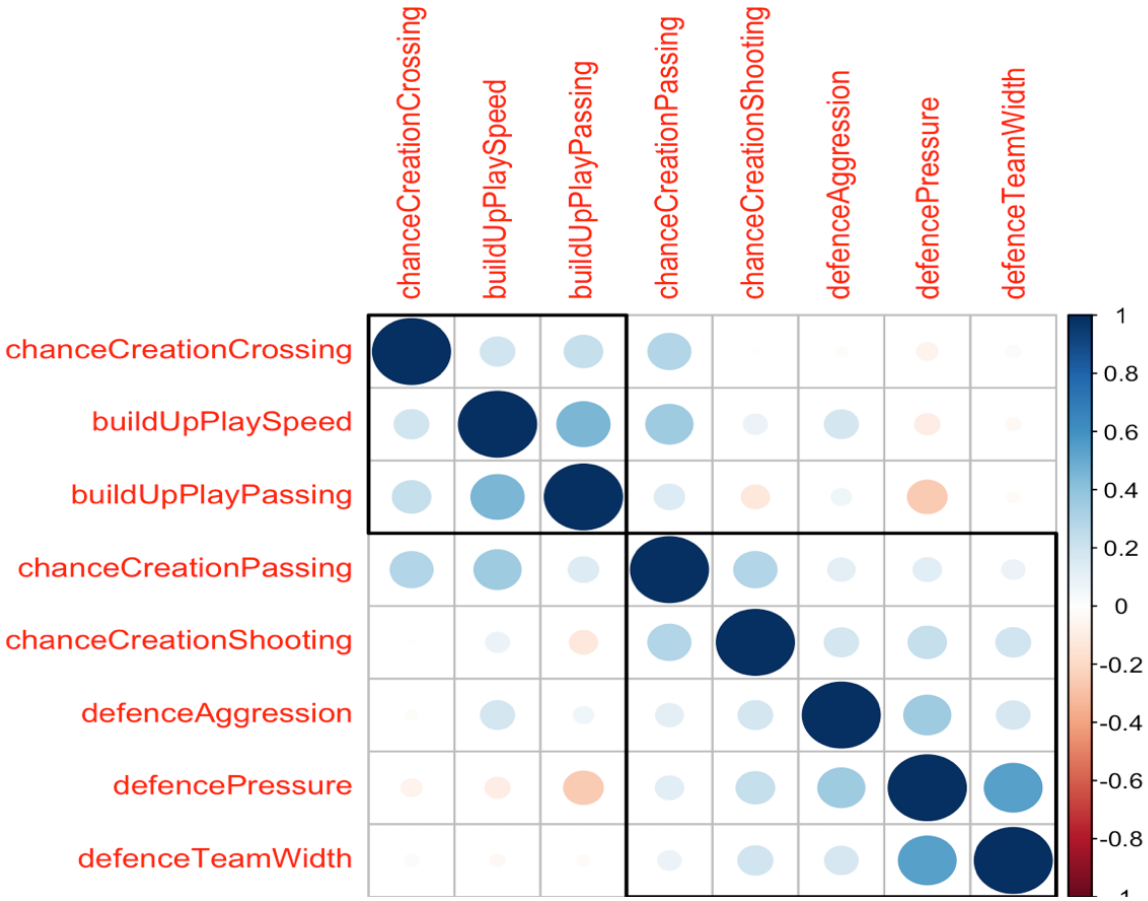
4.2.4 visualise features w.r.t. W/D/L density plot:



