CAPSTONE PROJECT

ANSARI ZAID

MAY A BATCH

CRICKET WIN PREDICTION

Table of content

1) INTRODUCTION

Brief introduction of problem statement and the need of solving it……………………………….1

2) Exploratory data analysis and business implications…………………………………………………1

Head of the data ……………………………………………………………………………………………….2

Data dictionary ………………………………………………………………………………………………..2, 3

Shape of the data ……………………………………………………………………………………………..3

Duplicates ………………………………………………………………………………………………………..3

Description of the data ………………………………………………………………………………………3, 4

Information about the data ……………………………………………………………………………….4

Univariate analysis ……………………………………………………………………………………………5

Hygiene check of the data ……………………………………………………………………………5

Box plot for continuous variables ………………………………………………………………….6, 7

Count plot for categorical variables ………………………………………………………………8

Bivariate analysis ……………………………………………………………………………………………..9, 16

Business insights from exploratory data analysis ……………………………………………….16

3) Data cleaning and preprocessing ……………………………………………………………………….16

Approach used for identifying and treating missing values and outlier's treatments …….16, 19

Variables transformation ………………………………………………………………………………….19

4) Model building ……………………………………………………………………………………………….20, 21

Logistic regression using stats model ………………………………………………………………..21, 22

Decision tree classifier(CART) ……………………………………………………………………………22, 23

Random forest ………………………………………………………………………………………………….23

Artificial neural network ……………………………………………………………………………………23

Efforts to improve model performance …………………………………………………………….23, 25

5) Model validation ……………………………………………………………………………………………..25

Model validation done on test set

Logistic regression ………………………………………………………………………………………….25, 26

Random forest ………………………………………………………………………………………………..26, 27

Decision tree(CART) ………………………………………………………………………………………..27, 28

Artificial neural network ………………………………………………………………………………….28, 29

Interpretation of the models …………………………………………………………………………..29, 30

6) Final recommendation given to BCCI ………………………………………………………………..30 ,31

List of figures

figure 1: Box plot of continuous variables ……………………………………6
figure 2: Count plot of categorical variables…………………………………8
figure 3: Correlation plot……………………………………………………………9
figure 4: Pair plot………………………………………………………………10
figure 5: Scatter plot…………………………………………………………11
figure 6: Box plot of extra bowls opponents………………………………11
figure 7: Box plot of extra bowls bowled ……………………………………12
figure 8: Bar plot of player highest wicket…………………………………….12
figure 9: count plot of average team age……………………………………….13
figure 10: Stack bar plot of offshore…………………………………………13
figure 11: Stack bar plot of First selection………………………………….13
figure 12: count plot of all-rounder in team………………………………….14
figure 13: count plot of opponents…………………………………………….15
figure 14: count plot of match light type………………………………………15
figure 15: Count plot of player scored zero…………………………………….16
figure 16: Box plot after removal of outliers…………………………………..19
figure 17: roc auc curve of logistic regression using grid search for test ….26
figure 18: roc auc curve of RF model using grid search ………………………..27
figure 19: roc auc of DT model for test set ………………………………….....28
figure 20: roc auc of ANN model for test set ………………………………………29

List of tables

Table 1: Interpretation of all the models …………………………………………….29

FINAL BUSINESS REPORT

1) INTRODUCTION

BCCI has hired an external analytics consulting firm for data analytics. The major objective of this tie up is to extract actionable insights from the historical match data and make strategic changes to make India win. Primary objective is to create Machine Learning models which correctly predicts a win for the Indian Cricket Team. Once a model is developed then you have to extract actionable insights and recommendation.

Also, below are the details of the next 5 matches, India is going to play. You have to predict the result of the matches and if you are getting prediction as a Loss then suggest some changes and re-run your model again until you are getting Win as a prediction. You cannot use the same strategy in the entire series, because opponent will get to know your strategy and they can come with counter strategy. Hence for all the below 5 matches you have to suggest unique strategies to make India win. The suggestions should be in-line with the variables that have been mentioned in the given data set. Do consider the feasibility of the suggestions very carefully as well.

1 Test match with England in England. All the match are day matches. In England, it will be rainy season at the time to match.
2 T20 match with Australia in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.
2 ODI match with Sri Lanka in India. All the match are Day and Night matches. In India, it will be winter season at the time to match

NEED OF THE PROJECT AND BUSINESS UNDERSTANDING

Need of the project is to create an automated product which creates strategies and in future whichever team we are playing in against it will show the strategy. As soon as the match is decided with the opponent the product will tell the strategy. It will show the strategy based on the model built on past data we have. In future whichever team is in with similar characteristics we can predict the result

Most of the cricket bookies uses win prediction. Over 90 percent of the cricket betting tips goes on to win. The top bookies have a success rate of prediction over 80%. Over 81% of the bets we can recommend to cricket fans who turned out to be the winner rather than losing. Most of the cricket fans can refer to this URL and also make recommended predictions

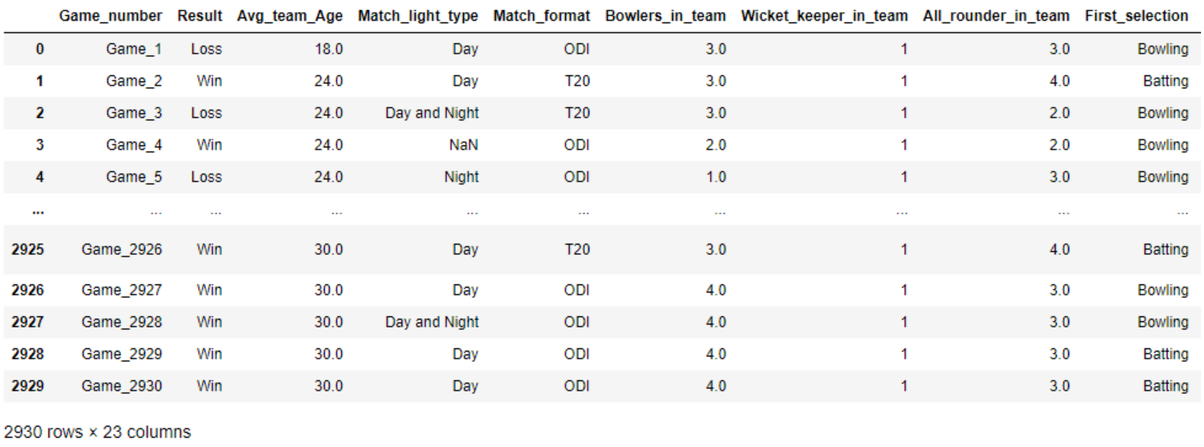
https://www.thetopbookies.com/betting-tips/cricket

The automated product we created can be sold on different platforms. For example, we can sell this product to BCCI so that it can benefit them by changing the characteristics accordingly in order to win the match. We can also sell this automated product on soccer platform by simply changing the characteristics of the data

2) EXPLORATORY DATA ANAYSIS AND BUSINESS IMPLICATIONS

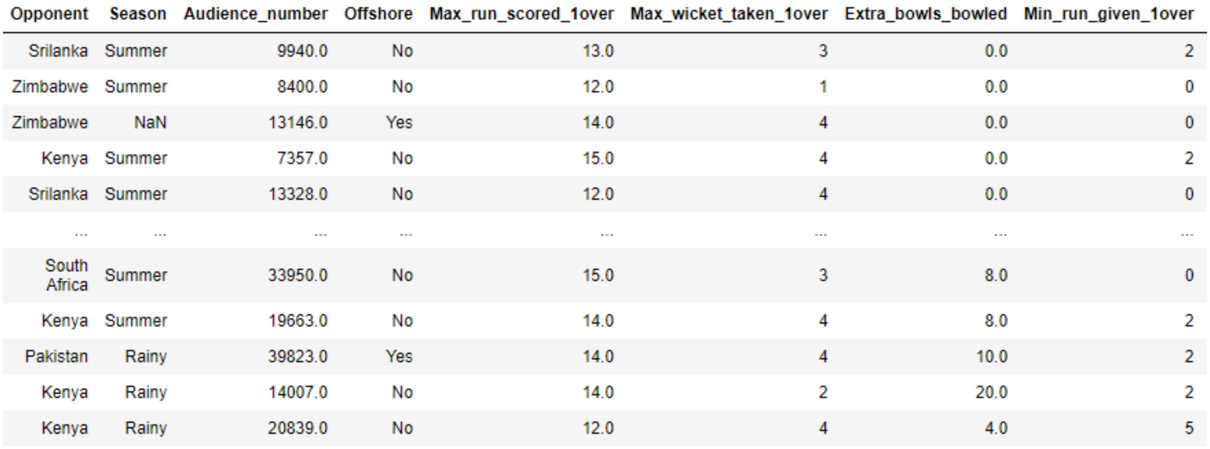
VISUAL INSPECTION OF THE DATA

HEAD OF THE DATA

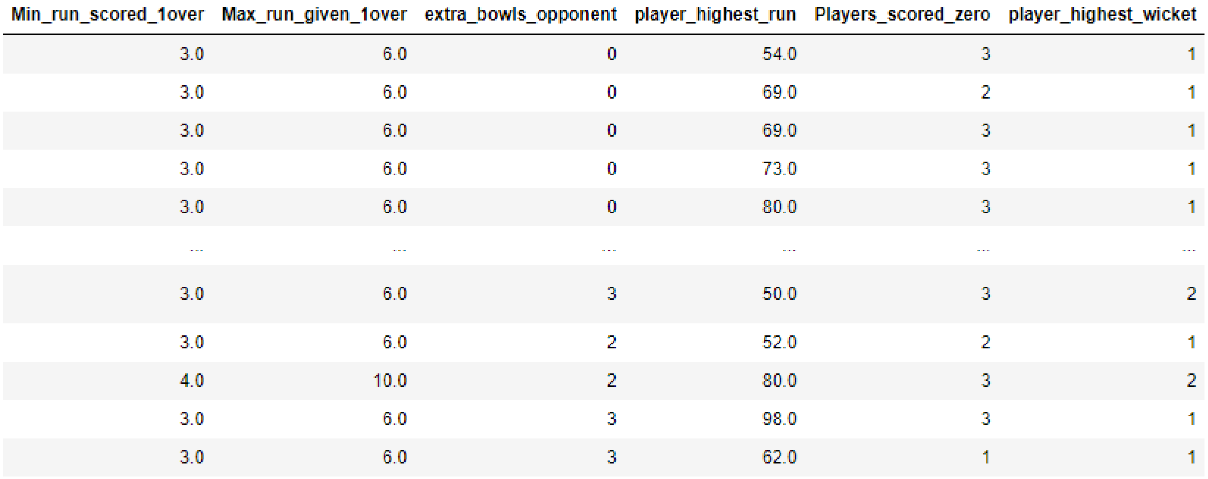


	Game_number	Result	Avg_team_Age	Match_light_type	Match_format	Bowlers_in_team	Wicket_keeper_in_team	All_rounder_in_team	First_selection
0	Game_1	Loss	18.0	Day	ODI	3.0	1	3.0	Bowling
1	Game_2	Win	24.0	Day	T20	3.0	1	4.0	Batting
2	Game_3	Loss	24.0	Day and Night	T20	3.0	1	2.0	Bowling
3	Game_4	Win	24.0	NaN	ODI	2.0	1	2.0	Bowling
4	Game_5	Loss	24.0	Night	ODI	1.0	1	3.0	Bowling
...	...	...	...	...	...	...	...	...	...
2925	Game_2926	Win	30.0	Day	T20	3.0	1	4.0	Batting
2926	Game_2927	Win	30.0	Day	ODI	4.0	1	3.0	Bowling
2927	Game_2928	Win	30.0	Day and Night	ODI	4.0	1	3.0	Bowling
2928	Game_2929	Win	30.0	Day	ODI	4.0	1	3.0	Batting
2929	Game_2930	Win	30.0	Day	ODI	4.0	1	3.0	Batting

2930 rows × 23 columns



Opponent	Season	Audience_number	Offshore	Max_run_scored_1over	Max_wicket_taken_1over	Extra_bowls_bowled	Min_run_given_1over
Srilanka	Summer	9940.0	No	13.0	3	0.0	2
Zimbabwe	Summer	8400.0	No	12.0	1	0.0	0
Zimbabwe	NaN	13146.0	Yes	14.0	4	0.0	0
Kenya	Summer	7357.0	No	15.0	4	0.0	2
Srilanka	Summer	13328.0	No	12.0	4	0.0	0
...	...	...	...	...	...	...	...
South Africa	Summer	33950.0	No	15.0	3	8.0	0
Kenya	Summer	19663.0	No	14.0	4	8.0	2
Pakistan	Rainy	39823.0	Yes	14.0	4	10.0	2
Kenya	Rainy	14007.0	No	14.0	2	20.0	2
Kenya	Rainy	20839.0	No	12.0	4	4.0	5



Min_run_scored_1over	Max_run_given_1over	extra_bowls_opponent	player_highest_run	Players_scored_zero	player_highest_wicket
3.0	6.0	0	54.0	3	1
3.0	6.0	0	69.0	2	1
3.0	6.0	0	69.0	3	1
3.0	6.0	0	73.0	3	1
3.0	6.0	0	80.0	3	1
...	...	...	...	...	...
3.0	6.0	3	50.0	3	2
3.0	6.0	2	52.0	2	1
4.0	10.0	2	80.0	3	2
3.0	6.0	3	98.0	3	1
3.0	6.0	3	62.0	1	1

