PRESENTATION OVERVIEW

UNDERSTANDING BUSINESS PROBLEM
BUSINESS INSIGHT AND OBJECTIVE

MODELS FOR PROBLEM SOLVING

- **EXPLORATORY DATA ANALYSIS**
- **MODEL BUILDING**
- **MODEL TUNING**

ANALYSIS INSIGHTS

- **MODEL EVALUATION AND VALIDATION**
- **MODEL INTERPRETATION**

BUSINESS RECOMMENDATIONS FOR WIN STRATEGY
(TEST:1; ODI:2; T20:2)

UNDERSTANDING PROBLEM STATEMENT

External analytics consulting firm was hired by BCCI for data analytics of Sports data. The sports data set with the information of games played previously were provided to develop strategies for winning cricket game by Indian cricket team.

BUSINESS INSIGHT

Cricket is liked by most of the people across the globe and has a huge finance involved in organising them. In turn, the BCCI should also earn profits by organising them. The BCCI with the winning strategies and recommendations provided by the analysed data of the past games played, can be used to train Indian team so that the Indian team can have a better win and in turn have profits for the BCCI from the tickets sold to audience, advertising companies, promotions for products, parking payments etc.

There is rapid expansion of sports data analytics every year. Technological advances in development of different machine learning models, evaluating models have gained insight to formulate strategies for sports so that chances of win can be increased or achieved.

OBJECTIVE

To propose different win strategies for Indian cricket to be played in various formats i.e., Test (1 strategy), ODI (2 strategies), T20 (2 strategies) for INDIAN CRICKET TEAM WIN. The strategies are to be provided for matches played in India against Sri Lanka (ODI), Australia(T20) during winter season as day and night matches and for test match to be played against England in England as day match in rainy season.

Data Set given: 'Sports Data.xlsx'
Target Variable: Result

DESCRIPTIVE STATISTICS OF CONTINUOUS VARIABLES

Numerical variables	count	mean	std	min	25%	50%	75%	max
Avg_team_Age	2833	29.24	2.26	12	30	30	30	70
Bowlers_in_team	2848	2.91	1.023	1	2	3	4	5
Wicket_keeper_in_team	2930	1	0	1	1	1	1	1
All_rounder_in_team	2890	2.72	1.09	1	2	3	4	4
Audience_number	2849	46267.96	48599.58	7063	20363	34349	57876	1399930
Max_run_scored_1over	2902	15.19	3.66101	11	12	14	18	25
Max_wicket_taken_1over	2930	2.713993	1.080623	1	2	3	4	4
Extra_bowls_bowled	2901	11.252671	7.780829	0	6	10	15	40
Min_run_given_1over	2930	1.95256	1.678332	0	0	2	3	6
Min_run_scored_1over	2903	2.762659	0.705759	1	2	3	3	4
Max_run_given_1over	2896	8.669199	5.003525	6	6	6	9.25	40
extra_bowls_opponent	2930	4.229693	3.626108	0	2	3	7	18
player_highest_run	2902	65.889387	20.331614	30	48	66	84	100

DESCRIPTIVE STATISTICS OF CATEGORICAL VARIABLES

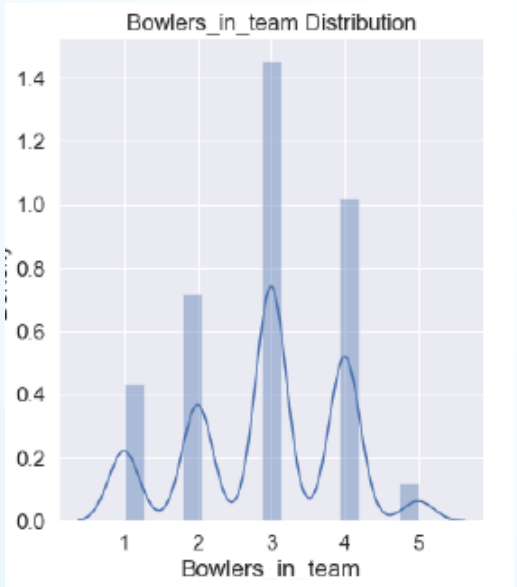
Categorical Varibales	count	unique	top	freq
Game_number	2930	2930	Game_799	1
Result	2930	2	Win	2457
Match_light_type	2878	3	Day	2041
Match_format	2860	4	ODI	1865
First_selection	2871	3	Bowling	1722
Opponent	2894	9	South Africa	640
Season	2868	3	Rainy	1309
Offshore	2866	2	No	2057
Players_scored_zero	2930	5	3	1730
player_highest_wicket	2930	6	1	1084