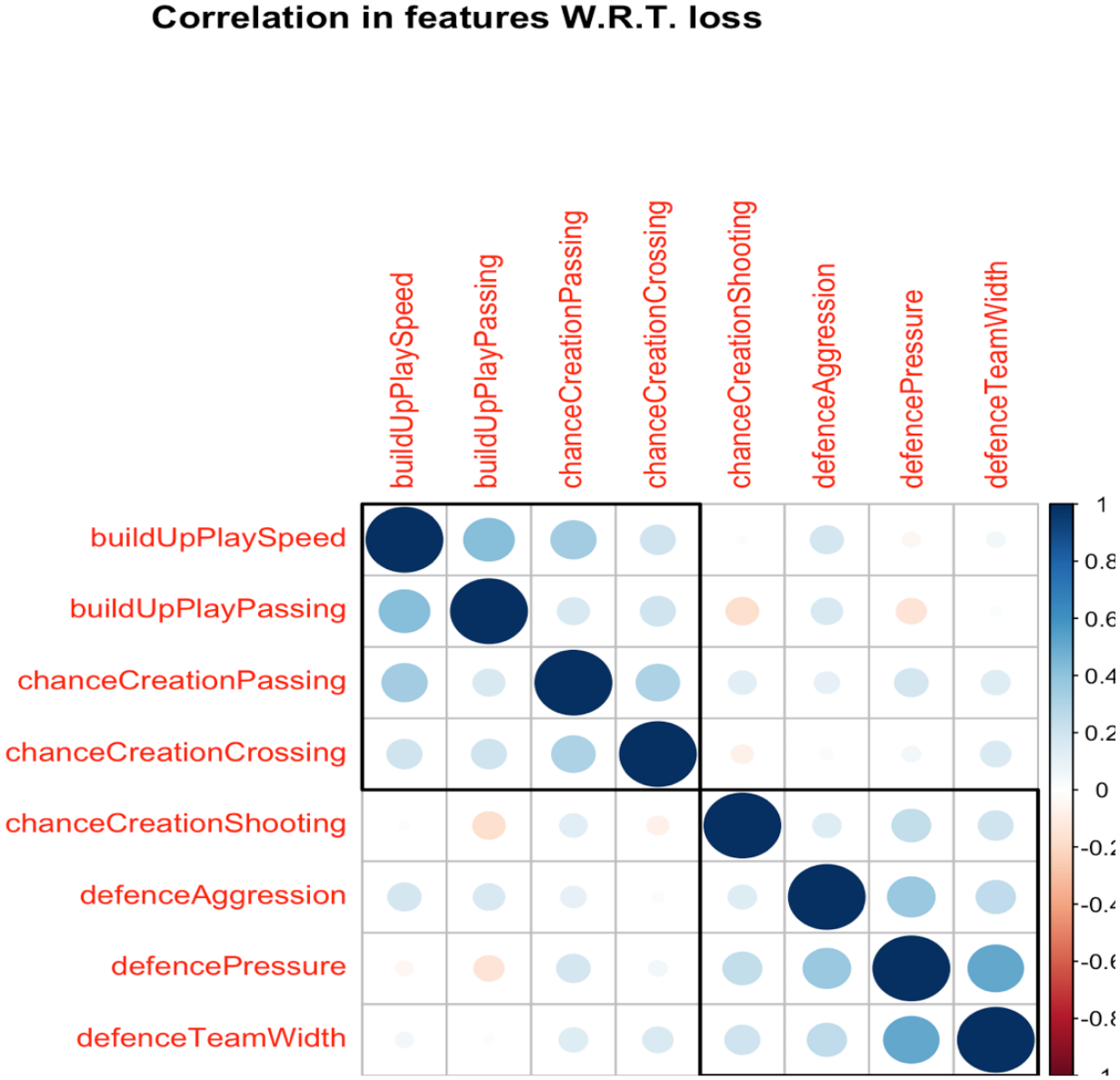
Here, from density plot, there is no strong relationship for win class and all features of teams but team will always loss when there is high defence aggression and pressure and play passing. and its draw when there is high creation shooting and creation passing.

4.3 Finding co-relation of all features of teams with loss and win class:

Correlation in features W.R.T. win



Here, Correlation in features W.R.T. win shows that play speed has positive co-relation with play passing, creation passing and defence aggression like when they highly co related with positive relation then there is high chance to win. moreover, crossing has also positive co relation for win class with creation passing. there is negative co relation between defence pressure and play passing in case of win.



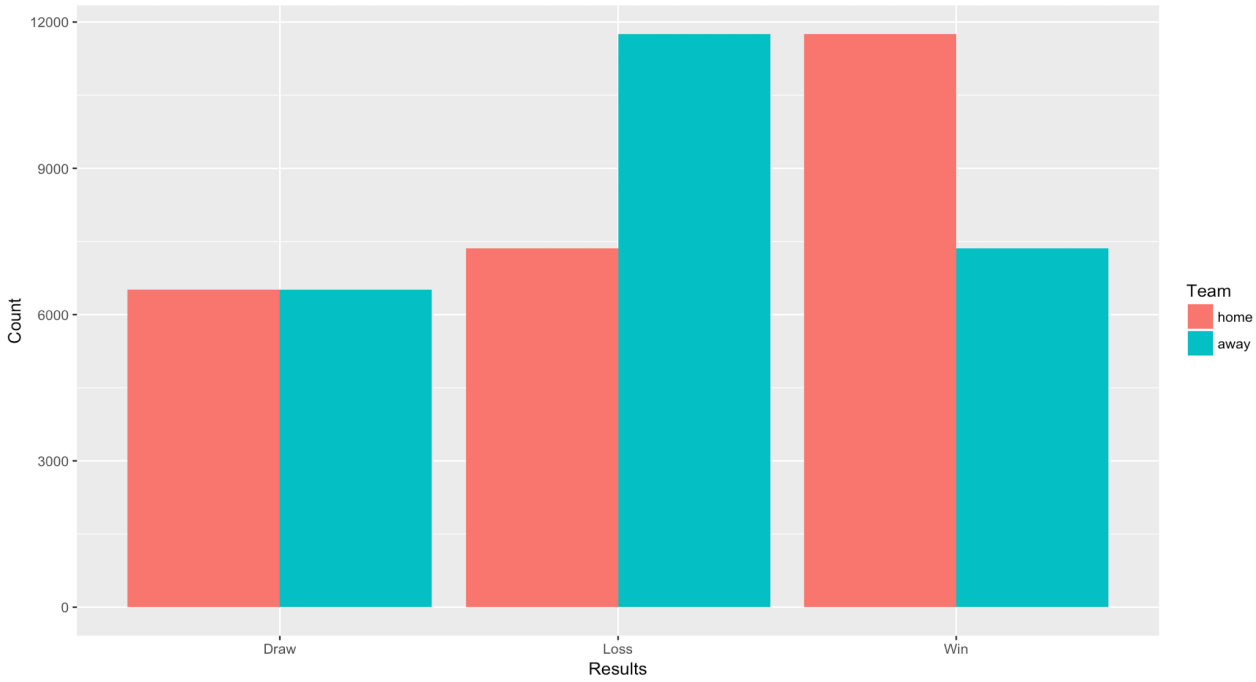
When observing against loss class, play speed has positive relation with play passing and creation passing. but play passing has negative co relation with creation shooting and defence pressure but defence pressure and defence aggression has positive co relation for loss class. which means when defence pressure and defence aggression are positively co related then there is more chance that team will loss the match.