DATA DICTIONARY

Variables	Description
Game_number	Unique ID for each match
Result	Final result of the match
Avg_team_Age	Average age of the playing 11 players for that match
Match_light_type	type of match: Day, night or day & night
Match_format	Format of the match: T20, ODI or test
Bowlers_in_team	how many full time bowlers has been player in the team
Wicket_keeper_in_team	how many full time wicket keeper has been player in the team
All_rounder_in_team	how many full time all-rounder has been player in the team
First_selection	First inning of team: batting or bowling
Opponent	Opponent team in the match
Season	What is the season of the city, where match has been played
Audience_number	Total number of audience in the stadium
Offshore	Match played within country or outside of the country
Max_run_scored_1over	Maximum run scored in 1 over by team
Max_wicket_taken_1over	Maximum wicket taken in 1 over by team
Extra_bowls_bowled	Total number of extras bowled by team
Min_run_given_1over	Minimum run given by the bowler in one over
Min_run_scored_1over	Minimum run scored in 1 over by team
Max_run_given_1over	Maximum run given by the bowler in one over
extra_bowls_opponent	Total number of extras bowled by opponent
player_highest_run	Highest score in the match by one player
Players_scored_zero	Number of player out on zero run
player_highest_wicket	Highest wickets taken by single player in match

SHAPE OF THE DATA

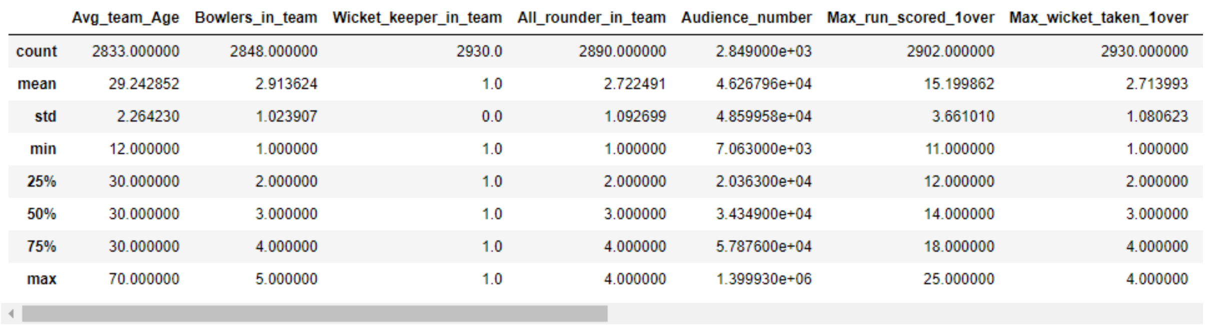
(2930, 23)

There are 2930 number of rows and 23 columns

DUPLICATES

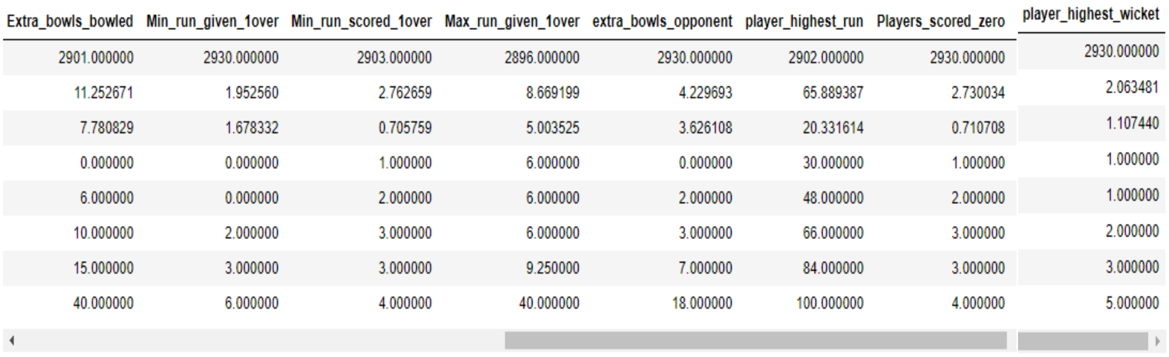
There are no duplicates entries

DESCRIPTION OF THE DATA



	Avg_team_Age	Bowlers_in_team	Wicket_keeper_in_team	All_rounder_in_team	Audience_number	Max_run_scored_1over	Max_wicket_taken_1over
count	2833.000000	2848.000000	2930.0	2890.000000	2.849000e+03	2902.000000	2930.000000
mean	29.242852	2.913624	1.0	2.722491	4.626796e+04	15.199862	2.713993
std	2.264230	1.023907	0.0	1.092699	4.859958e+04	3.661010	1.080623
min	12.000000	1.000000	1.0	1.000000	7.063000e+03	11.000000	1.000000
25%	30.000000	2.000000	1.0	2.000000	2.036300e+04	12.000000	2.000000
50%	30.000000	3.000000	1.0	3.000000	3.434900e+04	14.000000	3.000000
75%	30.000000	4.000000	1.0	4.000000	5.787600e+04	18.000000	4.000000
max	70.000000	5.000000	1.0	4.000000	1.399930e+06	25.000000	4.000000

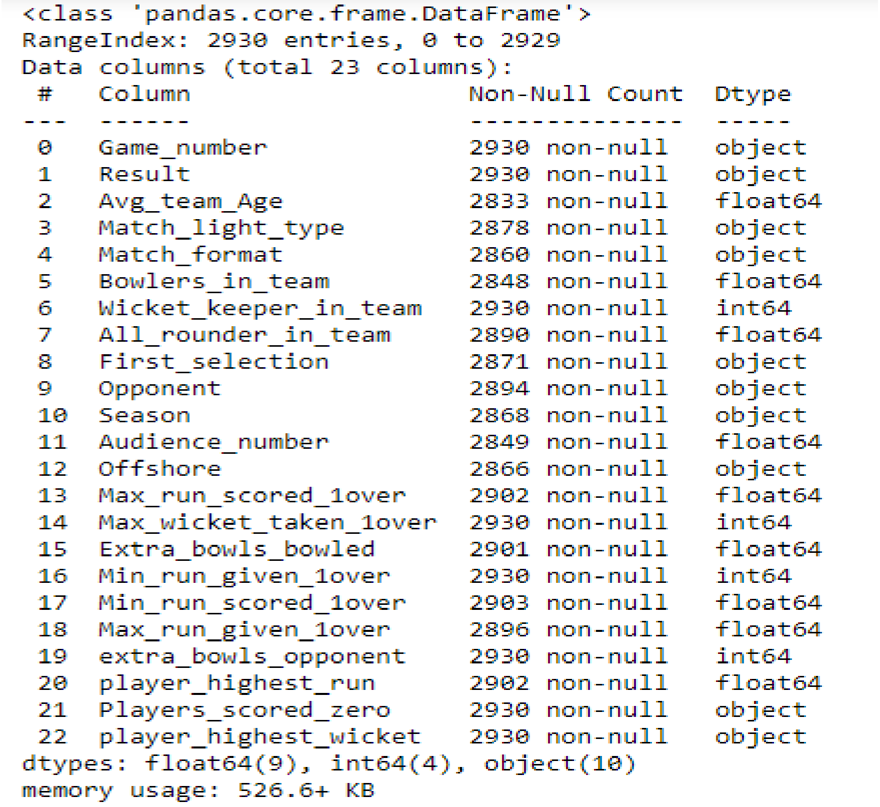
```
(8, 15)
```



Extra_bowls_bowled	Min_run_given_1over	Min_run_scored_1over	Max_run_given_1over	extra_bowls_opponent	player_highest_run	Players_scored_zero	player_highest_wicket
2901.000000	2930.000000	2903.000000	2896.000000	2930.000000	2902.000000	2930.000000	2930.000000
11.252671	1.952560	2.762659	8.669199	4.229693	65.889387	2.730034	2.063481
7.780829	1.678332	0.705759	5.003525	3.626108	20.331614	0.710708	1.107440
0.000000	0.000000	1.000000	6.000000	0.000000	30.000000	1.000000	1.000000
6.000000	0.000000	2.000000	6.000000	2.000000	48.000000	2.000000	1.000000
10.000000	2.000000	3.000000	6.000000	3.000000	66.000000	3.000000	2.000000
15.000000	3.000000	3.000000	9.250000	7.000000	84.000000	3.000000	3.000000
40.000000	6.000000	4.000000	40.000000	18.000000	100.000000	4.000000	5.000000

```
(8, 15)
```

UNDERSTANDING THE ATTRIBUTES



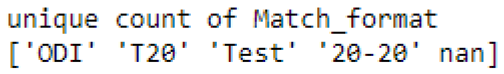
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2930 entries, 0 to 2929
Data columns (total 23 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Game_number            2930 non-null   object 
 1   Result                 2930 non-null   object 
 2   Avg_team_Age           2833 non-null   float64
 3   Match_light_type       2878 non-null   object 
 4   Match_format           2860 non-null   object 
 5   Bowlers_in_team        2848 non-null   float64
 6   Wicket_keeper_in_team  2930 non-null   int64  
 7   All_rounder_in_team    2890 non-null   float64
 8   First_selection        2871 non-null   object 
 9   Opponent               2894 non-null   object 
 10  Season                 2868 non-null   object 
 11  Audience_number        2849 non-null   float64
 12  Offshore               2866 non-null   object 
 13  Max_run_scored_1over   2902 non-null   float64
 14  Max_wicket_taken_1over 2930 non-null   int64  
 15  Extra_bowls_bowled     2901 non-null   float64
 16  Min_run_given_1over    2930 non-null   int64  
 17  Min_run_scored_1over   2903 non-null   float64
 18  Max_run_given_1over    2896 non-null   float64
 19  extra_bowls_opponent   2930 non-null   int64  
 20  player_highest_run     2902 non-null   float64
 21  Players_scored_zero    2930 non-null   object 
 22  player_highest_wicket  2930 non-null   object 
dtypes: float64(9), int64(4), object(10)
memory usage: 526.6+ KB
```

There are 10 variables having object data types

There are 13 variables having integer and float data types

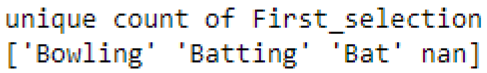
A) UNIVARIATE ANALYSIS

HYGIENE CHECK OF THE DATA



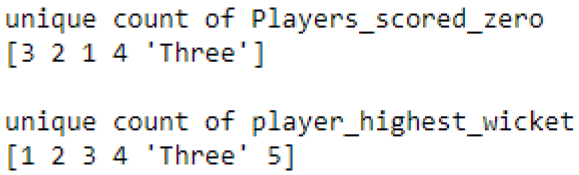
```
unique count of Match_format
['ODI' 'T20' 'Test' '20-20' nan]
```

We merge ’20-20’ with the ‘T20’



```
unique count of First_selection
['Bowling' 'Batting' 'Bat' nan]
```

We merge ‘Bat’ with the ‘Batting’

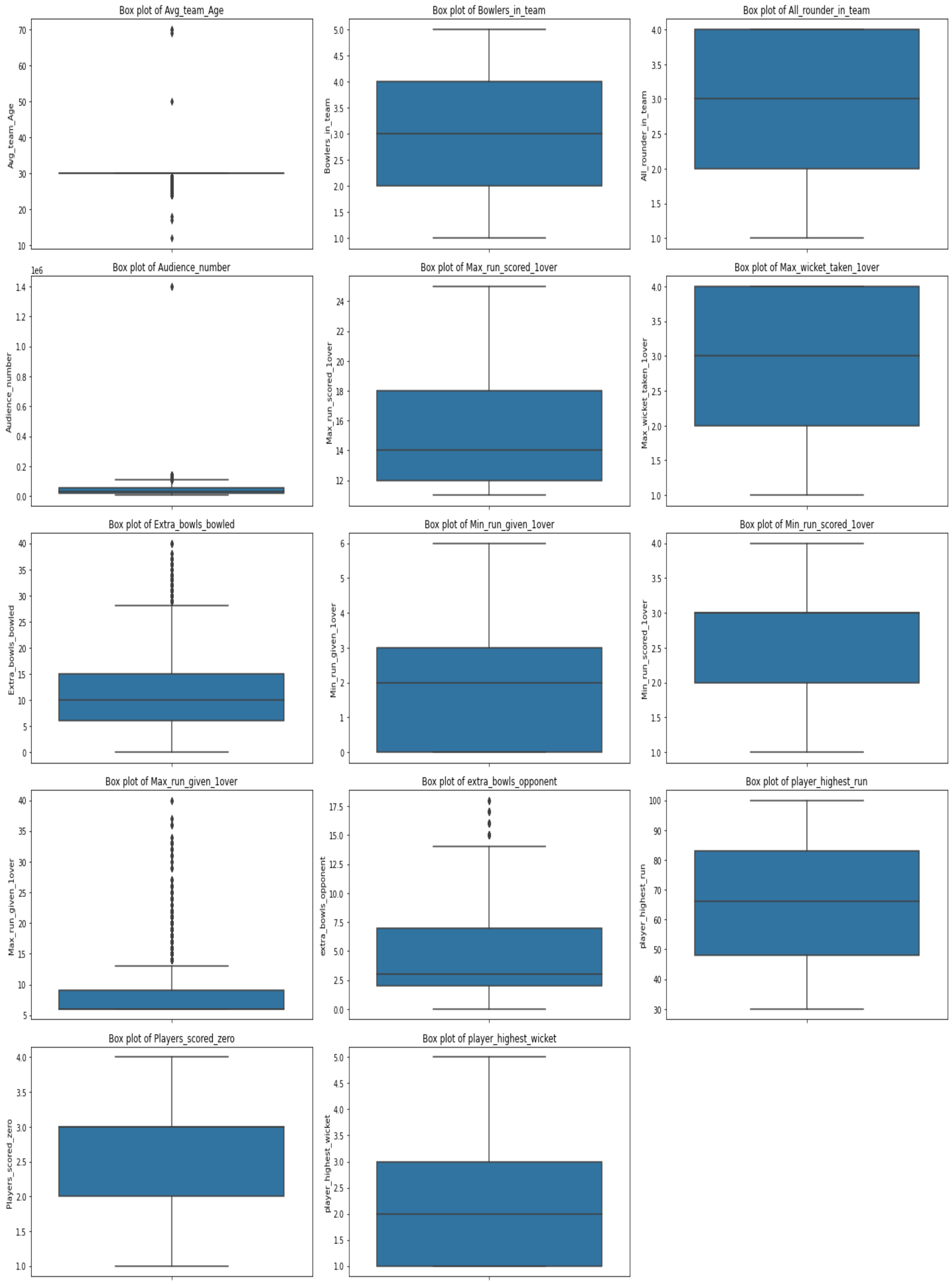


