CAPSTONE PROJECT:-
Cricket Win Prediction

Submitted to

By
Subhadeep Seal

In Partial Fulfillment of PGP-DSBA

CONTENTS

SL.NO.	TITLE	PAGE NO.
1.	Introduction , Problem Statement understanding & need to solve it	3-4
2.	Workflow and tools	5
3.	Data and Variable Description	6
4.	Data Preparation: Missing values & Outlier Detection & Treatment	7-10
5.	Exploratory Data Analysis (EDA): Univariate, Bivariate & Multivariate Analysis	11-13
6.	Feature Selection	14
7.	Model Building	15
8.	Model Validation	16-17
9.	Conclusions	18
10.	Recommendations	19

INTRODUCTION

- Main aim is to create Machine Learning models which correctly predicts a win for the Indian Cricket Team.
- Developing a model to extract and provide actionable insights and recommendation.

PROBLEM STATEMENT UNDERSTANDING

BCCI has contracted an external analytics consulting firm to assist with its data analytics. In this partnership, strategic changes are made to make India win by extracting actionable insights from historical match data. Objective is to build Machine Learning models that correctly predict wins for the Indian Cricket Team. The next step is to extract actionable insights and recommendations from the model.

India will also play the following 10 matches in the next 10 days. It is important to predict the outcome of the matches, and if you get a loss, suggest some changes, and re-run your model until you get a win. The same strategy cannot be used throughout the series, as opponent will become accustomed to it and come up with their own counter strategy. As such, for all the below 5 matches, you must suggest unique strategies to help India win. There should be suggestions that correspond with the variables in the dataset. Be sure to carefully consider whether these suggestions could be implemented. Total no. of matches will be 5.

NEED FOR THE STUDY/PROJECT

- BCCI aims to make data-driven strategic decisions to improve India's win rate. This model supports strategic planning for upcoming matches.
- India is one the successful cricket team in all the formats that is Test, ODI and T20 matches. India plays all the formats throughout the year.
- It is necessary to be world best team which will help the cricket council to maintain the standards and also yield more revenue
- With the above said intention historical data is provided with certain parameters.
- We need to build a accurate model which can predict the future matches.
- The critical need would be if it is a loss then we have to change the parameter accordingly such that India will win the match.
- The tweaking of parameters based on opponent and other parameters has to be predicted.
- The metric to measure the success of this project would be to make the team/council to choose the right given parameters and make India win every match.
- The data for next 5 matches which India going to play is provided. The model has to be built on the historical data and predict the 5 matches.
- Then parameters for each match has to be tweaked as the opponent will understand the strategy as well.

The following are the matches to be predicted,

- Test match with England in England. All the matches are day matches. In England, it will be rainy season at the time to match.
- T20 match with Australia in India. All the matches 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 matches are Day and Night matches. In India, it will be winter season at the time to match.
- The study is a supervised learning and it specifies the class to which data elements belong to and is best used when the output has finite and discrete values.