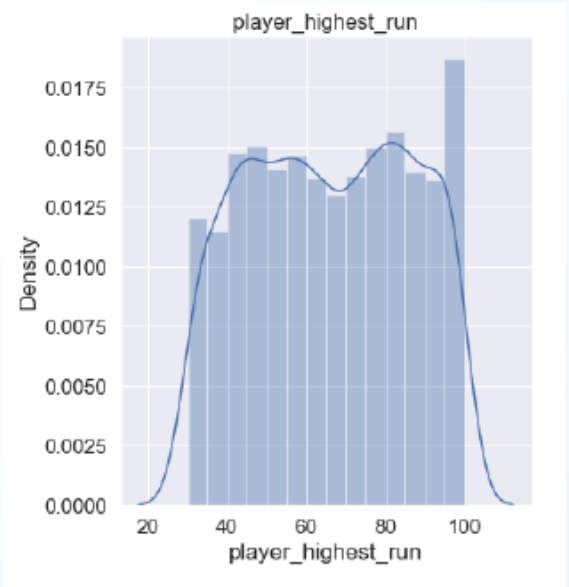
EXPLORATORY DATA ANALYSIS

DESCRIPTION	INFERENCE
DATASET SHAPE	23 COLUMN VARIABLES AND 2390 ROWS
DUPLICATES	NO DUPLICATES IN DATASET
VARIABLES	The Categorical variables comprise of Game_number; Result; Match_light_type; Match_format; First_selection; Opponent; Season; Offshore; Players_scored_zero; Player_highest_wicket. The Numerical or continuous variables comprise of Avg_team_Age ; Bowlers_in_team; Wicket_keeper_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.
UNWANTED VARIABLES	REMOVED 'Game_number'
MISSING VALUES	789 VALUES. AFTER IMPUTATION ZERO MISSING VALUES
OUTLIERS	PRESENT IN DATASET FOR FEW VARIABLES TREATED BY IQR
VARIABLE TRANSFORMATION	LABEL ENCODING FOR THE CATEGORICAL VARIABLES SCALING OF DATA FOR CONTINUOUS VARIABLES