```
unique count of Players_scored_zero
[3 2 1 4 'Three']

unique count of player_highest_wicket
[1 2 3 4 'Three' 5]
```

We replace ‘three’ with the integer value 3 and convert the data type from object to numerical

figure 1: box plot of continuous variables



Variable **avg_team_age** is close to normal distribution as mean and median are similar though it has outliers

Variable **Bowlers_in_team** is normally distributed as mean and median are almost same, no outliers are present

Variable **All_rounder_in_team** is slightly left skewed as there is difference in mean and median and also 75 percentile and the maximum value are equal, no outliers are present

Variable **audience_number** is right skewed as mean is effected due to outliers present

Variable **Max_run_scored_1over** is slightly right skewed, no outliers are present

Variable **Max_wicket_taken_1over** is slightly left skewed as mean is less than median and also 75 percentile and the maximum value are same, no outliers are present

Variable **Extra_bowls_bowled** is right skewed as mean is higher than median, outliers are present

Variable **Min_run_given_1over** is close to normal and minimum value and 25 percentile are equal, no outliers are present

Variable **Min_run_scored_1over** is slightly left skewed, 50 and 75 percentile are equal, no outliers are present

Variable **Max_run_given_1over** is right skewed, minimum value,25,50 percentile have same values, outliers are present

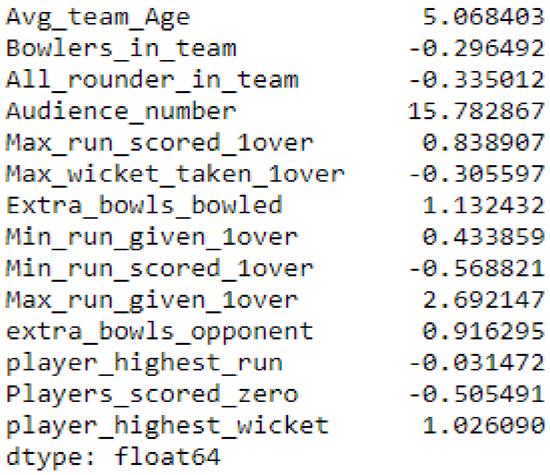
Variable **extra_bowls_opponent** is right skewed, outliers are present

Variable **player_highest_run** is normally distributed, no outliers are present

Variable **Players_scored_zero** is slightly left skewed, 50 and 75 percentile have the same value, no outliers are present

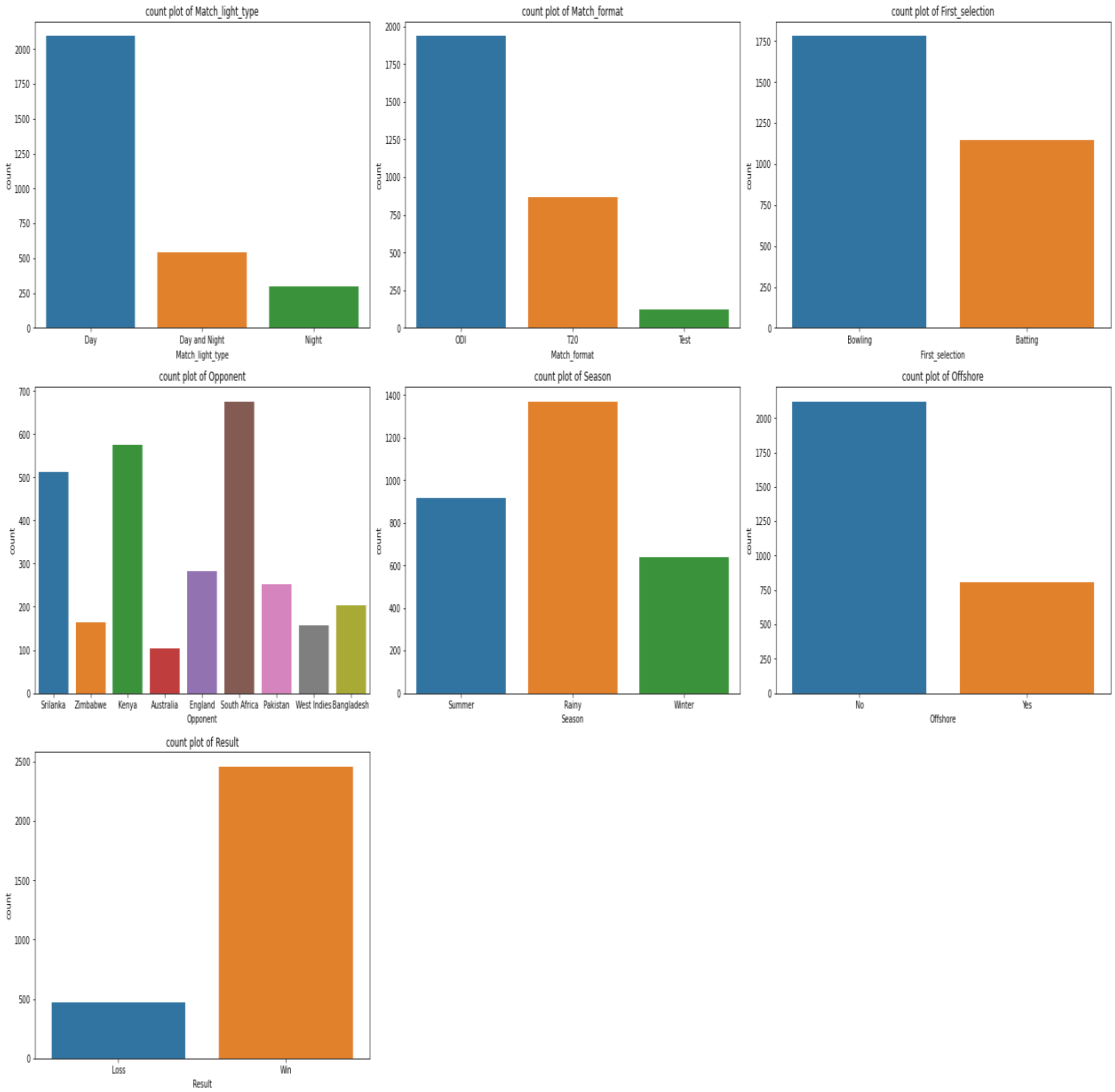
Variable **player_highest_wicket** is normally distributed, minimum value and 25 percentile are equal

Skewness