WORKFLOW & TOOLS

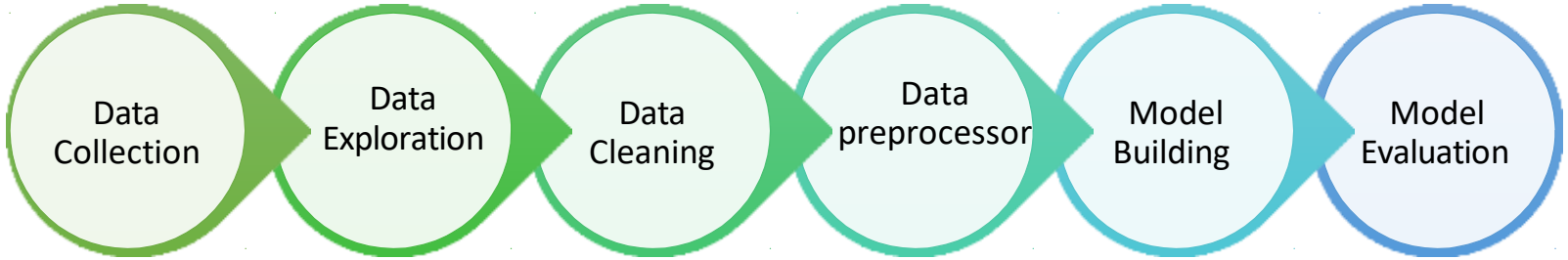


Fig 1: Workflow Diagram

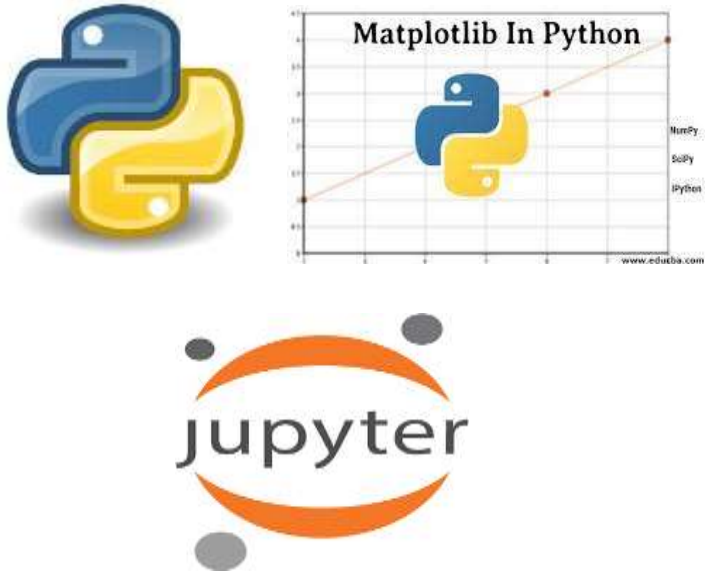


Fig 2: Environment & Tools Used

DATA & VARIABLES DESCRIPTION

- Total 2929 matches result for all three format (T20, ODI, Test) have taken for this case study.
- We have total 23 variable present in the dataset in which 3 datatypes are given float64(9), int64(4), object (10). We have 'Result' as target variable.

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_1ove r	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

Table 1: Data Description

DATA PREPARATION

Missing Value Treatment

- Any data entry which has a null value in any of the predictors was not considered in model building.

Label Encoding& Feature selection

- Convert the categorical variable into numerical by using dummy encoding
- Result, Match format, Match light..etc.
- Feature selection done using Filter technique (Chi Square Test)

Outlier Treatment

- Treated the outlier for **'Avg_team_Age'** variable as there are some invalid entry present in the variable
- For other variable outlier seems to be real value.

Scaling & Data Imbalance

- Scaling performed before fitting KNN model to reduce the variation.
- Synthetic Minority is performed over sampling Technique to overcome Imbalance in data issue.

Missing Value treatment

The data set has many null values. This is detected by using the is null() function. There are a total of 789 null values out of 67390 entries. These missing values were treated using KNN Imputer as they don't contribute towards Model Building.

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

dtype: int64

Before Imputation

Result	0
Avg_team_Age	0
Match_light_type	0
Match_format	0
Bowlers_in_team	0
Wicket_keeper_in_team	0
All_rounder_in_team	0
First_selection	0
Opponent	0
Season	0
Audience_number	0
Offshore	0
Max_run_scored_1over	0
Max_wicket_taken_1over	0
Extra_bowls_bowled	0
Min_run_given_1over	0
Min_run_scored_1over	0
Max_run_given_1over	0
extra_bowls_opponent	0
player_highest_run	0
Players_scored_zero	0
player_highest_wicket	0

dtype: int64

After Imputation

Table 2 – Null Values in the Data Set

Total Percentage of missing values in the dataset is: 1.17 %, which is negligible. But in this data set Avg_team_age column has 97 missing values. So deleting the missing value rows is not a good choice. So we will impute the null values accordingly.

All categorical variables are imputed with the mode values and non-categorical variables are imputed with Median values.

Outlier Detection

The best way to check the outliers is box plot. Hence box plot is drawn for all the numeric variables.