4.4 PREDICTION MODEL: To predict the results of matches

4.4.1 Why home and away feature is an important to predict results:

First of all, I wanted to know whether the type of team

(home or away) impact the results of matches or not. So, should I consider this as good feature to predict the match results.



Here it is clear from the bar graph that home and away is good feature to consider for prediction model as when in case of home teams wins more matches than away and less loss than away teams.

4.4.2 DECISION TREE:

Confusion matrix for decision tree when considering two class (win and loss) left side vs 3 classes (draw as well) right side

```
     Loss  Win
Loss 3566 1947
Win  2180 3753
Confusion Matrix and Statistics

          Reference
Prediction Loss  Win
      Loss 3566 1947
      Win  2180 3753

               Accuracy : 0.6394
                 95% CI : (0.6306, 0.6482)
    No Information Rate : 0.502
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.279
 Mcnemar's Test P-Value : 0.0003046
```

```
Confusion Matrix and Statistics

          Reference
Prediction Draw Loss  Win
      Draw    0    0    0
      Loss 1899 3493 2053
      Win  2036 2239 3675

Overall Statistics

               Accuracy : 0.4656
                 95% CI : (0.4577, 0.4735)
    No Information Rate : 0.3723
    P-Value [Acc > NIR] : < 2.2e-16
```

figure: decision tree stats for two classes *figure: decision tree stats for three classes*

Left side figure gives the statistics about the two classes win and loss and right side when there are three classes win, loss and draw. It is obvious that it is always difficult to predict draw class. Therefore, accuracy rate is higher (approx. 64%) in case of two classes than that of (approx. 47%) when trying to predict draw class as well.
For two classes: it gives 7319 correct predictions/results for win and loss class out of 11446 and giving 4127 wrong results for win and loss class. decision tree start node is weather team is home or away.

For three classes: Here, decision tree model's accuracy rate is 46.56% for three classes loss, win and draw, but this is not predicting draw class accurately. it gives 7168 correct predictions/results for win and loss class out of 15395. it is very clear that it is very difficult to predict draw class and giving 3935 wrong results for draw class. decision tree start node is weather team is home or away.

4.4.3 SVM: Support vector machine model: Confusion matrix for SVM when considering two class (win and loss) left side vs 3 classes (draw as well) right side