```
Avg_team_Age             5.068403
Bowlers_in_team         -0.296492
All_rounder_in_team     -0.335012
Audience_number         15.782867
Max_run_scored_1over     0.838907
Max_wicket_taken_1over  -0.305597
Extra_bowls_bowled       1.132432
Min_run_given_1over      0.433859
Min_run_scored_1over    -0.568821
Max_run_given_1over      2.692147
extra_bowls_opponent     0.916295
player_highest_run      -0.031472
Players_scored_zero     -0.505491
player_highest_wicket    1.026090
dtype: float64
```

figure 2: count plot of categorical variables



Match_light_type: The number of matches played during the day is the highest and the lowest matches played during the night

Match_format: The number of ODI matches played is the highest and the lowest played is for the test format

First selection: The highest number of the matches played with bowling first and lowest with the batting as the first selection

Opponent: highest number of matches are played against south Africa and lowest matches played against Australia

Season: highest number of matches are played during rainy season and lowest matches are played during winter season

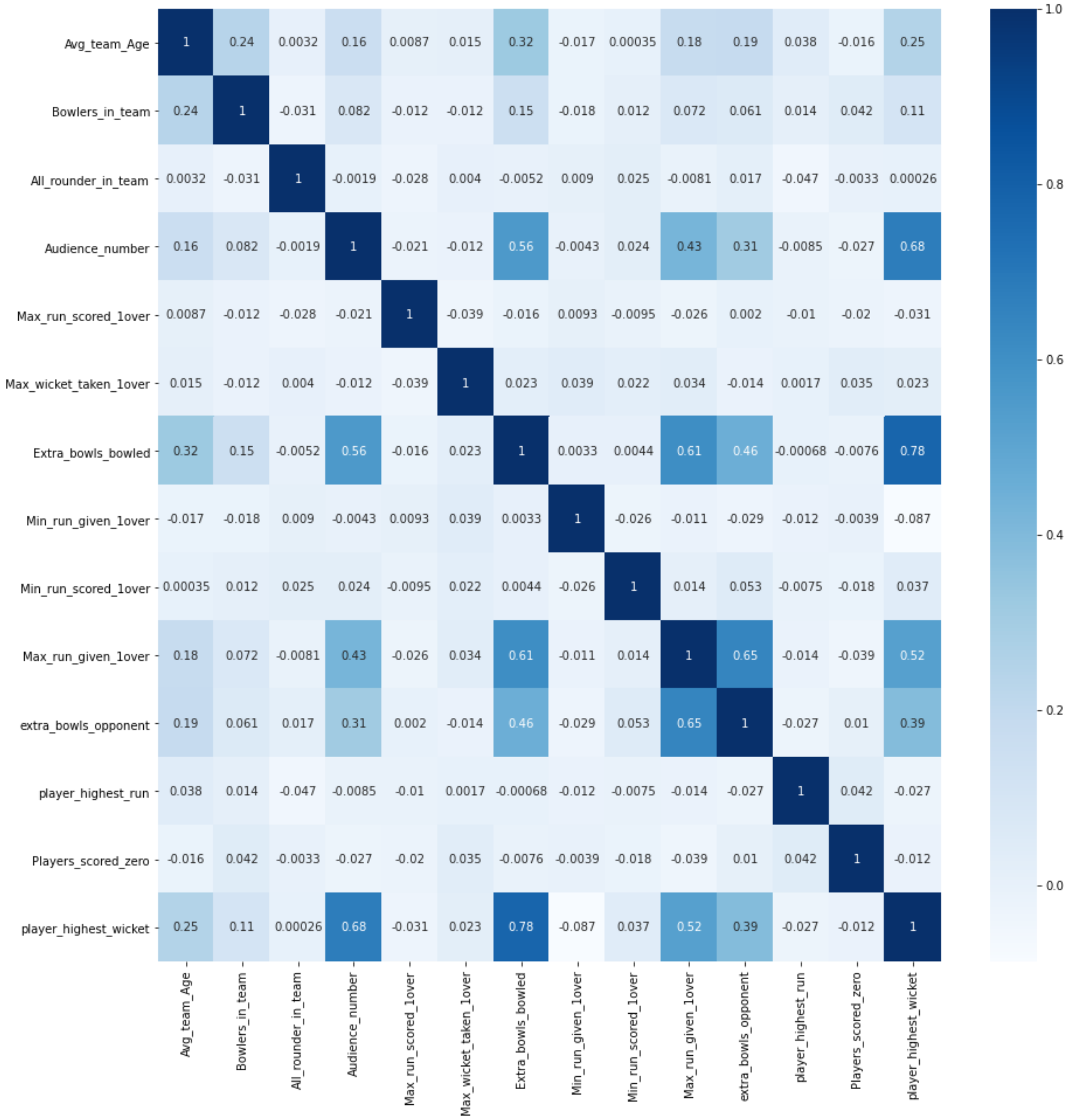
Offshore: Less matches are played offshore, most of the matches are played on the home ground

Result: Among all the matches most of the matches are won, less matches are lost

B) BIVARIATE ANALYSIS

CORRELATION PLOT

FIGURE 3: CORRELATION PLOT



Variables **extra bowls bowled** and **Players_highest_wicket** has the highest correlation 0.78.

Variables **Audience number** and **Players_highest_wicket** has the strong correlation 0.68.

Variables **Maximum runs given in one over** and **extra bowls opponents** has the correlation of 0.65

Variables **Maximum runs given in one over** and **extra bowls bowled** has the correlation of 0.61

FIGURE 4: PAIRPLOT

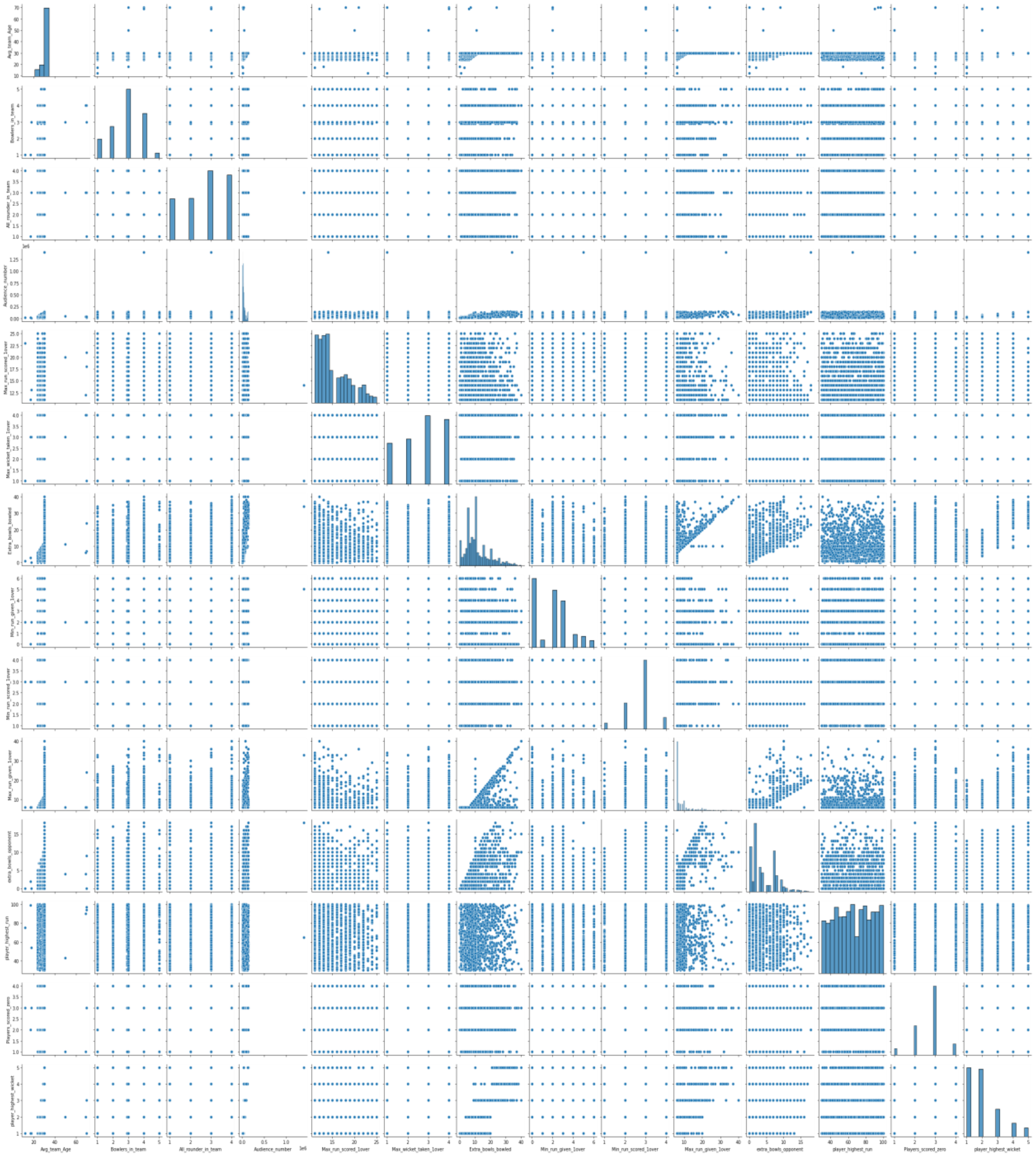
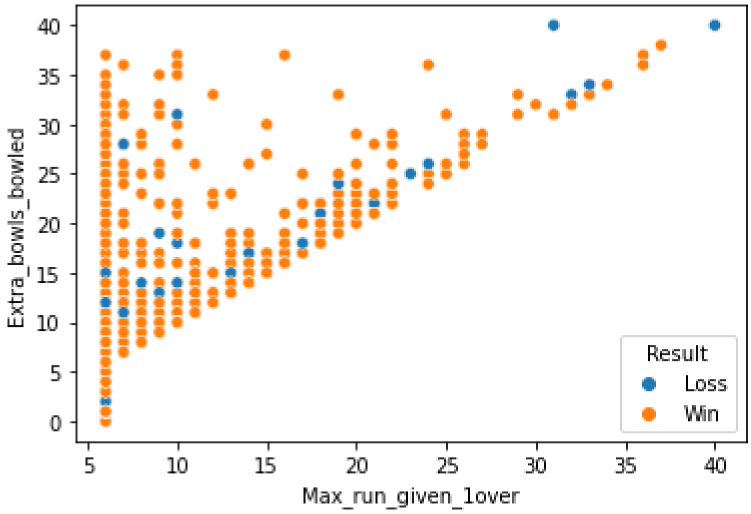
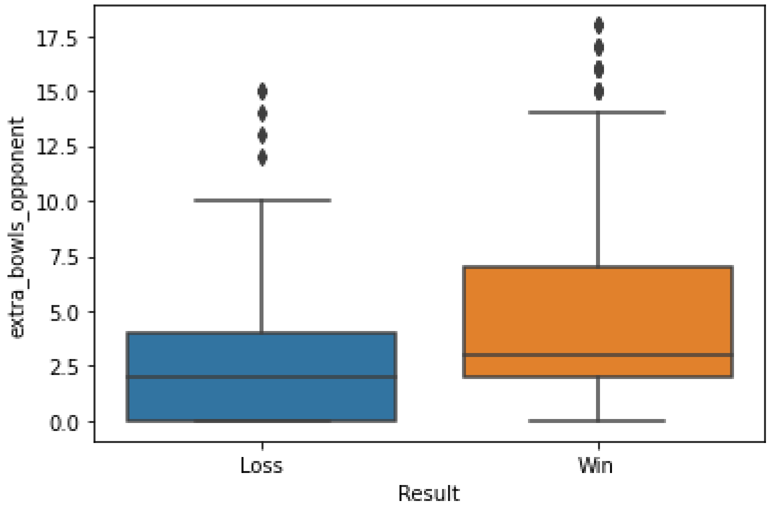


FIGURE 5: Scatter plot



It shows the linear relationship between maximum run given in one over and extra bowls bowled. Highest value of both the variables result in loss

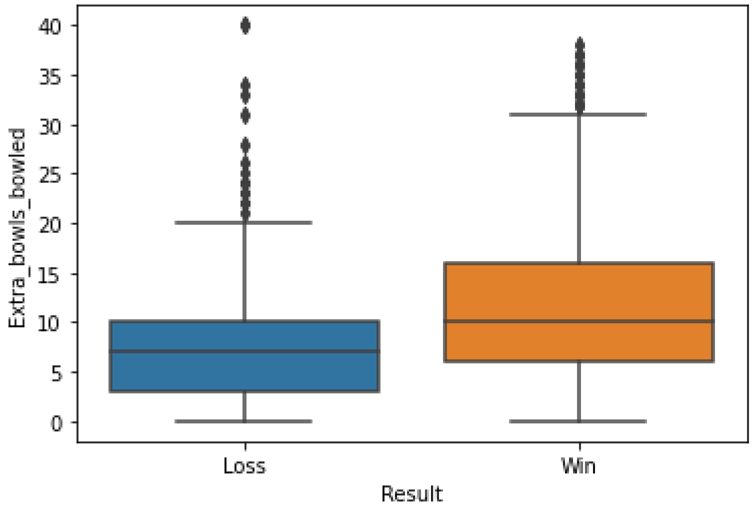
figure 6: box plot



The median of extra bowls opponents is higher for win, If the extra bowls opponents is higher than 16 then India will win the match as per the data set

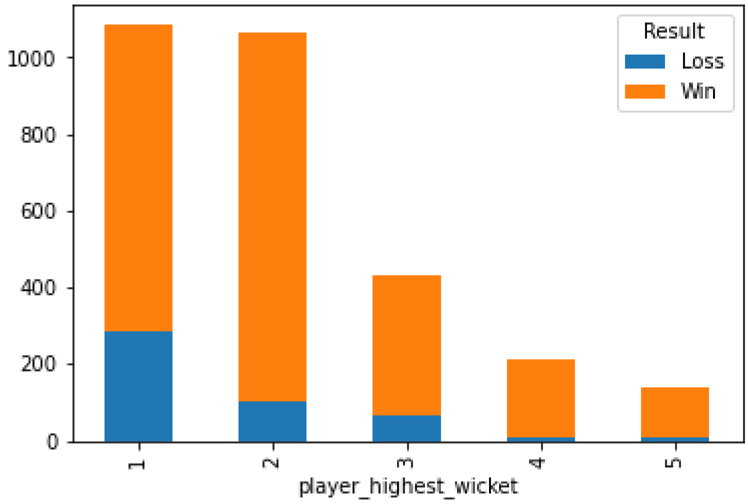