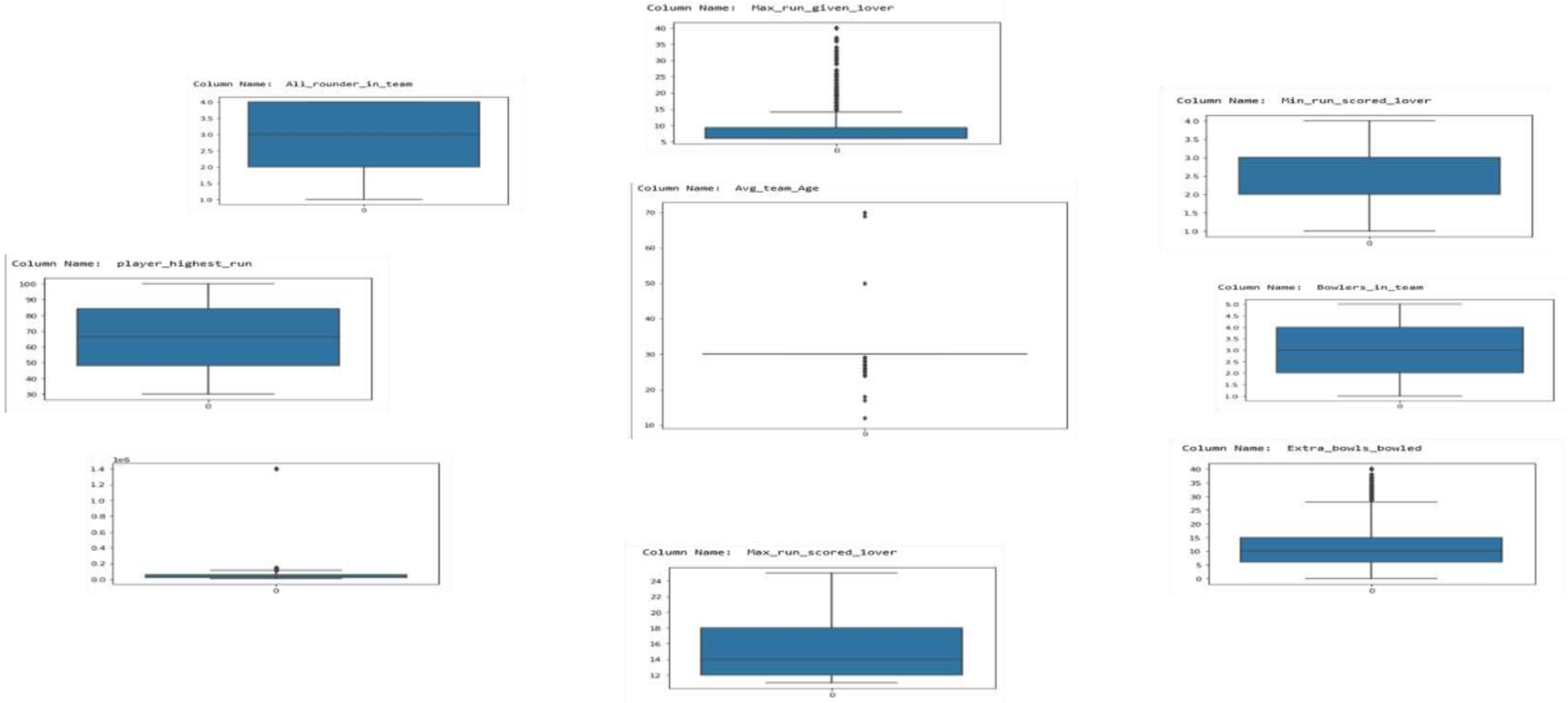


Figure 2 – Box plot on the data set

Outlier Treatment

•Variables like Avg_team_age, player_highest_run, Max_run_given_1Over, Extra_bows_bowled and Audience number has outliers.

•After analyzing the features, only Avg_team_age and Audience variables require outlier treatment in which Audience variable is of no importance in Model building.

•Therefore, only '**Avg_team_age**' is only treated for outliers.

•Rest of the variables based on the data collected is appearing like outliers in the dispersion. But those are all possible in the Cricket and can't blindly treat them which will exploit the data.

•An outlier function is used by calculation the IQR values for treating the outlier.