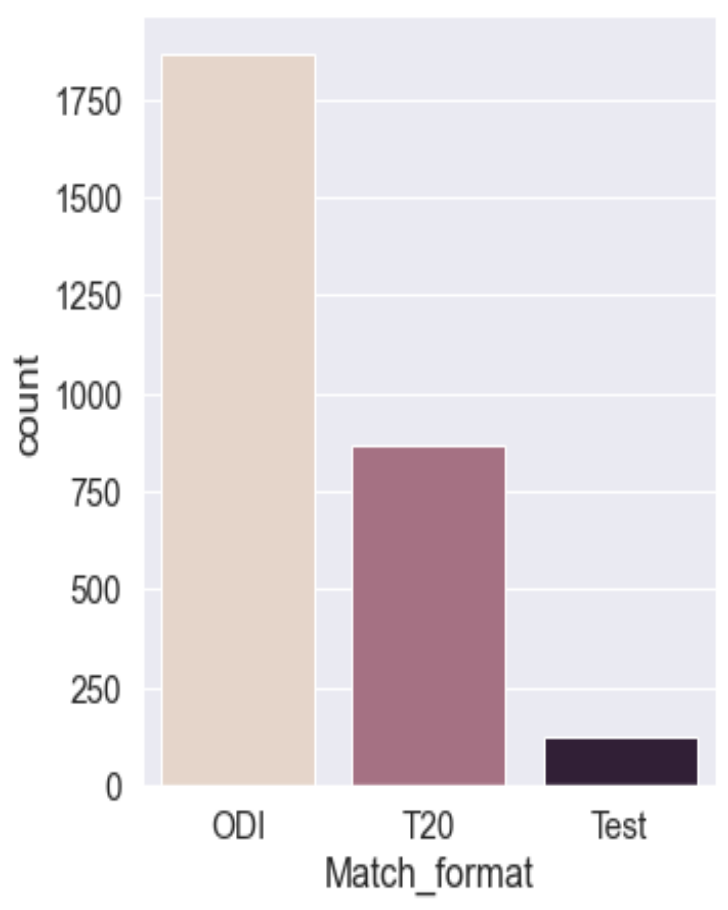
UNIVARIATE ANALYSIS OF DATA

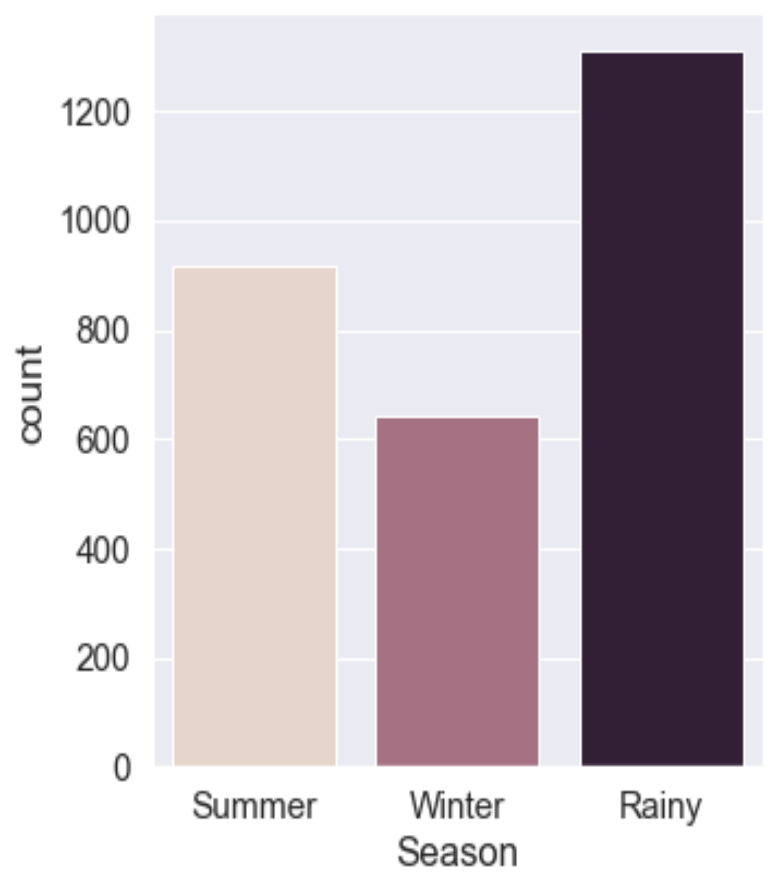
DISTRIBUTION	VARIBALES
Normal distribution	Bowlers_in_team, Wicket_keeper_in_team, player_highest_run
Left skewed	Avg_team_Age, All_rounder_in_team, Max_wicket_taken_1over, Min_run_scored_1over
Right skewed	Audience_number, Extra_bowls_bowled, Min_run_given_1over, Max_run_scored_1over, Max_run_given_1over, extra_bowls_opponent