if the extra bowls opponents are higher than 10 then chances of winning the match is more

figure 7: Box plot



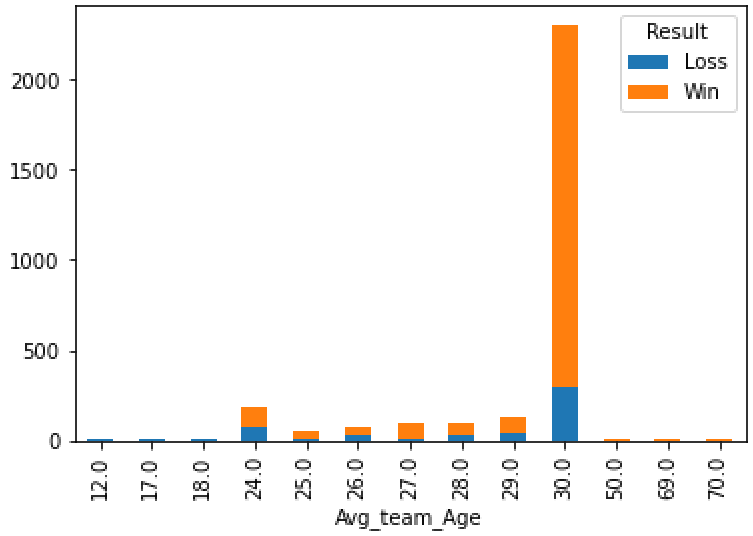
As per data when you bowl 40 extra bowls India is definitely will lose the match

figure 8: bar plot



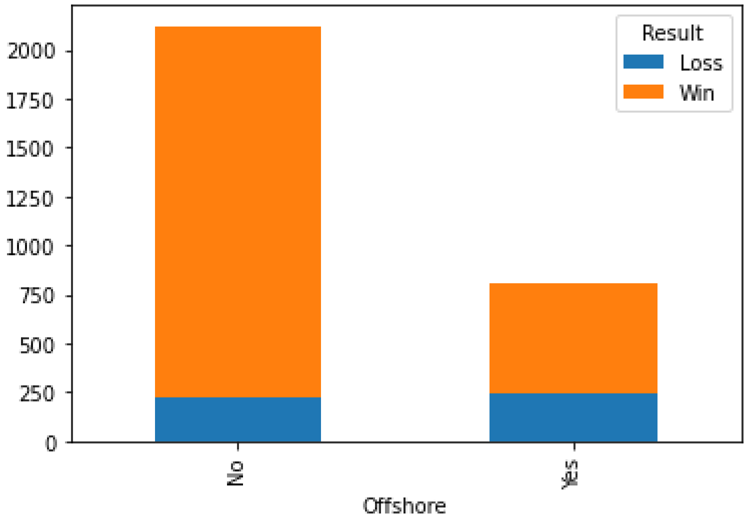
The chances of winning the match is high when the wicket taken by a single player is 2

figure 9: bar plot



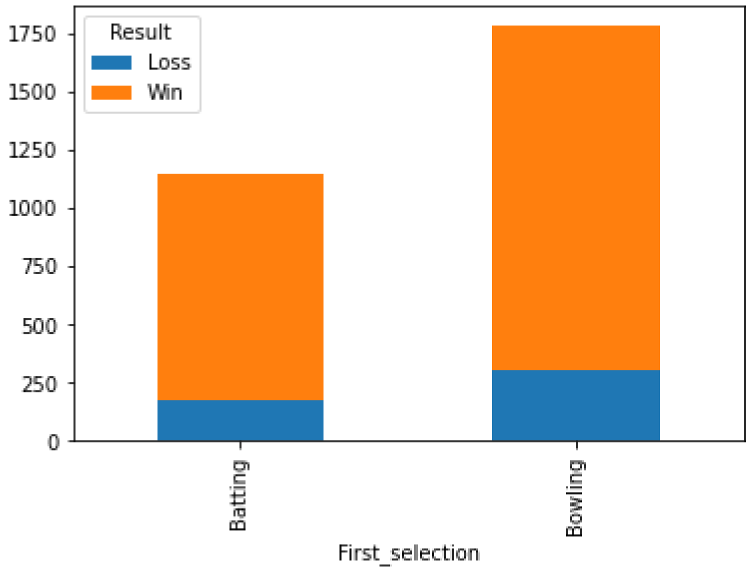
India has the highest win with average team age at 30

figure 10: Bar plot



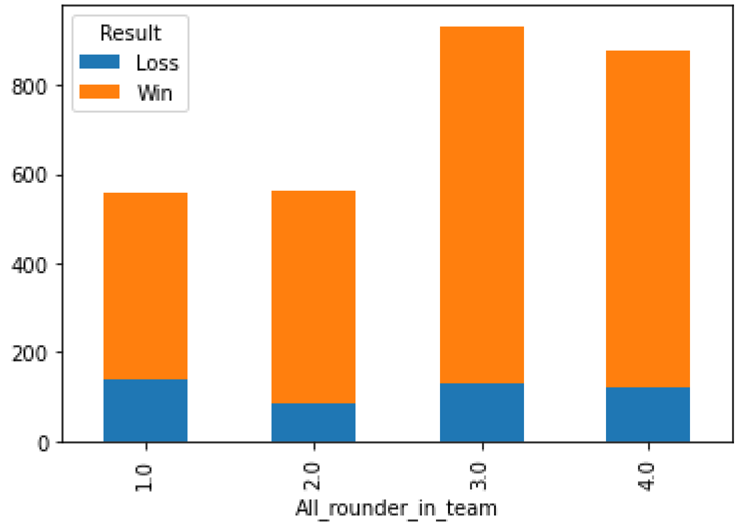
The performance of the Indian team is good when played on the home ground

figure 11: bar plot

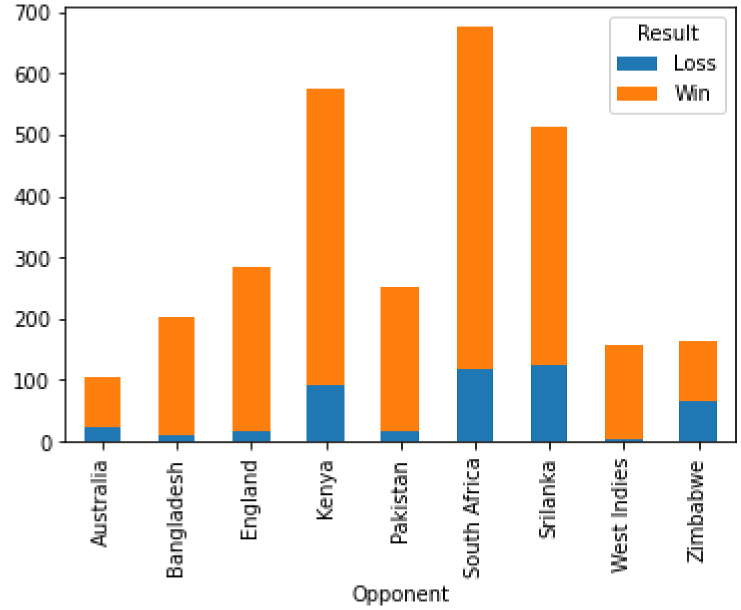


Most of the matches are won when selected to bowl first

figure 12: Bar plot



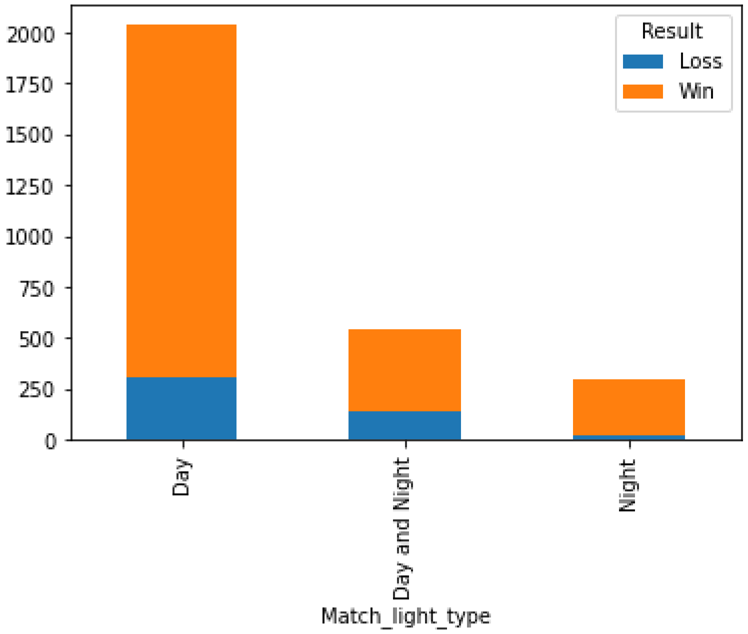
Winning percentage is higher when matches are played with 3 to 4 all-rounders in the team figure 13: bar plot



India team is performing well against west indies, Bangladesh, England and Pakistan as the winning rate is high

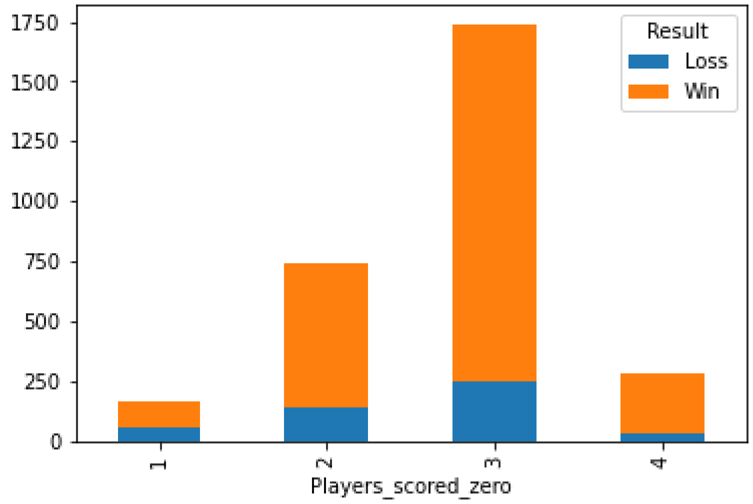
India team wining rate is less against south Africa, srilanka, Zimbabwe and Australia

figure 14: bar plot



Indian team has a good performance in ODI format as compared to both the formats

Winning rate is higher in ODI format and lesser in T20 and test figure 15: bar plot



Indian team has the highest wining rate when there are 3 player scored zero and lowest when there is one player scored zero

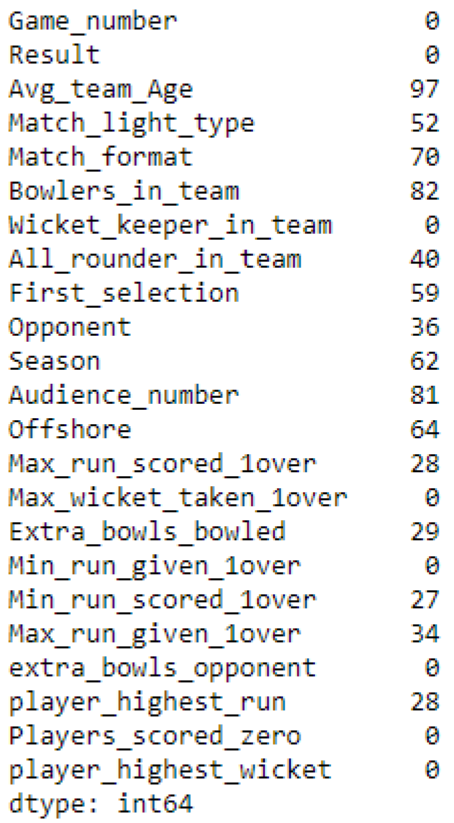
Business insights from analysis

- Variables maximum runs given in one over and extra bowls bowled have a good relation which results in win or loss of the match. More number of runs given in one over and extra bowls bowled chances of losing is high
- Variable extra bowls opponent can add value to the prediction. More extra bowls opponent more is chance of the winning
- Variable offshore also predict the win. If the number of matches played on home ground are more chances of winning is high
- Variable first selection also impact on the result. If the first selection is bowling the chances of winning is high
- Variable all-rounder in team has impact on the result. Having 3 to 4 all-rounders in the team may result into win

3) DATA CLEANING AND PREPROCESSING

Approach used for identifying and treating missing values and outliers

Approach used for identifying missing values is is null condition