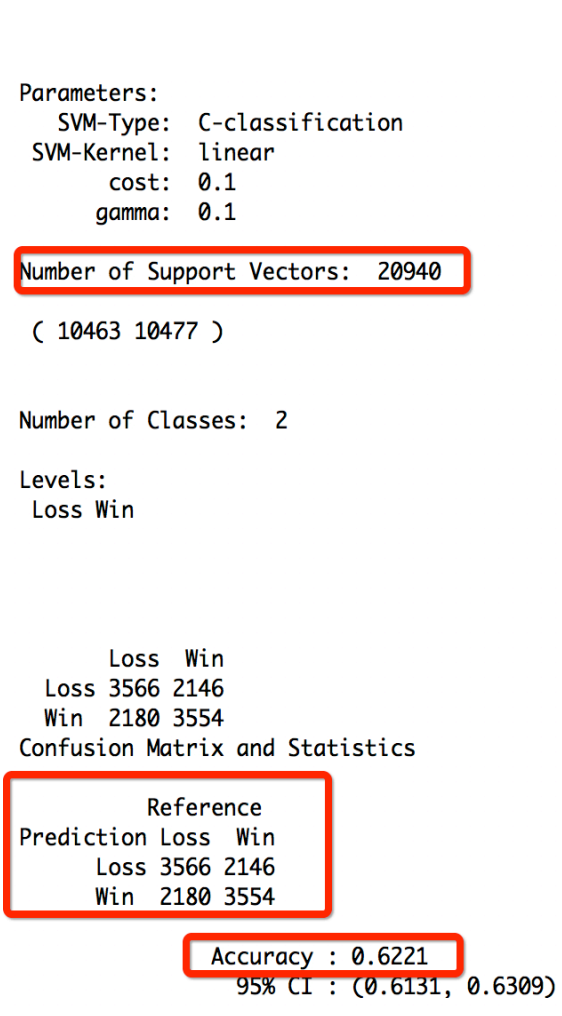


figure: SVM stats for two classes

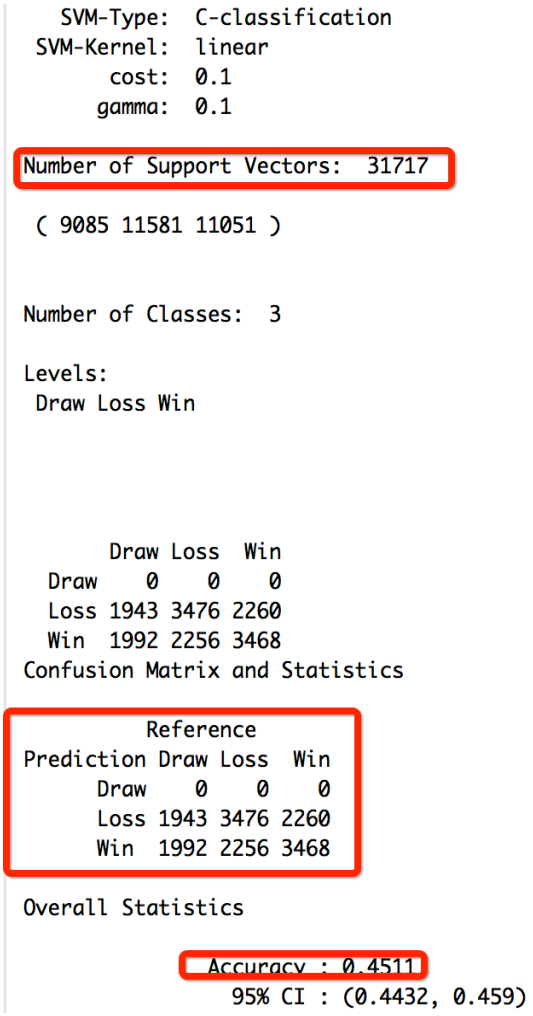


figure: SVM stats for three classes

Left side figure gives the statistics about the two classes win and loss and right side when there are three classes win, loss and draw. It is obvious that it is always difficult to predict draw class. Therefore, accuracy rate is higher (approx. 62%) in case of two classes than that of (approx. 45%) when trying to predict draw class as well.
For 2 classes: SVM model's accuracy rate is 62.11% which is lower than the decision tree model. there are 20940 support vectors for this model and it gives 7120 correct prediction results out of 15395.
For 3 classes: SVM model's accuracy rate is 45.11% which is lower than the decision tree model. there are 31717 support vectors for this model and it gives 6944 correct prediction results out of 11446.

4.4.4 Random forest model:

All the statistics for random forest model

when considering two class (win and loss) left side vs 3 classes (draw as well) right side.

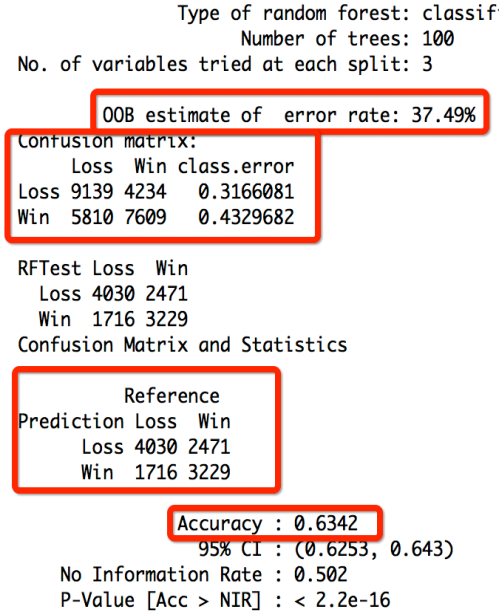


figure: random forest stats for two classes

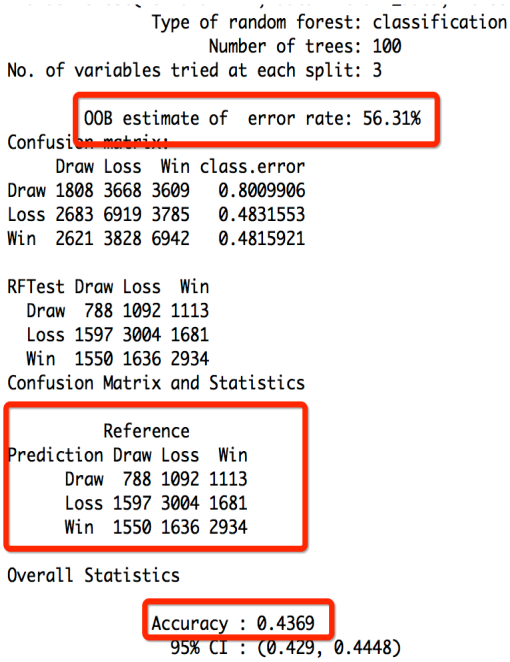


figure: random forest stats for three classes

Left side figure gives the statistics about the two classes win and loss and right side when there are three classes win, loss and draw.