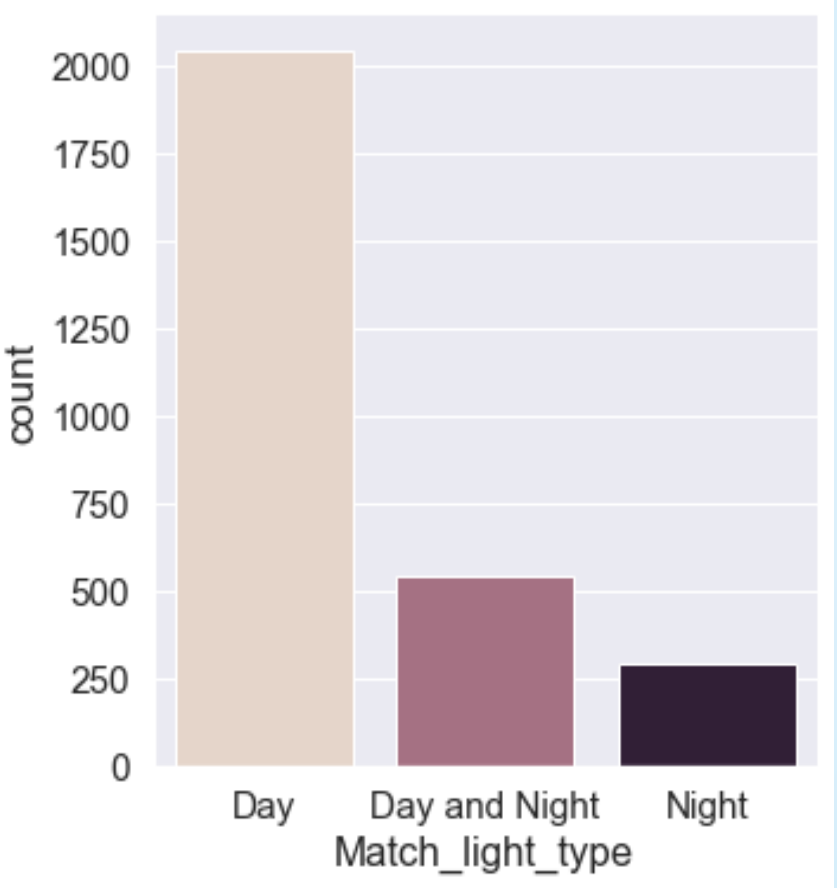




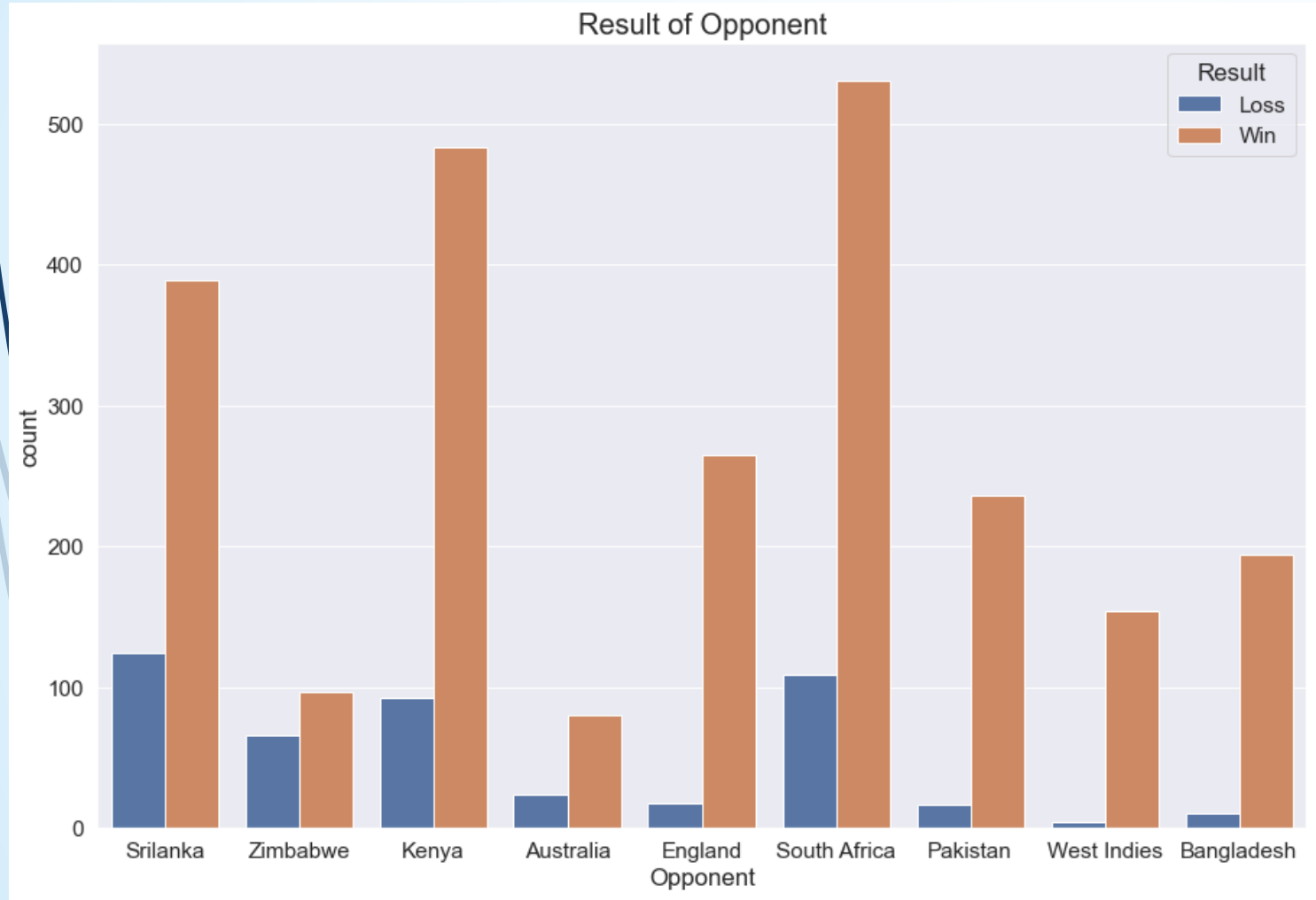