```
Game_number               0
Result                    0
Avg_team_Age             97
Match_light_type         52
Match_format             70
Bowlers_in_team          82
Wicket_keeper_in_team     0
All_rounder_in_team      40
First_selection          59
Opponent                 36
Season                   62
Audience_number          81
Offshore                 64
Max_run_scored_1over     28
Max_wicket_taken_1over    0
Extra_bowls_bowled       29
Min_run_given_1over       0
Min_run_scored_1over     27
Max_run_given_1over      34
extra_bowls_opponent      0
player_highest_run       28
Players_scored_zero       0
player_highest_wicket     0
dtype: int64
```

For numerical variable

Variable **Avg_team_Age** has 97 missing values and is replaced by the median value as the variable is normally distributed mean and median are same

Variable **Bowlers_in_team** has 82 missing value and is replaced by the mean value as the variable is slightly left skewed

Variable **All_rounder_in_team** has 40 missing value and is replaced by the median value as the variable is slightly left skewed

Variable **Audience_number** has 81 missing value and is replaced by the median value as the variable is right skewed

Variable **Max_run_scored_1over** has 28 missing values and is replaced by the median value

Variable **Extra_bowls_bowled** has 29 missing value and is replaced by the median value

Variable **Min_run_scored1over** has 27 missing value and is replaced by the median value

Variable **Max_run_given_1over** has 34 missing values and is replaced by the median value

Variable **player_highest_run** has 28 missing values and is replaced by the median value

For categorical variable

Variable **Match_light_type** has 52 missing values and it is replaced by the mode value the most occurring value

Variable **Match_format** has 70 missing values and it is replaced by the mode

Variable **First_selection** has 59 missing values and it is replaced by the mode

Variable **Opponent** has 36 missing values and it is replaced by the mode

Variable **Season** has 62 missing values and it is replaced by the mode

Variable **Offshore** has 64 missing values and it is replaced by the mode

Outliers treatment

Outliers have effect on mean. It increases the mean. If it is a distance based calculations, then the modelling may be effected. So it is necessary to treat outliers

Box plot is used to find the outliers

Variables **avg_team_age, audience number, extra bowls bowled, maximum run given in one over, extra bowls opponent** have outliers.

Inter quantile range (IQR) method is used to treat the outliers

IQR = Q3 – Q1

Q1 is the first quantile which is 25 percentile

Q3 is the third quantile which is 75 percentile

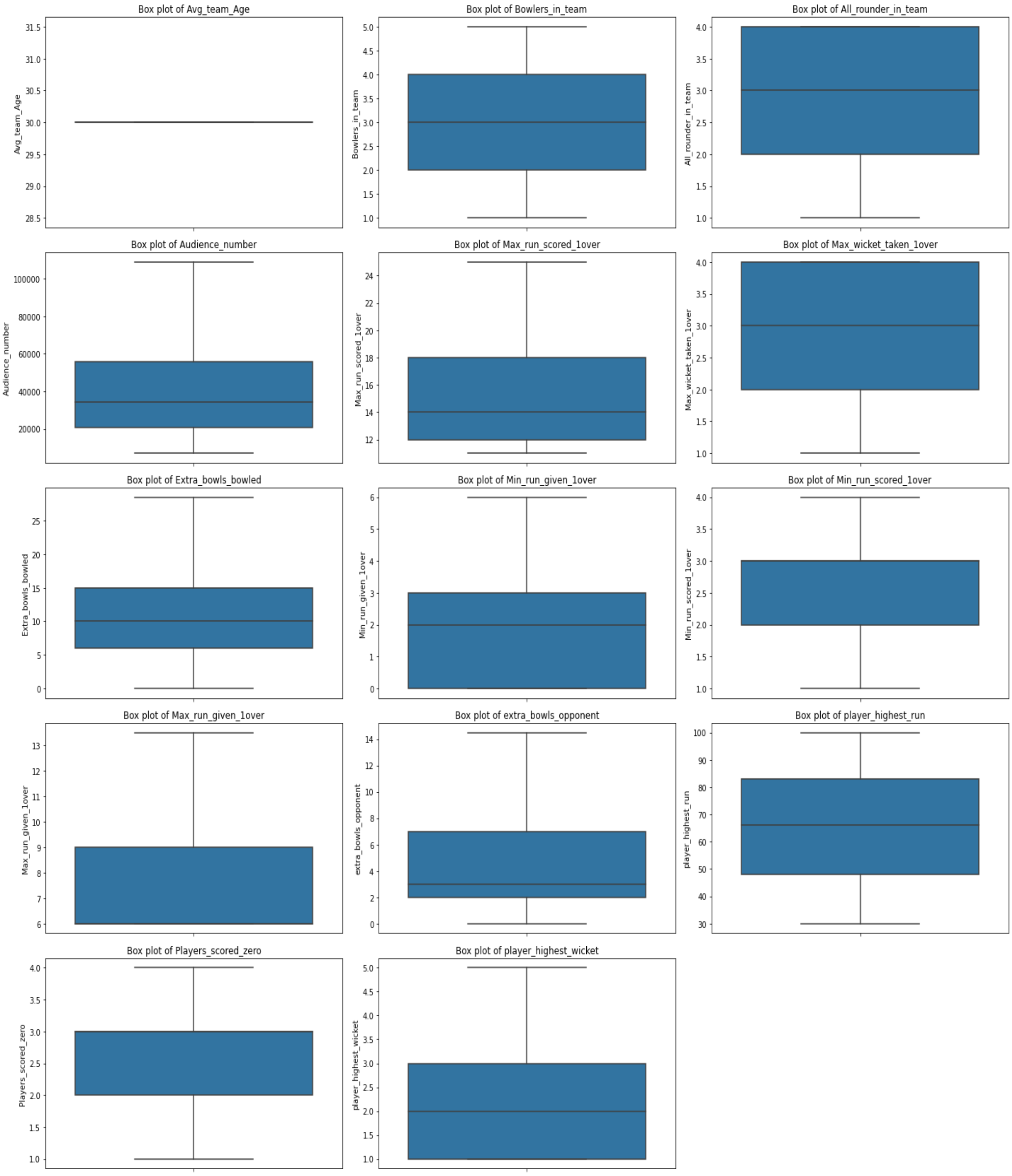
Lower limit = Q1 – 1.5*IQR

Upper limit = Q3 + 1.5*IQR

Outliers are capped to the upper and lower limit

After treating outliers,

figure 16: box plot after outlier treatment



Box plot of Avg_team_Age

Box plot of Bowlers_in_team

Box plot of All_rounder_in_team

Box plot of Audience_number

Box plot of Max_run_scored_1over

Box plot of Max_wicket_taken_1over

Box plot of Extra_bowls_bowled

Box plot of Min_run_given_1over

Box plot of Min_run_scored_1over

Box plot of Max_run_given_1over

Box plot of extra_bowls_opponent

Box plot of player_highest_run

Box plot of Players_scored_zero

Box plot of player_highest_wicket

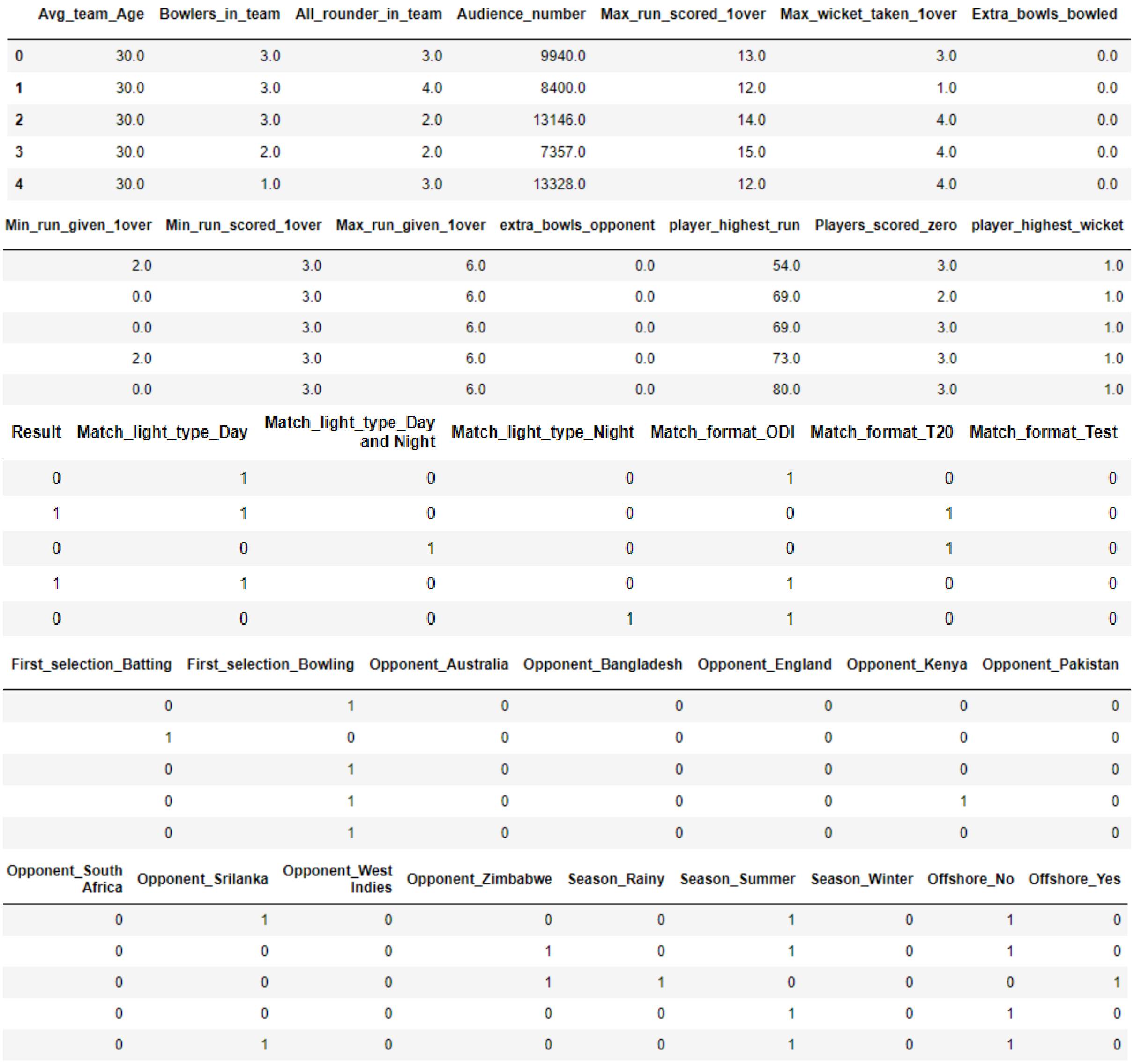
Variable transformation

Variable player scored zero and player highest wicket was an object data type it is convert into integer data type

Variable game number and wicket keeper in team has been as it does not add any value to the analysis

4) Model building

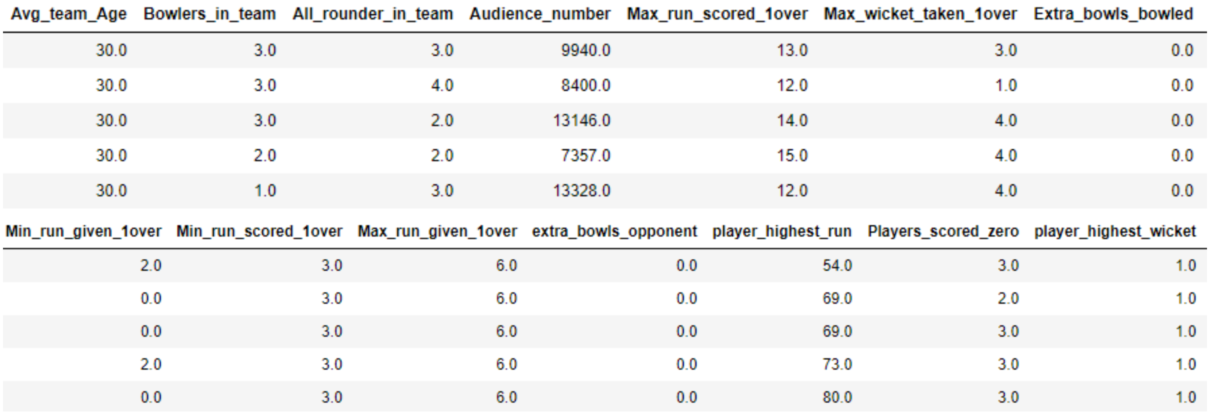
Data using one hot encoding



	Avg_team_Age	Bowlers_in_team	All_rounder_in_team	Audience_number	Max_run_scored_1over	Max_wicket_taken_1over	Extra_bowls_bowled	Min_run_given_1over	Min_run_scored_1over	Max_run_given_1over	extra_bowls_opponent	player_highest_run	Players_scored_zero	player_highest_wicket	Result	Match_light_type_Day	Match_light_type_Day and Night	Match_light_type_Night	Match_format_ODI	Match_format_T20	Match_format_Test	First_selection_Batting	First_selection_Bowling	Opponent_Australia	Opponent_Bangladesh	Opponent_England	Opponent_Kenya	Opponent_Pakistan	Opponent_South Africa	Opponent_Srilanka	Opponent_West Indies	Opponent_Zimbabwe	Season_Rainy	Season_Summer	Season_Winter	Offshore_No	Offshore_Yes
0	30.0	3.0	3.0	9940.0	13.0	3.0	0.0	2.0	3.0	6.0	0.0	54.0	3.0	1.0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0
1	30.0	3.0	4.0	8400.0	12.0	1.0	0.0	0.0	3.0	6.0	0.0	69.0	2.0	1.0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0
2	30.0	3.0	2.0	13146.0	14.0	4.0	0.0	0.0	3.0	6.0	0.0	69.0	3.0	1.0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1
3	30.0	2.0	2.0	7357.0	15.0	4.0	0.0	2.0	3.0	6.0	0.0	73.0	3.0	1.0	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0
4	30.0	1.0	3.0	13328.0	12.0	4.0	0.0	0.0	3.0	6.0	0.0	80.0	3.0	1.0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0

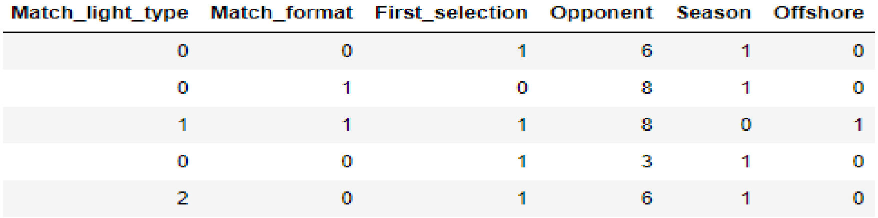
shape of the data (2930,37)

Data using label encoding



Avg_team_Age	Bowlers_in_team	All_rounder_in_team	Audience_number	Max_run_scored_1over	Max_wicket_taken_1over	Extra_bowls_bowled
30.0	3.0	3.0	9940.0	13.0	3.0	0.0
30.0	3.0	4.0	8400.0	12.0	1.0	0.0
30.0	3.0	2.0	13146.0	14.0	4.0	0.0
30.0	2.0	2.0	7357.0	15.0	4.0	0.0
30.0	1.0	3.0	13328.0	12.0	4.0	0.0

Min_run_given_1over	Min_run_scored_1over	Max_run_given_1over	extra_bowls_opponent	player_highest_run	Players_scored_zero	player_highest_wicket
2.0	3.0	6.0	0.0	54.0	3.0	1.0
0.0	3.0	6.0	0.0	69.0	2.0	1.0
0.0	3.0	6.0	0.0	69.0	3.0	1.0
2.0	3.0	6.0	0.0	73.0	3.0	1.0
0.0	3.0	6.0	0.0	80.0	3.0	1.0



Match_light_type	Match_format	First_selection	Opponent	Season	Offshore
0	0	1	6	1	0
0	1	0	8	1	0
1	1	1	8	0	1
0	0	1	3	1	0
2	0	1	6	1	0

Shape of the data (2930,21)

We split the result with all independent variables

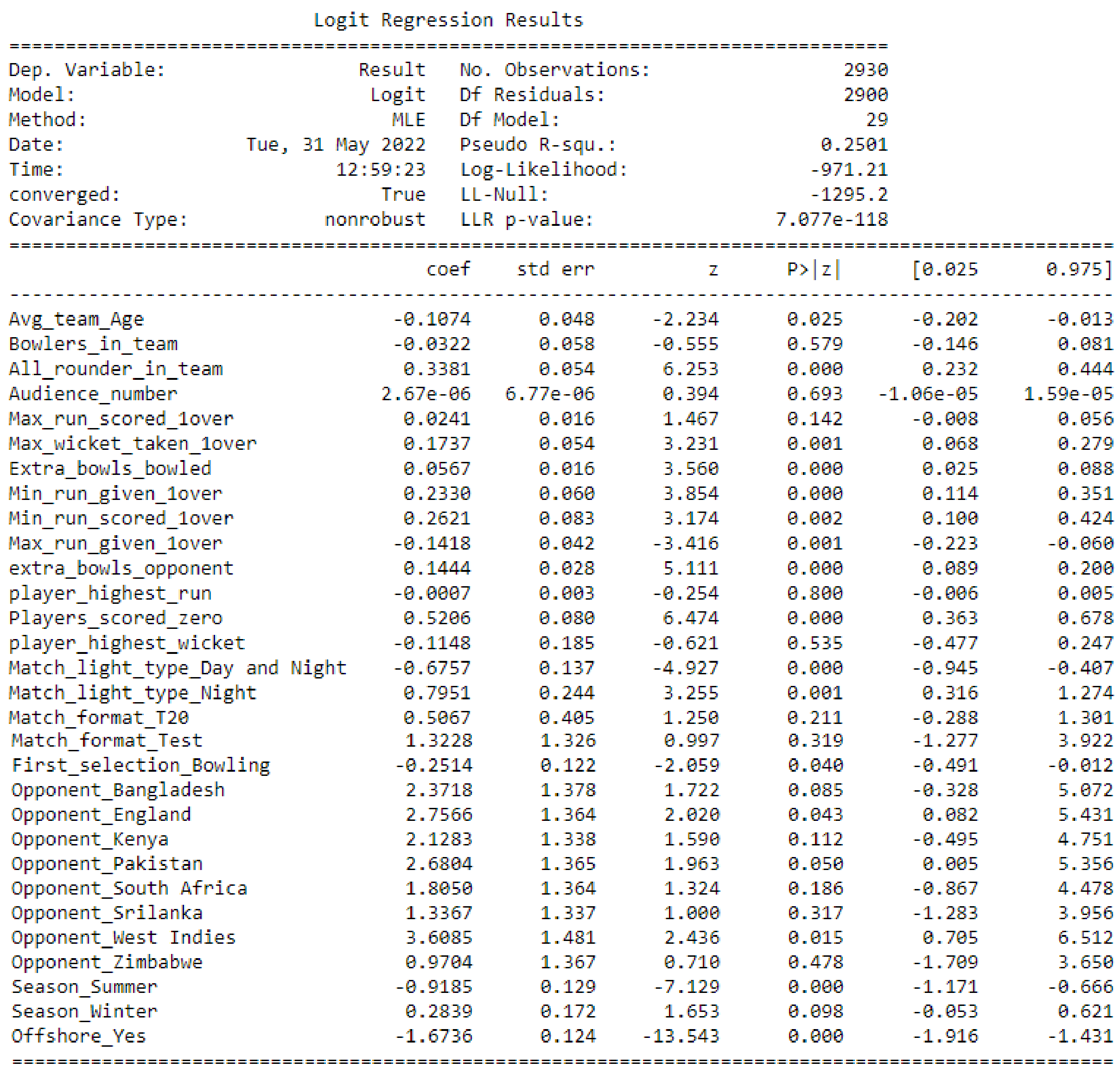
We then split the data into 70 and 30. 70 for training and 30 for testing using train test split

Building various model

1 logistic regression using stats model

For logistics regression one hot encoding is done

Logistic regression model internally uses linear equation to find the intercept and coefficient and then it is converted to the classes using activation function. It uses sigmoid curve to calculate the probability depending on the defined threshold. Any value greater than threshold will be considered as 1 and the value less than threshold will be considered as 0. Threshold value is usually 0.5 and it can be adjusted accordingly. It uses log of odds to convert into the probability. Log of odds is the linear equation having intercept and coefficient.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                 Result   No. Observations:                 2930
Model:                          Logit   Df Residuals:                     2900
Method:                           MLE   Df Model:                           29
Date:                Tue, 31 May 2022   Pseudo R-squ.:                  0.2501
Time:                        12:59:23   Log-Likelihood:                -971.21
converged:                       True   LL-Null:                       -1295.2
Covariance Type:            nonrobust   LLR p-value:                 7.077e-118
==================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Avg_team_Age                    -0.1074      0.048     -2.234      0.025      -0.202      -0.013
Bowlers_in_team                 -0.0322      0.058     -0.555      0.579      -0.146       0.081
All_rounder_in_team              0.3381      0.054      6.253      0.000       0.232       0.444
Audience_number                2.67e-06   6.77e-06      0.394      0.693   -1.06e-05    1.59e-05
Max_run_scored_1over             0.0241      0.016      1.467      0.142      -0.008       0.056
Max_wicket_taken_1over           0.1737      0.054      3.231      0.001       0.068       0.279
Extra_bowls_bowled               0.0567      0.016      3.560      0.000       0.025       0.088
Min_run_given_1over              0.2330      0.060      3.854      0.000       0.114       0.351
Min_run_scored_1over             0.2621      0.083      3.174      0.002       0.100       0.424
Max_run_given_1over             -0.1418      0.042     -3.416      0.001      -0.223      -0.060
extra_bowls_opponent             0.1444      0.028      5.111      0.000       0.089       0.200
player_highest_run              -0.0007      0.003     -0.254      0.800      -0.006       0.005
Players_scored_zero              0.5206      0.080      6.474      0.000       0.363       0.678
player_highest_wicket           -0.1148      0.185     -0.621      0.535      -0.477       0.247
Match_light_type_Day and Night  -0.6757      0.137     -4.927      0.000      -0.945      -0.407
Match_light_type_Night           0.7951      0.244      3.255      0.001       0.316       1.274
Match_format_T20                 0.5067      0.405      1.250      0.211      -0.288       1.301
Match_format_Test                1.3228      1.326      0.997      0.319      -1.277       3.922
First_selection_Bowling         -0.2514      0.122     -2.059      0.040      -0.491      -0.012
Opponent_Bangladesh              2.3718      1.378      1.722      0.085      -0.328       5.072
Opponent_England                 2.7566      1.364      2.020      0.043       0.082       5.431
Opponent_Kenya                   2.1283      1.338      1.590      0.112      -0.495       4.751
Opponent_Pakistan                2.6804      1.365      1.963      0.050       0.005       5.356
Opponent_South Africa            1.8050      1.364      1.324      0.186      -0.867       4.478
Opponent_Srilanka                1.3367      1.337      1.000      0.317      -1.283       3.956
Opponent_West Indies             3.6085      1.481      2.436      0.015       0.705       6.512
Opponent_Zimbabwe                0.9704      1.367      0.710      0.478      -1.709       3.650
Season_Summer                   -0.9185      0.129     -7.129      0.000      -1.171      -0.666
Season_Winter                    0.2839      0.172      1.653      0.098      -0.053       0.621
Offshore_Yes                    -1.6736      0.124    -13.543      0.000      -1.916      -1.431
==================================================================================================
```