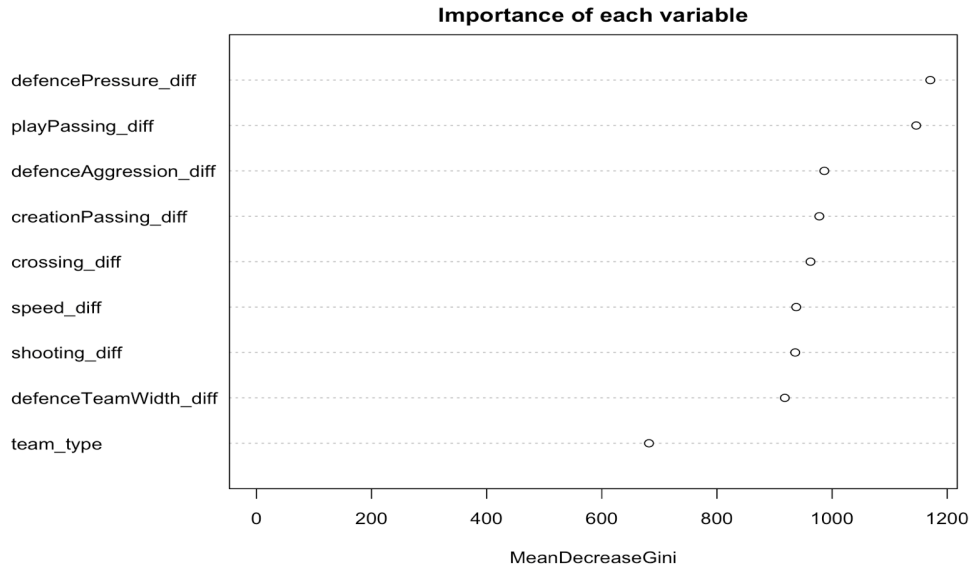
It is obvious that it is always difficult to predict draw class. Therefore, accuracy rate is higher (approx. 63%) in case of two classes than that of (approx. 43%) when trying to predict draw class as well. And also error rate is lower for two classes than that of three classes.

For 2 classes: Random forest model gives 63.42% accuracy rate and error rate is 37.49%. Error rate for win and loss is 43.29% and 31.66% respectively. This model is giving total of 6726 correct results out of 11446.

For 3 classes: Random forest model gives 43.69% accuracy rate and error rate is 56.31%. it gives highest error rate (approx. 80%) for draw class as it is unable to predict correctly. Error rate for win and loss is 48.15% and

48.31% respectively. This model is giving total of 7259 correct results out of 15395.

Importance of each variable is as following:



	MeanDecreaseGini
team_type	681.9648
speed_diff	937.7286
playPassing_diff	1146.1493
creationPassing_diff	977.8724
crossing_diff	962.5399
shooting_diff	935.9808
defencePressure_diff	1170.5885
defenceAggression_diff	986.5458
defenceTeamWidth_diff	917.9337

Here, above figure depicts that defence team width difference is most important feature and team type is least important.

4.5 Comparison between all of three models:

Out of above three prediction models, decision tree model always gives highest accuracy rate but decision tree and SVM completely giving wrong results for draw class however, random forest is only model which is predicting all of three classes but giving 43.7%

accuracy rate. So, random forest can be considered as good prediction model when we considering all of three classes and decision tree gives more accuracy when its two classes win and loss.

5 DISCUSSION

In this section, I am going to discuss about the implications of findings, significance of findings, limitations of this project and future work or potential work in this context.

5.1 Critical Analysis of Findings

- Hence, the important features which can affect the match results can be whether team is home or away as home team contributes towards win class.
- Similarly, as discussed in section 4.2 and 4.3, the defense pressure, defense aggression, play passing and defense team width also impact the match results which can be also considered as good features for prediction model.
- Out of above three predictions models, decision tree model gives highest accuracy rate but decision tree and SVM completely giving wrong results for draw class however, random forest is only model which is predicting all of three classes but giving 43.7% accuracy rate. so, random forest can be considered as good prediction model when we considering all of three classes and decision tree gives more accuracy when its two classes win and loss

5.2 Significance of findings

1. By understanding the characteristics of teams will help the football clubs and scouting to identify strengths and weaknesses of their teams and opposite teams so that they can use these insights to improve the training level and increase the awareness of team against opposition to prepare them for next challenges.
2. By identifying the weaknesses of teams, Coaches can implement new fitness training programme to improve the performance of

teams.

3. Deep comparisons of different team's characteristics and abilities can be useful for clubs to make decision of choosing teams and development or selection of teams and players.
4. The trend of performance by tracking the performance of particular players can give idea about the future performance of that player.
5. Moreover, many professional betting agencies are making a lot of money by predicting match results. So, it means betting odds are calculated to increase the profits and minimize their risks (wrong predictions). So, by analysing the factors which affects the results of matches, teams and players performance can be used to predict the results of matches and performance of team and players in which betting industries are highly interested in.
6. Moreover, the knowledge is hidden within the soccer clubs/management level, it's not disclose to the general public. The project aimed to provide insights to team's characteristics / strength/weakness to general public.

5.3 Limitations of the Project

- The main limitation of this project is that, it is very always difficult to predict draw results of matches and the prediction models are giving very low accuracy rate. Therefore, the prediction models would not able to predict three classes win, loss and draw classes accurately.
- Moreover, there was data about players was missing in match table where results of matches are given in the dataset. So, I was unable to analyse the features of players w.r.t results of matches.

5.4 Future Work

If there will be information about the players and features of players available in match table where results of matches are given. Then, there will be possibility of analysing the features of players attributes or characteristics that what factors of players features affects the results of matches which will help to find out what kind of features and factors can affect the results of matches that will able to help to implement prediction models for results of matches.