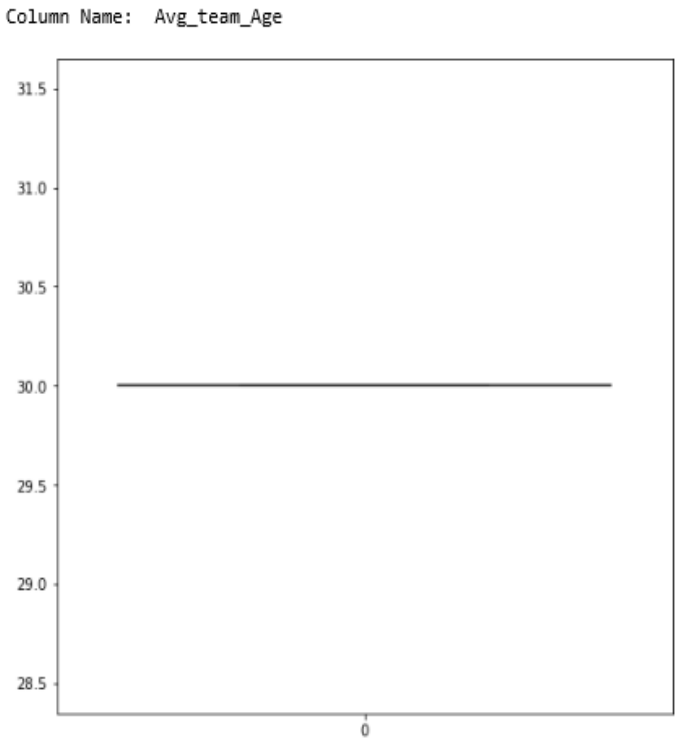


Figure 3 – Box plot for 'Avg_team_Age' column after outlier treatment

Exploratory Data Analysis- Univariate

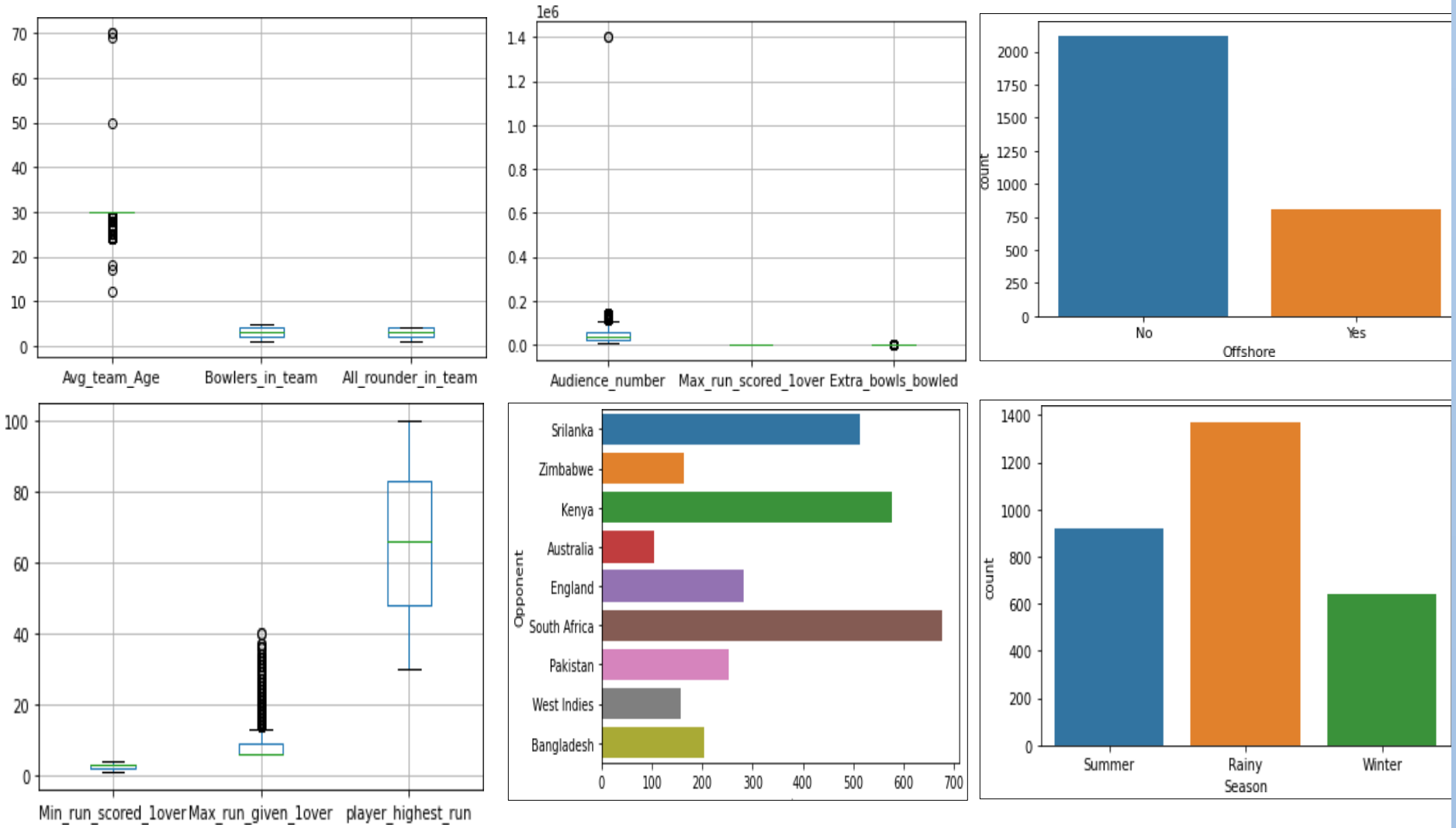
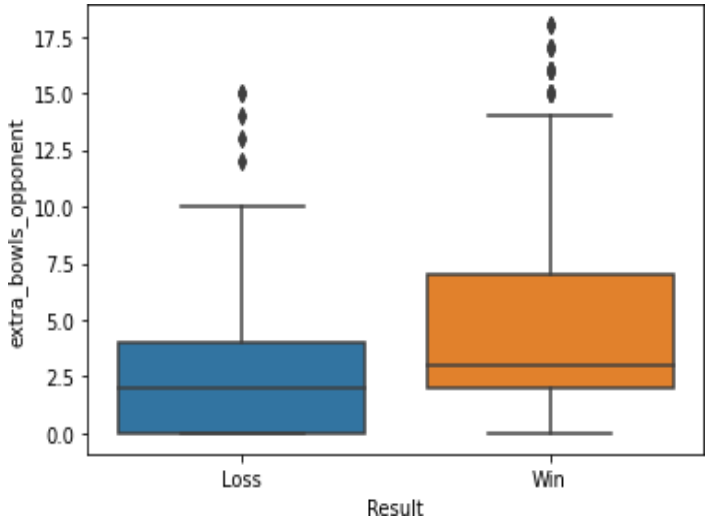


Figure 4 – Box plot & Histogram of Categorical Variables

Insights:

- Around 72% of matches are played in India and only 28% are played out of India.
- Most of the matches are played in Rainy Season.
- Majority of the matches are played against South Africa (676).
- There are some outliers present in the predictors such as Avg_team_age, max_run_1over, audience_number.
- Around 71% ODI Matches are played in Day light.