Significant variables are whose p value is less than alpha which is 0.05
Avg_team_Age
All_rounder_in_team
Max_wicket_taken_1over
Extra_bowls_bowled
Min_run_given_1over
Min_run_scored_1over
Max_run_given_1over
Extra_bowls_opponent
Players_scored_zero
Match_light_type_Day and Night
Match_light_type_Night
Season_Summer
Offshore_yes

2 Decision tree classifier (CART)

For decision tree we use label encoding

Parameters used for building model

Criterion = ‘gini’

Max_depth = ‘none’

Min_sample_leaf = ‘1’

Min_sample_split = ‘2’

Max_features = None

3 Random forest classifier

Making random forest model using random forest classifier with parameters estimator which is number of trees, max_features which is the number of variables, min_sample_leaf is the minimum sample of the terminal node, minimum_sample_split is the minimum split of the parent node and fitting the data in the model.

Parameters used for model building

n_estimators = 100

max_features=7

random_state =0

4 Artificial neural network

To build artificial neural networks first we need to scale both training and testing data
By using MLP classifier we pass the parameters
Hidden_layer_sizes: it consists of number of perceptron’s and each perceptron are connected to the inputs.
max_iter: weights are adjusted at every iteration and the losses are reduced.
Solver: solver gradient method is used for calculation of weights.
Tolerance: tolerance of losses at every iteration.

Parameters used for model building

hidden_layer_sizes=100,

max_iter=2500,

solver='sgd',

verbose=True,

random_state=0,

tol=0.01

Efforts to improve the model performance

Logistic regression using stats model

Taking threshold as 0.5

If the value is greater than 0.5 it is will be 1 and less than 0.5 it will 0

To increase the precision value, we need to decrease the false positive cases we can do by shifting the threshold point to the right

In this case precision value should be high, so we reduce the false positive rate by shifting the threshold value from 0.5 to 0.6

Taking threshold as 0.6

If the value is greater than 0.6 it is will be 1 and less than 0.6 it will 0

We can further increase the precision value we shifting the threshold from 0.6 to 0.7

Taking threshold as 0.7

We can consider threshold point as 0.7 where we get the maximum precision value which is 0.91 and lower false positive cases

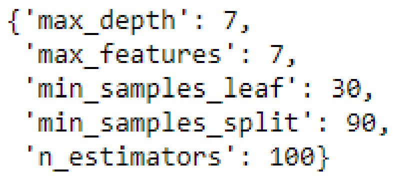
We can use this threshold point for predictions

Ensemble modeling

Random forest model uses ensemble technique

Random forest using grid search. Grid search cross validation is a technique to find the best combination of parameters

Getting the best parameters using grid search



```
{'max_depth': 7,
 'max_features': 7,
 'min_samples_leaf': 30,
 'min_samples_split': 90,
 'n_estimators': 100}
```

Decision tree model (CART) using regularisation

Parameters used in model building

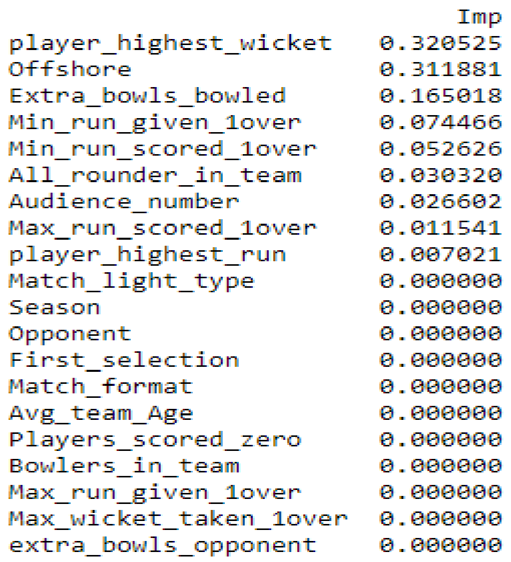
criterion='gini'

max_depth=10

min_samples_leaf=30

min_samples_split =300

Feature importance



```
                          Imp
player_highest_wicket   0.320525
Offshore                0.311881
Extra_bowls_bowled      0.165018
Min_run_given_1over     0.074466
Min_run_scored_1over    0.052626
All_rounder_in_team     0.030320
Audience_number         0.026602
Max_run_scored_1over    0.011541
player_highest_run      0.007021
Match_light_type        0.000000
Season                  0.000000
Opponent                0.000000
First_selection         0.000000
Match_format            0.000000
Avg_team_Age            0.000000
Players_scored_zero     0.000000
Bowlers_in_team         0.000000
Max_run_given_1over     0.000000
Max_wicket_taken_1over  0.000000
extra_bowls_opponent    0.000000
```

Artificial neural network using grid search

After using grid search cross validation, we get best parameters

{'hidden_layer_sizes': 50, 'max_iter': 2500, 'solver': 'sgd', 'tol': 0.01}

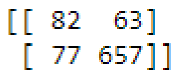
5) Model validation

Model validation is done on the test set by checking the performance metrics of each models

As the data is imbalance we cannot select the accuracy we need to select precision as false positive cases are not accepted

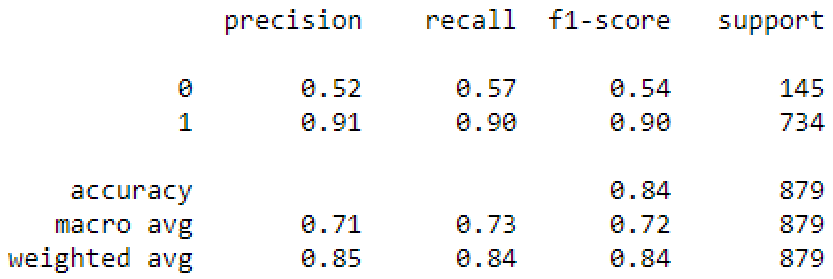
Logistic regression using stats model

Confusion metrics



```
[[ 82  63]
 [ 77 657]]
```

Classification report



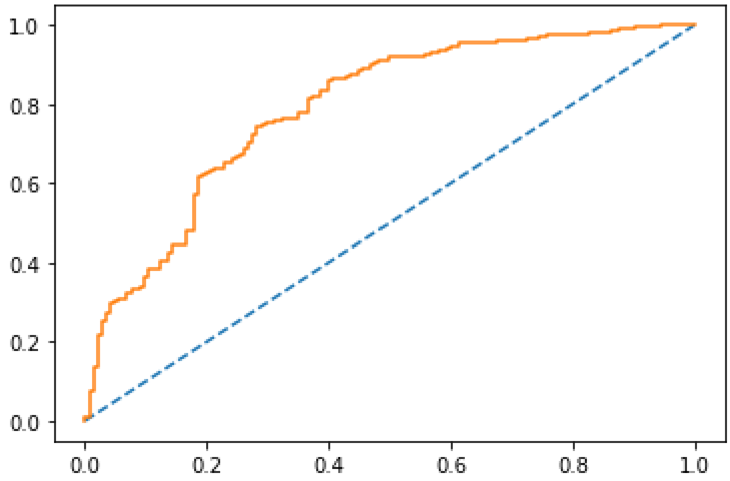
```
              precision    recall  f1-score   support

           0       0.52      0.57      0.54       145
           1       0.91      0.90      0.90       734

    accuracy                           0.84       879
   macro avg       0.71      0.73      0.72       879
weighted avg       0.85      0.84      0.84       879
```

Area under curve

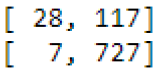
figure 17: roc auc curve of logistic regression using grid search for test



AUC: 0.80

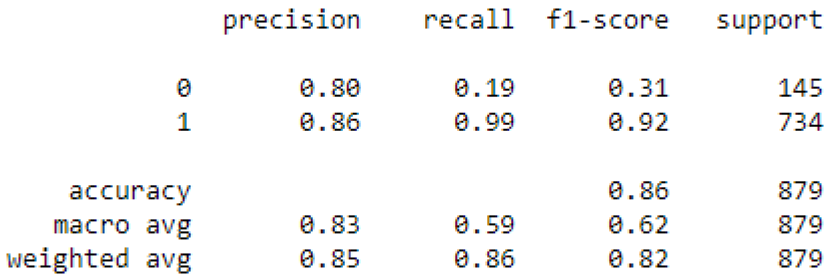
Random forest

Confusion metrics after grid search for test set



```
[ 28, 117]
[  7, 727]
```