6 CONCLUSION

To cap it all, I able to answer all the research questions which I mentioned in section 1. With the help of analysis and visualisations done in R, I found there are some factors that affects the results of matches like if team is playing at their home ground then there are more chances to win the match. Moreover, the defense pressure, defense aggression, play passing and defense team width also impact the match results which can be also considered as good features for prediction model. Because, correlation in features W.R.T. win depicts that play speed has positive co-relation with play passing, creation passing and defence aggression like when they highly co related with positive relation then there is high chance to win and when defence pressure and defence aggression are positively co related then there is more chance that team will loss the match. Also, team will always loss when there is high defence aggression and pressure and play passing. and its draw when there is high creation shooting and creation passing. Decision tree model can be considered as good model to predict results as it is giving high accuracy rate.

7 REFERENCES

- Collignon, H. & Sultan, N. The Sports Market: Major trends and challenges in an industry full of passion. Retrieved from: https://www.atkearney.com/documents/10192/6f46b880-f8d1-4909-9960-cc605bb1ff34
- Mr Neutral (January 21, 2014). Soccer Betting: A Worldwide Gambling Industry Worth Billions. Retrieved from: http://www.caughtoffside.com/2014/01/21/soccer-betting-a-worldwide-gambling-industry-worth-billions
- Kang, C.H. & Hwang, J.R. & Li, K.J. Trajectory analysis for soccer players. In Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on, pages 377–381. IEEE, 2006.
- Carling C, Williams AM, Reilly T. What match analysis tells us about succeessful strategy and tactics in soccer Handbook of soccer match analysis: a systematic approach to improving performance. London: Routledge; 2005. pp. 108–128.
- Mackenzie R, Cushion C. Performance analysis in football: a critical review and implications for future research. J Sports Sci. 2013;31(6):639–676. doi: 10.1080/02640414.2012.746720.
- Yiannakos A, Armatas V. Evaluation of the goal scoring patterns in European Championship in Portugal 2004. Int J Perform Anal Sport. 2006;6(1):178–188.
- GamblingSites.com(n.d.). What Affects the Outcome of Football Games? Retrieved August 2017, from https://www.gamblingsites.com/football-betting/strategy/what-affects-outcome
- Irving, P.G. and Goldstein, S.R. (1990) 'Effects of Home-field Advantage on Peak Performance of Baseball Pitchers', Journal of Sport Behavior 13: 23–7.
- Smith, D.R. (2003) 'The Home Advantage Revisited: Winning and Crowd Support in an Era of National Publics', Journal of Sport and Social Issues 27: 346–71.
- Smith, D.R., Ciacciarelli, A., Serzan, J. and Lambert, D. (2000) 'Travel and the Home Advantage in Professional Sports', Sociology of Sport Journal 17: 364–85.
- Schreck, T. Daniel A. Keim & Oliver Deussen. Feature-Driven Visual Analytics of Soccer Data.

- Ray, S. (2017). Essentials of Machine Learning Algorithms. Retrieved from: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms
- Brownlee, J. (2014). Feature Selection with the Caret R Package. Retrieved from: https://machinelearningmastery.com/feature-selection-with-the-caret-r-package
- Brownlee, J. (2016). Better Understand Your Data in R Using Visualization. Retrieved from: https://machinelearningmastery.com/data-visualization-in-r
- FeedBurner (n.d.). Correlation matrix: A quick start guide to analyze, format and visualize a correlation matrix using R software. Retrieved from: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software
- Arsanjani, J.J. & Helbich, M. & Tayyebi, A. & Birenboim, A. (1 january 2017). How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects**.** Retrieved from: www.mdpi.com/journal/data
- The DSDM Agile Project Framework (2014 Onwards) (n.d.). Agile Business Consortium. Retrieved from https://www.agilebusiness.org/content/principles
- V Dhar, Data Science and Prediction, Communications of the ACM, vol 56, No 12, December 2013.
- BS ISO 21500:2012: Guidance on project management (2012). British Standards Institute
- Byrne, C. (2017). Development Work ows for Data Scientists. Retrieved from: https://resources.github.com/downloads/development-workflows-data-scientists.pdf
- Barter, R. (2017). A Basic Data Science Workflow. Retrieved from: http://www.rebeccabarter.com/blog/2017-08-16-data-science-workflow
- Schedlbauer, M. (2011, February 22). Requirements Prioritization Strategies. Retrieved from https https://www.projecttimes.com/articles/requirements-prioritization-strategies.html

8 REFLECTIONS

- **What were the things/activities you thought you did best in this project?**
  I have used DSDM project management approach for this project, which enabled me to successfully complete the project on time without compromising the quality. With the help of GitHub and Slack, I was able to have communications with my supervisor conveniently. It enabled me to get in touch with my supervisor when I need his assistance when he was away or out of country to get feedback. Moreover, the prioritization of tasks helped me to achieve all the promised deliverable on time.

- **What were the things/activities you thought you did least well in this project?**
  I struggled in writing report because of my other two unit's deliverables due on same day. But, I am thankful to Dr. Guido, who gave extension for the completion of this report.