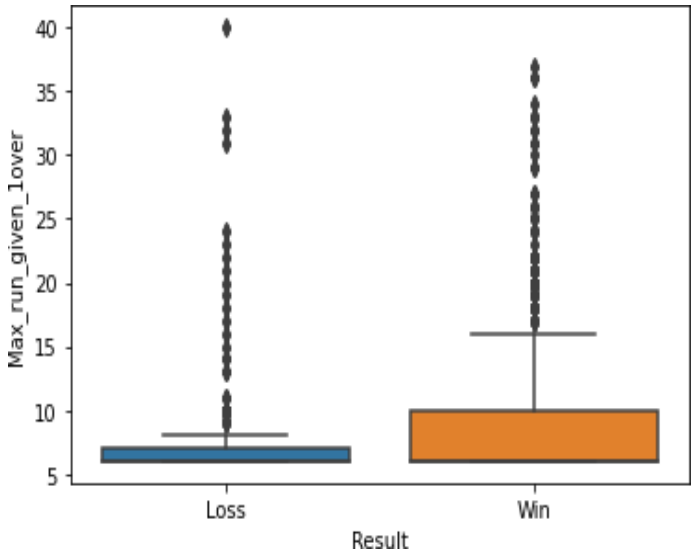
Exploratory Data Analysis- Bivariate



Result	Loss	Win
First_selection		
Batting	172	977
Bowling	301	1480

Result	Loss	Win
Offshore		
No	227	1894
Yes	246	563

Result	Loss	Win
Opponent		
Australia	24	80
Bangladesh	10	194
England	18	265
Kenya	93	483
Pakistan	17	236
South Africa	117	559
Srilanka	124	389
West Indies	4	154
Zimbabwe	66	97



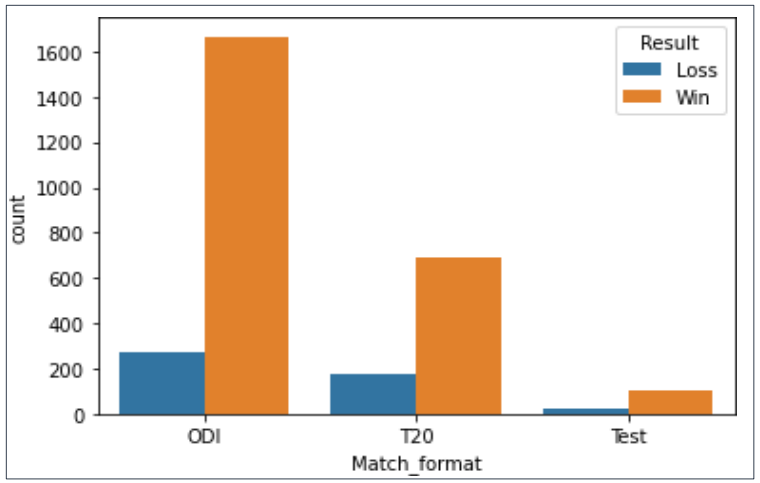


Figure 5 – Box plot for Categorical Variables VS Target Variables

Insights:

- During the bowling contest, India won 51% Matches while Batting 33%.
- Extra ball bowled by the opponent increases winning opportunities for the team.
- On average, 19% of the time when playing outside the country, and 65% when playing within the country, team manage to win.
- Team has won 57% of ODI matches and 24% of T20 matches.
- Inexperienced team (young player) has the higher chances to lose the match