Classification report for test set



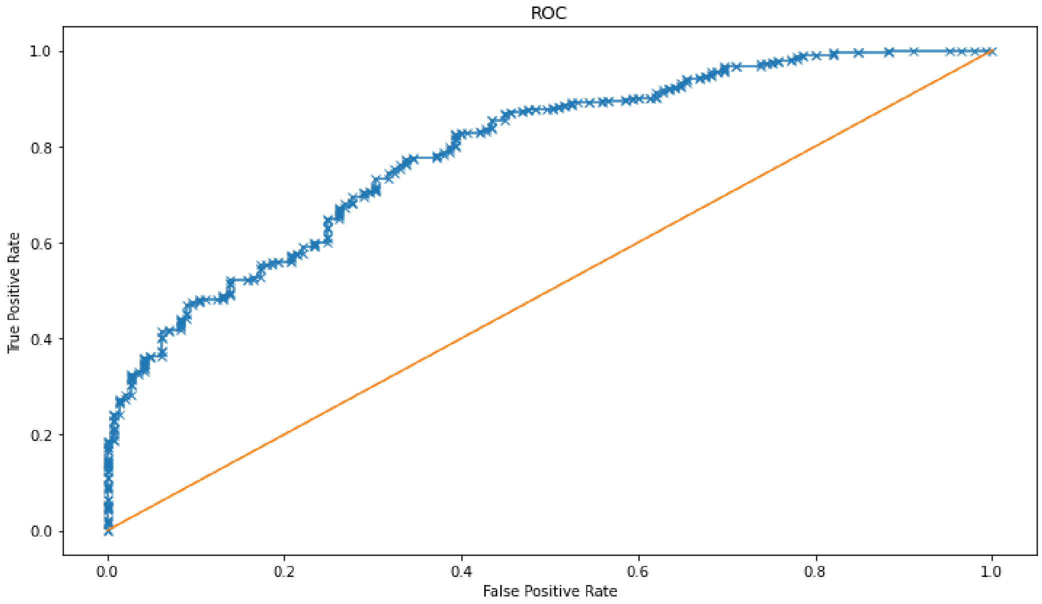
```
              precision    recall  f1-score   support

           0       0.80      0.19      0.31       145
           1       0.86      0.99      0.92       734

    accuracy                           0.86       879
   macro avg       0.83      0.59      0.62       879
weighted avg       0.85      0.86      0.82       879
```

ROC AUC curve

figure 18: roc auc curve of RF model using grid search



```
Area under Curve is 0.7920980926430516
```

Decision tree model (CART) using regularisation

Confusion metrics for test set

```
[ 43, 102]
[ 45, 689]
```

Classification report for test set

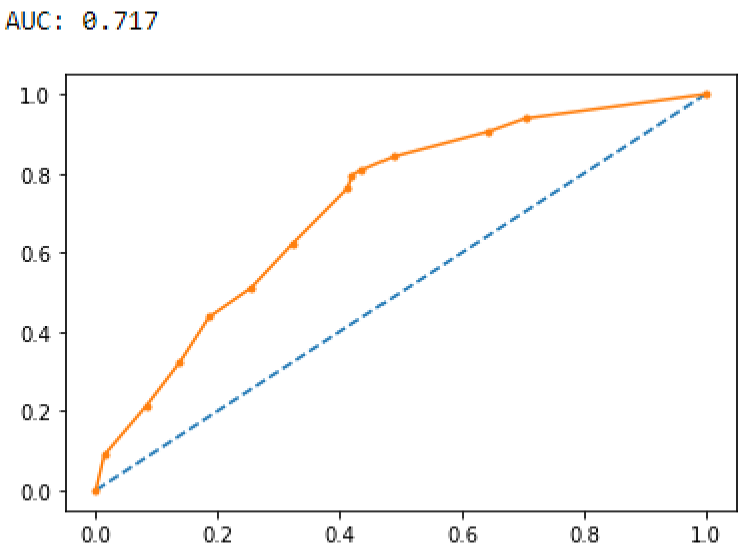
```
              precision    recall  f1-score   support

           0       0.49      0.30      0.37       145
           1       0.87      0.94      0.90       734

    accuracy                           0.83       879
   macro avg       0.68      0.62      0.64       879
weighted avg       0.81      0.83      0.82       879
```

AUC ROC curve for test set

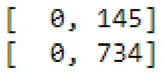
figure 19: roc auc of DT model for test set



```
AUC: 0.717
```

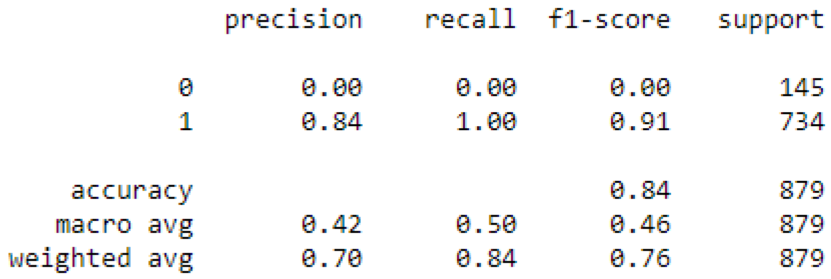
Artificial neural network

Confusion metrics for test set



```
[  0, 145]
[  0, 734]
```

Classification report for test set



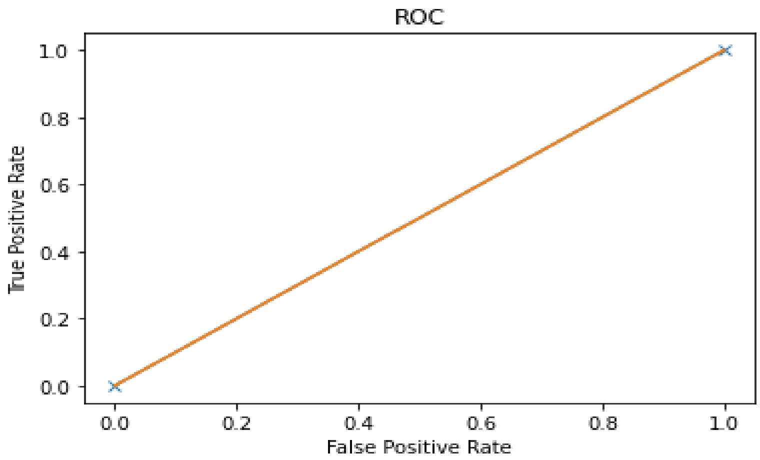
```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       145
           1       0.84      1.00      0.91       734

    accuracy                           0.84       879
   macro avg       0.42      0.50      0.46       879
weighted avg       0.70      0.84      0.76       879
```

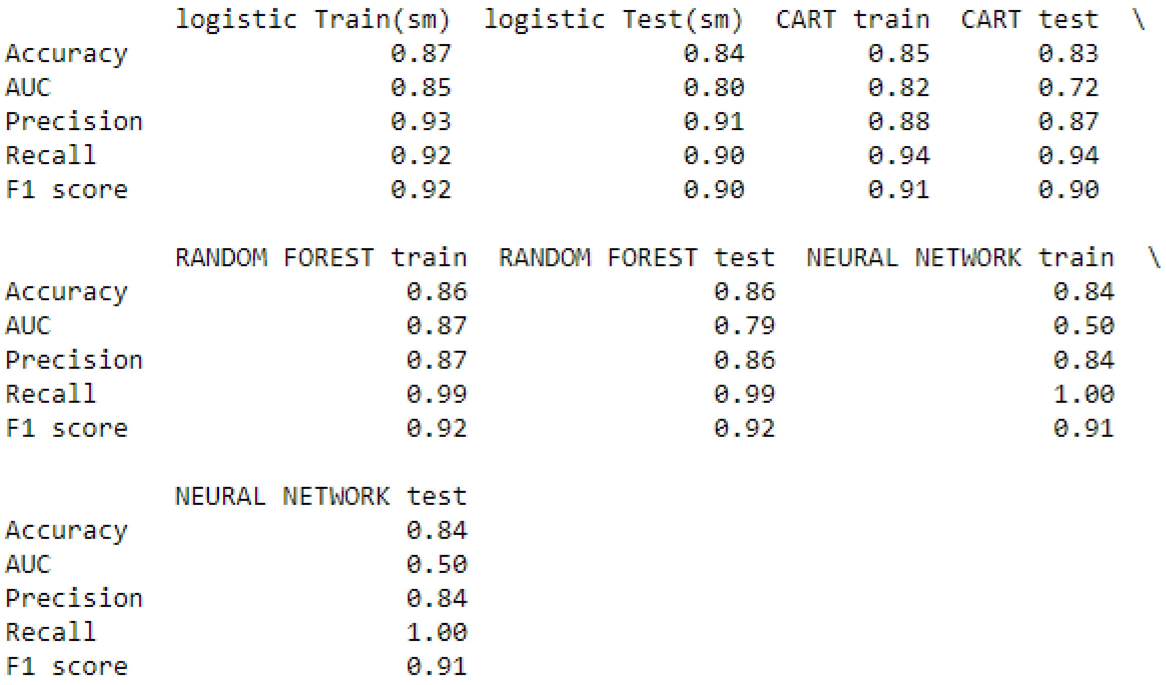
AUC ROC curve

figure 20: roc auc of ANN model for test set



Interpretation of the models

Table 1: interpretations of all the models



```
          logistic Train(sm)  logistic Test(sm)  CART train  CART test  \
Accuracy                0.87               0.84        0.85       0.83
AUC                     0.85               0.80        0.82       0.72
Precision               0.93               0.91        0.88       0.87
Recall                  0.92               0.90        0.94       0.94
F1 score                0.92               0.90        0.91       0.90

          RANDOM FOREST train  RANDOM FOREST test  NEURAL NETWORK train  \
Accuracy                 0.86                0.86                  0.84
AUC                      0.87                0.79                  0.50
Precision                0.87                0.86                  0.84
Recall                   0.99                0.99                  1.00
F1 score                 0.92                0.92                  0.91

          NEURAL NETWORK test
Accuracy                 0.84
AUC                      0.50
Precision                0.84
Recall                   1.00
F1 score                 0.91
```

Logistic regression using stats model

Accuracy is 0.87 and 0.84 for train and test set

As false positive cases are not acceptable. Hence precision is important. It has precision 0.93 and 0.91 for train and test set. It is right fit model.

Area under curve for train and test is 0.85 and 0.80

CART model

Accuracy is 0.85 and 0.83 for train and test set

As false positive cases are not acceptable. Hence precision is important. It has precision 0.88 and 0.87 for train and test set. It is right fit model.

Area under curve for train and test is 0.82 and 0.72

RANDOM FOREST

Accuracy is 0.86 and 0.86 for train and test set

As false positive cases are not acceptable. Hence precision is important. It has precision 0.87 and 0.86 for train and test set. It is right fit model.

Area under curve for train and test is 0.87 and 0.79

Artificial neural network

Accuracy is 0.84 and 0.84 for train and test set

As false positive is not acceptable. Hence precision is important. It has precision 0.84 and 0.84 for train and test set. It is right fit model.

Area under curve for train and test is 0.5 and 0.5

From above all the models we select logistic regression using stats model. It has highest accuracy and precision value. Logistic regression using stats model gives us approach to make the strategy by making changes to the values of the variables with respect to the coefficient

6) Final recommendations given to the BCCI

1 Test match with England in England. All the match are day matches. In England, it will be rainy season at the time to match.

Strategy to be followed

Average team age: Team average age should not be above 34. above age 34 we may lose the match
All-rounder's in team: There should be at least 3 all-rounders' in team
First selection: The first selection should be bowling
Bowlers in team: There should be at least one bowler in the team
Players scored zero: There should be no player scored zero

2 T20 match with Australia in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

1st Strategy to be followed

Average team age: Team average age should be 31. If it is greater 31 then we may lose the match
All-rounder's in team: There should be at least 3 all-rounders' in team, less than 3 all-rounders may lose the match
First selection: The first selection should be bowling
Bowlers in team: There should be 3 bowlers in team
Extra bowls opponents: It should be greater than 14. if it is 14 we may lose the match

Maximum runs given in one over: It should be less than 9 runs. If it 9 or greater then we may lose the match
Player scored zero: There should be at least 2 players scored zero

2nd Strategy to be followed
Average team age: Team average age should be 32. If it is greater than 32 we may lose the match.
All rounder's in team: There should be at least one all-rounder's in team. Playing with no all rounders can lose the match
First selection: The first selection should be batting
Bowlers in team: There should be 2 bowlers in team
Extra bowls opponents: It should be greater than 14. If it is 14 we may lose the match
Maximum runs given in one over: It should not be greater than 4. If it is greater than 4 then we may lose the match
Player scored zero: There should be at least 3 players scored zero

2 ODI match with Sri Lanka in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

1st strategy to be followed
Average team age: Team average age should be less than 36. if it is 36 we may lose the match
All-rounder's in team: There should be at least 2 all-rounder's in team.
First selection: The first selection should be bowling
Bowlers in team: There should be 2 bowlers in team
Maximum runs given in one over: It should be less than 13. If it is 13 or greater than that we may lose the match
Extra bowls opponents: It should be greater than 4. If it is 5 we may lose the match
Player scored zero: We should be having at least two player scored zero. With one players scored zero we may lose the match

2nd strategy to be followed
Average team age: Team average age should be 34. If it is greater than 34 we may lose the match
All-rounder's in team: There should be at least 3 all rounder's in team.
First selection: The first selection should be batting
Bowlers in team: There should be 3 bowlers in team
Maximum runs given in one over: It should be 23. more than 23 runs will result into lost
Extra bowls opponents: It should be greater than 19. If it is 19 we may lose the match
Player scored zero: We should be having at least one player scored zero. With no players scored zero we may lose the match