- **Were there any specific problems or challenges you encountered? How did you handle them?**
  Firstly, I got disappointed when there was information about the players in match table where results of matches were given. Therefore, I was unable to analyse the features of players w.r.t. results of matches and how and what factors of players attributes affects the match results. Thereafter, I moved to team's level to analyse the features at team level as there was information about teams given.
- **What did you find as the hardest part of this project?**
  Implementation of prediction model was hardest part of this project as it took too much time and there were errors during the implementation. Also, this dataset has 8 tables and I had to calculate many this myself and sometimes all the statistics went very complicated and confusing because there was not everything given directly. I had to extract the things from different table and to make sense of those things I joint them with other tables logically.
  The database was bit tricky and some values from dataset was missing which were actually required for analysis.
- **Which areas of your own professional knowledge and skills, or your personal attributes do you feel require further development?**

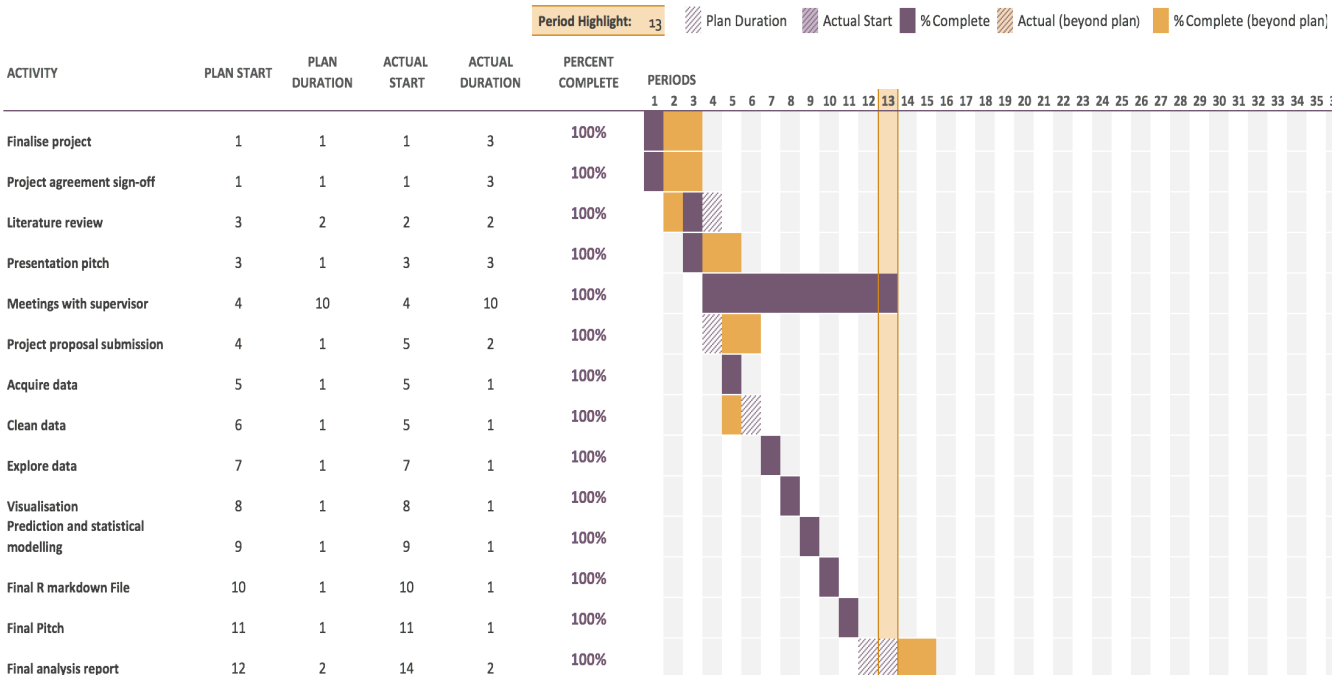
I need to work on how to implement perfect prediction models and how to manage all things at the same time to complete them and mange successfully on the time.

- **What was the most important thing you learned doing this project?**
  Overall, it was very good practice that how to analyse the big dataset which having 8 tables and how to extract new things from data and manipulate the data. It was good practice of analysis and visualisation of data and implementation of prediction models.
- **What have you learnt about project management and research within professional practice?**
  The project management play a very important role to achieve all the targeted goals. It enables me to breakdown of all tasks that need to be performed to get all goals. Through professional practice I learnt that how to complete important tasks on time. Also, communication with the supervisor is also very important to achieve the targeted goals.

- **What kind of opportunities or next steps you see based on your learnings doing this project?**
  This project gave me experience that how to work on big and complicated datasets. In future, I would love to work on various datasets for analysis and visualisation and implement predict models. I would like to explore more dataset on matches or football matches.

9. Appendix A

Project Planner



ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Finalise project	1	1	1	3	100%
Project agreement sign-off	1	1	1	3	100%
Literature review	3	2	2	2	100%
Presentation pitch	3	1	3	3	100%
Meetings with supervisor	4	10	4	10	100%
Project proposal submission	4	1	5	2	100%
Acquire data	5	1	5	1	100%
Clean data	6	1	5	1	100%
Explore data	7	1	7	1	100%
Visualisation	8	1	8	1	100%
Prediction and statistical modelling	9	1	9	1	100%
Final R markdown File	10	1	10	1	100%
Final Pitch	11	1	11	1	100%
Final analysis report	12	2	14	2	100%

10. Appendix B: Logsheets

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

1. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
6. The log sheet is NOT a deliverable for the project but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Harmandeep Kaur Bhullar **Date:** 1st **Aug, 2017** **Meeting No:1**

Project title: European Soccer database: Data analysis and research project **UNIT:** IFN701

☐ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Guido Zuccon **Supervisor's Signature:** Guido Zuccon

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1.