Exploratory Data Analysis- Multivariate

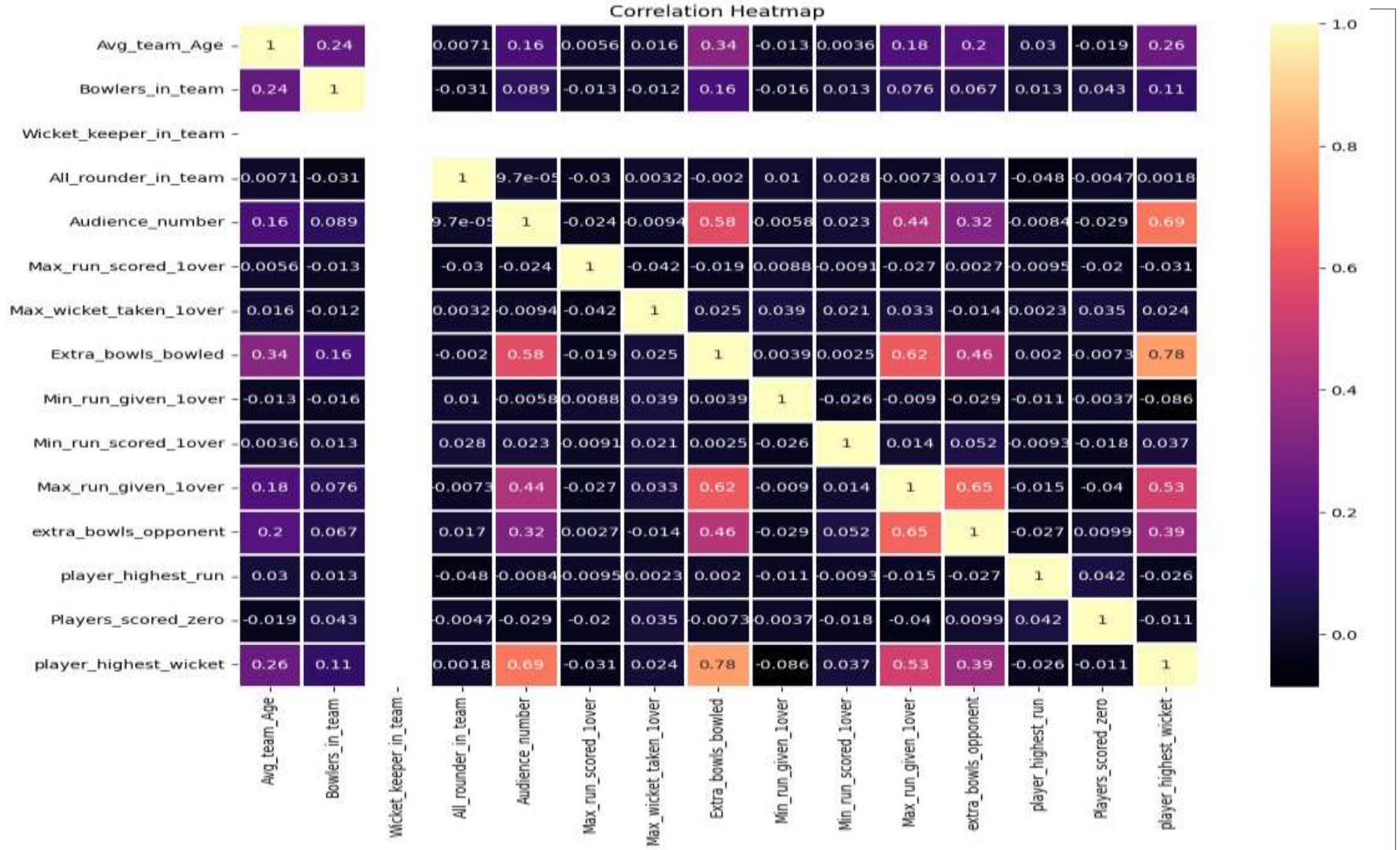


Figure 6 – Box plot for Categorical Variables VS Target Variables

Insights:

- Can see multicollinearity among some variable.
- Extra bowls opponent highly correlated with maximum run given in an over.
  We can also see negative correlation between extra Bowls bowled and all rounder in team.

FEATURE SELECTION

- Chi-square test is used to determine the relationship between the predictor and target variable.
- In Feature selection, we aim to select the features which are highly dependent on the target variable.
- Higher the chi-square value indicate that the feature is more dependent on the target variable and can be selected for model training.
- Chi-square score for Game number is null. So, we eliminate non significant variable Game number.
- After second iteration we find Wicket keeper as non significant variable as per chi-square test and same we can in the heat map. So, both the variable have been eliminated to train our model with remaining predicto

	Feature	Scores
5	Audience_number	1.539290e+06
9	Extra_bowls_bowled	3.373231e+02
13	extra_bowls_opponent	1.676524e+02
6	Offshore	8.085816e+01
26	Opponent_Zimbabwe	5.582450e+01
...	...	...
2952	Game_number_Game_988	NaN
2954	Game_number_Game_99	NaN
2959	Game_number_Game_994	NaN
2960	Game_number_Game_995	NaN
2962	Game_number_Game_997	NaN

2965 rows x 2 columns

Eliminating Game number

	Feature	Scores
5	Audience_number	1.539290e+06
9	Extra_bowls_bowled	3.373231e+02
13	extra_bowls_opponent	1.676524e+02
6	Offshore	8.085816e+01
26	Opponent_Zimbabwe	5.582450e+01
32	player_highest_wicket_1	5.165594e+01
27	Season_Summer	4.525627e+01
12	Max_run_given_1over	3.913104e+01
29	Players_scored_zero_1	3.360147e+01
10	Min_run_given_1over	3.077176e+01
15	Match_light_type_Day and Night	2.109785e+01
19	Opponent_Bangladesh	1.890085e+01
33	player_highest_wicket_2	1.749915e+01
25	Opponent_West Indies	1.614492e+01
24	Opponent_Srilanka	1.599199e+01
20	Opponent_England	1.482618e+01
34	player_highest_wicket_4	1.319061e+01
17	Match_format_T20	1.012452e+01
35	player_highest_wicket_5	1.009267e+01
28	Season_Winter	9.650935e+00
16	Match_light_type_Night	8.998566e+00
31	Players_scored_zero_4	8.890099e+00
3	All_rounder_in_team	7.763370e+00
22	Opponent_Pakistan	7.249808e+00
0	Avg_team_Age	6.308481e+00
18	Match_format_Test	2.171684e+00
8	Max_wicket_taken_1over	2.142275e+00
11	Min_run_scored_1over	1.345619e+00
23	Opponent_South Africa	1.116496e+00
30	Players_scored_zero_2	1.017460e+00
4	First_selection	9.019972e-01
1	Bowlers_in_team	8.729197e-01
21	Opponent_Kenya	5.556134e-01
14	player_highest_run	4.414944e-01
7	Max_run_scored_1over	7.549787e-02
2	Wicket_keeper_in_team	0.000000e+00