2.

3.

Items for discussion (noted by student before mandatory supervisory meeting):

1. Gain knowledge in project and dataset
2. Gain insights what kind of work to be done
3. Context of literature review

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Load data and clean the data
2. Get familiar with data and work on project plan
3. create a repository on gitHub and join IFN701 group on slack

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

7. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
8. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
9. A log sheet is to be brought by the STUDENT to each supervisory session.
10. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
11. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
12. The log sheet is NOT a deliverable for the project, but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Harmandeep Kaur Bhullar **Date: 25th Aug. 2017** **Meeting No: 2**

Project title: European Soccer database: Data analysis and research project **UNIT:** IFN701

☐ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Guido Zuccon **Supervisor's Signature:** Guido Zuccon

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Literature review
2. Explored the data and cleaned
3. Working on project plan

Items for discussion (noted by student before mandatory supervisory meeting):

1. Get feedback on project plan and project presentation
2. Discuss about research questions
3. Dr. Guido arranged a lab on Level 10

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. explore more data by visualizing the features using boxplot.
2. Implement changes in project plan as per suggestions given by Dr. Guido

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

13. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
14. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
15. A log sheet is to be brought by the STUDENT to each supervisory session.
16. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
17. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
18. The log sheet is NOT a deliverable for the project, but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Harmandeep Kaur Bhullar **Date: 23 th Sep. 2017** **Meeting No: 3**

Project title: European Soccer database: Data analysis and research project **UNIT:** IFN701

☐ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Guido Zuccon **Supervisor's Signature:** Guido Zuccon

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. visualized features using boxplot

Items for discussion (noted by student before mandatory supervisory meeting):

1. Discuss everything what is done so far to get feedback
2. Other way to visualize features and how to improve boxplots

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. improve visualization and add more visualization
2. write explanation(theory) for everything done in Rmd file

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

19. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
20. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
21. A log sheet is to be brought by the STUDENT to each supervisory session.
22. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
23. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
24. The log sheet is NOT a deliverable for the project, but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Harmandeep Kaur Bhullar **Date: 23 th Sep. 2017** **Meeting No: 4**

Project title: European Soccer database: Data analysis and research project **UNIT:** IFN701

☐ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Guido Zuccon **Supervisor's Signature:** Guido Zuccon

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. improved visualization by taking mean of features and implemented more visualizations
2. analysis and explanation written in RMD file

Items for discussion (noted by student before mandatory supervisory meeting):

1. Discuss everything done so far to get feedback
2. Discuss the missing data for players which restrict to analyze the features of players and makes it unable to predict results from players features
3. Prediction models

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Analyze the team level features and start working on prediction models by considering team features as data about team features is given
2. Implement decision tree, SVM and random forest

Note. A student should make an appointment to meet the supervisor in advance, usually at least 1 week prior.

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

25. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
26. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
27. A log sheet is to be brought by the STUDENT to each supervisory session.
28. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
29. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
30. The log sheet is NOT a deliverable for the project, but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Harmandeep Kaur Bhullar **Date:** 10th Oct. 2017 **Meeting No: 5**

Project title: European Soccer database: Data analysis and research project **UNIT:** IFN701

☐ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Guido Zuccon **Supervisor's Signature:** Guido Zuccon

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Analyzed and visualized team level features
2. Prediction model Decision tree

Items for discussion (noted by student before mandatory supervisory meeting):

1. Discuss everything to get feedback
2. Errors in random forest model

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

1. Write explanation for everything done in RMD
2. Solve errors in random forest and complete SVM

Note: As I mentioned in my project proposal the formal meetings will be after 15 days or as required. Therefore, I only did 5 formal meetings because most of the time we communicate by ***Slack.***

Project Log Sheet – Supervisory Session

Note on use of the project log sheet:

31. This log sheet is designed for all formal meetings, of which there must be at minimum SEVEN (7) during the course of the project (SEVEN mandatory supervisory sessions).
32. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant section of the form, effectively forming an agenda for the session.
33. A log sheet is to be brought by the STUDENT to each supervisory session.
34. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
35. It is recommended that students bring along log sheets of previous meetings during each supervisory session.
36. The log sheet is NOT a deliverable for the project, but it is an important record of a student's organization and learning experience. The students will be asked to hand in the log sheets as an appendix of the final report, with sheets dated and numbered consecutively. This is an important part of evidence on how you managed your project during the semester.

Student's Name: Harmandeep Kaur Bhullar **Date:** 17th Oct. 2017 **Meeting No: 6**

Project title: European Soccer database: Data analysis and research project **UNIT:** IFN701

☐ Journal entry logged into Blackboard (Optional)

Supervisor's Name: Dr. Guido Zuccon **Supervisor's Signature:** Guido Zuccon

Update on progress since last meeting, and challenges faced if any (noted by student before mandatory supervisory meeting):

1. Implemented SVM and random forest
2. Start working on final presentation

Items for discussion (noted by student before mandatory supervisory meeting):

4. Discuss everything to get feedback
5. Get suggestions for presentation and report writing

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

3. Finish with presentation and share on GitHub to get feedback
4. Start working on final report side by side

Note: As I mentioned in my project proposal the formal meetings will be after 15 days or as required. Therefore, I only did 5 formal meetings because most of the time we communicate by ***Slack.***