Eliminating Wicket keeper in team

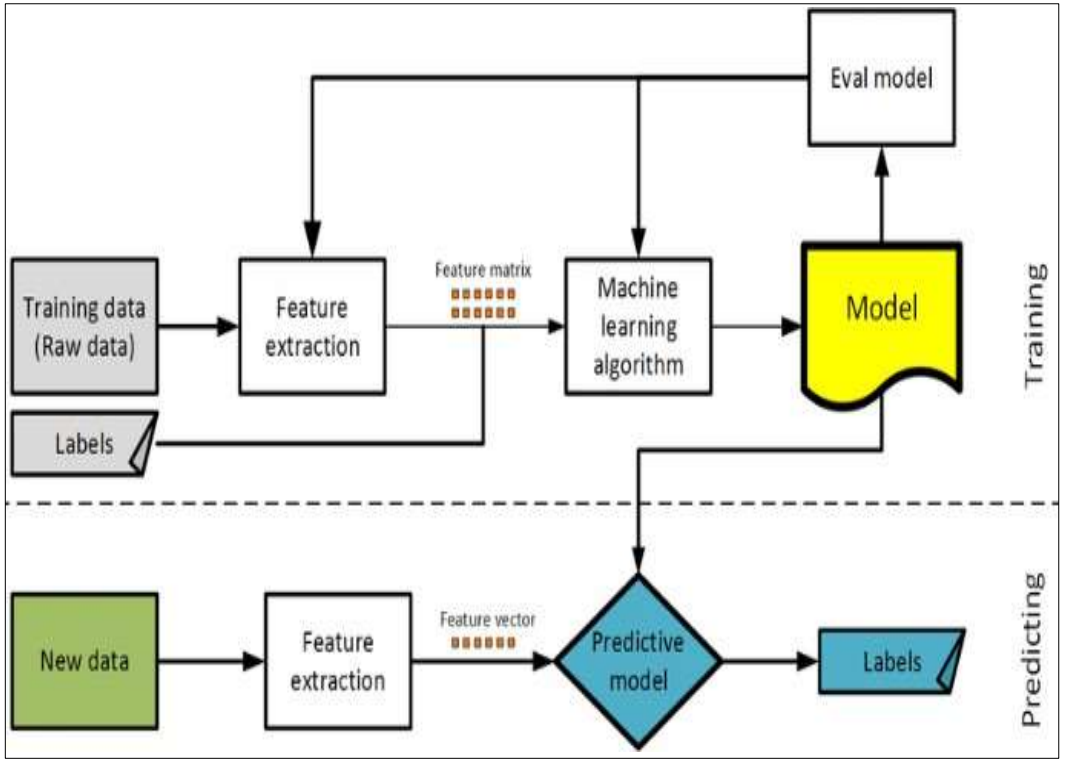
Variables
Result
Avg_team_Age
Match_light_type
Match_format
Bowlers_in_team
All_rounder_in_team
First_selection
Opponent
Season
Audience_number
Offshore
Max_run_scored_1over
Max_wicket_taken_1over
Extra_bowls_bowled
Min_run_given_1over
Min_run_scored_1over
Max_run_given_1over
extra_bowls_opponent
player_highest_run
Players_scored_zero
player_highest_wicket

Target Variable

Model Building

➢ Target variable –**Result** is categorical in nature.
➢ For this case study following classification model will be used:

- Logistic Regression
- Adaboost
- Random Forest Classifier
- XG boost
- KNN
- SVM
- Naive Baye's



Evaluation Parameters

Accuracy	Recall	Precision	F1-score
• Correctly classified point in test data and total number of points in the test data	• The ratio of total positive predicted by the model to the total actual positive.	• Out of the total positive, what percentage are predicted positive.	• It is the harmonic mean of precision and recall. It takes both false positive and false negatives into account

Model Validation

- All model except Random Forest model is giving 96.59% Accuracy and 100% of recall.

- In this case our most important matrix is Recall because we must predict winning for the Indian team and must reduce the false predicted records.

Model Selected:

- Tree based models perform better due to feature predictions.

- It sequentially add the misidentified predictors and under-fitted predictions to the ensemble, ensuring the errors identified previously are corrected.

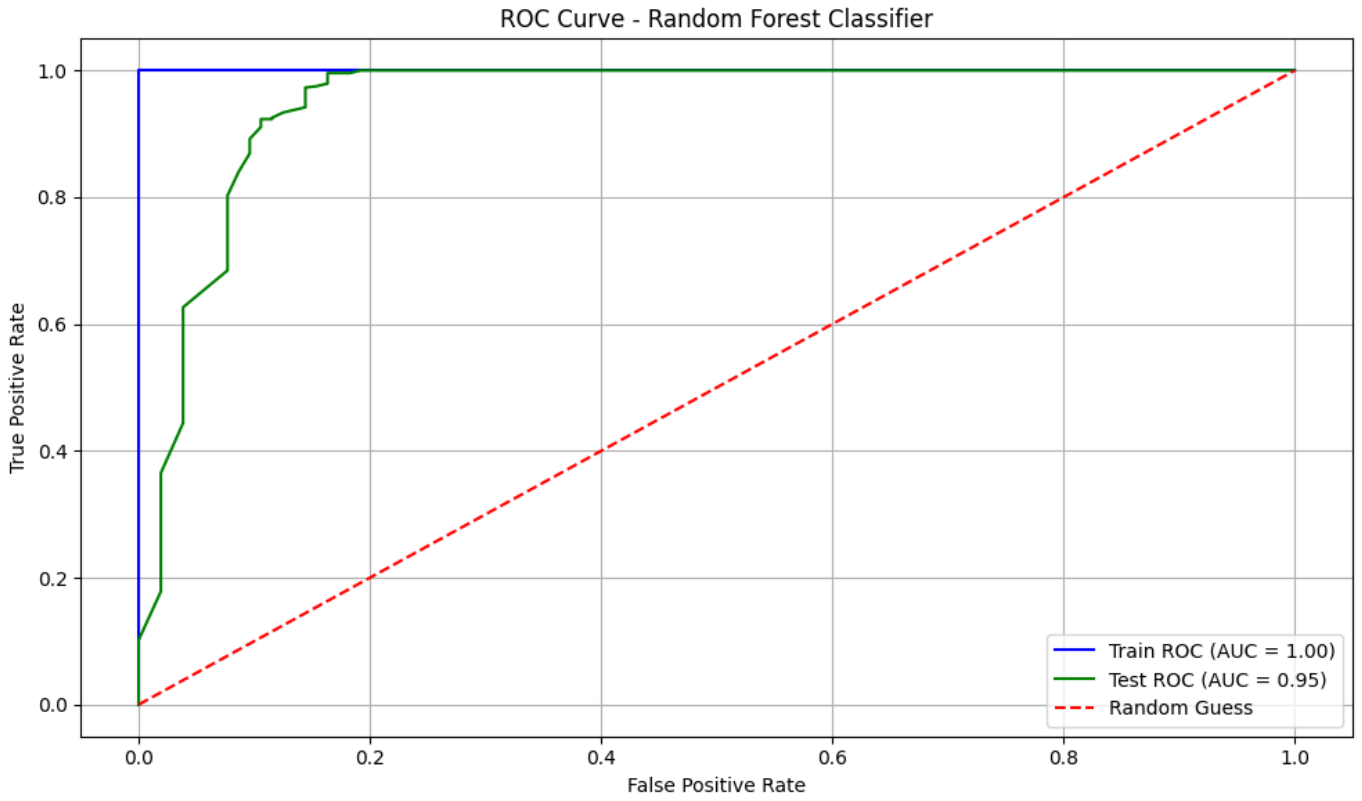
- Random Forest model has less False '+ve' and False '-ve' for both win and loss Classes. Comparing to other model it has AUC, Recall and Accuracy for both Train and Test.

	Model	Accuracy	Precision	Recall	F1 Score	AUC
0	Random Forest	0.9659	0.9602	1.0000	0.9797	0.9550
1	XGBoost	0.9642	0.9619	0.9959	0.9786	0.9449
2	SVM	0.8667	0.8649	0.9959	0.9257	0.8624
3	KNN	0.8635	0.8722	0.9772	0.9217	0.8361
4	AdaBoost	0.8669	0.8673	0.9896	0.9244	0.8270
5	Logistic Regression	0.8447	0.8523	0.9813	0.9122	0.7955
6	Naive Bayes	0.7778	0.8788	0.8507	0.8645	0.7000

Table 3 – Model Comparison

Model Evaluation Random Forest on Test Data

- For test data set we got 96.59% Accuracy and 100% Recall.
- We received very less false positive rate for this model.



MODEL	RM-Test
Accuracy	96.59
F1 Score	97.97
Recall	100
Precision	96.02
AUC	95.50

Table 4 – Best performing Test set

- In the first batting situation, the team wins 33% of the time and looses only 0.06% of the time.
- In the first bowling situation, team wins 51% of the time and looses 10% of the time.

- When playing in daylight, winning chances increase by 60%.
- In ODIs, the team has won 1666 games out of 1935.

- Out off 2121 matches India only lost 227 Matches while playing within the country.
- India has a high probability of winning at home Ground.
- India's Win rate improves by 18% batting first at home during winter.

- Playing with more than 2 all-rounders in a team increases winning chances by more than 50%.
- Team won 1201 matches in Rainy condition out off 1371.
- Its show Rain helps bowler to get some extra advantage in his bowling.

- Try to collect some more predictor like total score & bowling style for a better model.
- Try to add more than 3 all rounder in the team that will improve the team performance.
- If team opt for bowling first with an Avg team age of 30, with 4 bowlers in the team has higher chance to win against England in test match in Rainy season in England .
- If team opt for bowling first with an Avg team age of 30, minimum 3 bowlers in the team, scoring average 15 runs per over has higher chance to win against Australia in T20 match in Winter season in India.
- If team opt for Batting first with an Avg team age of 30, with 3 bowlers in the team and at least one player should score century has higher chance to win against Sri Lanka in ODI match in Winter season in India.

THANK YOU