COUNT PLOTS FOR CATEGORICAL DATA

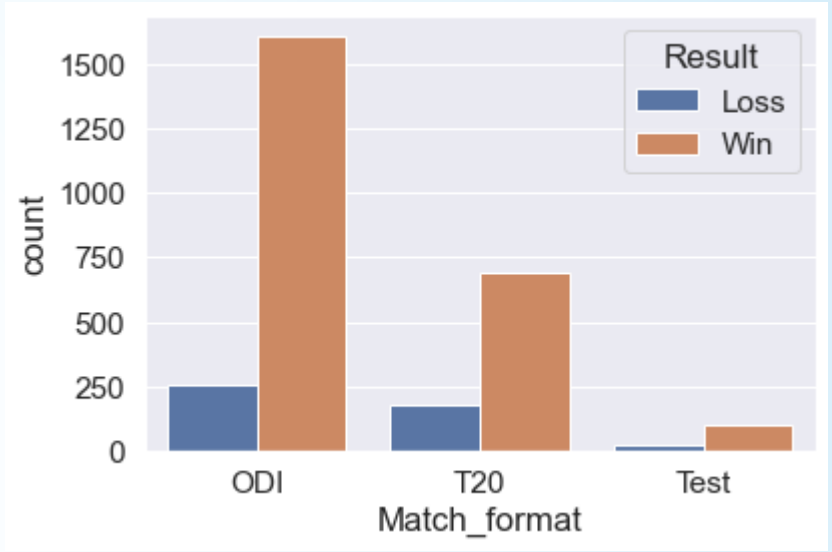


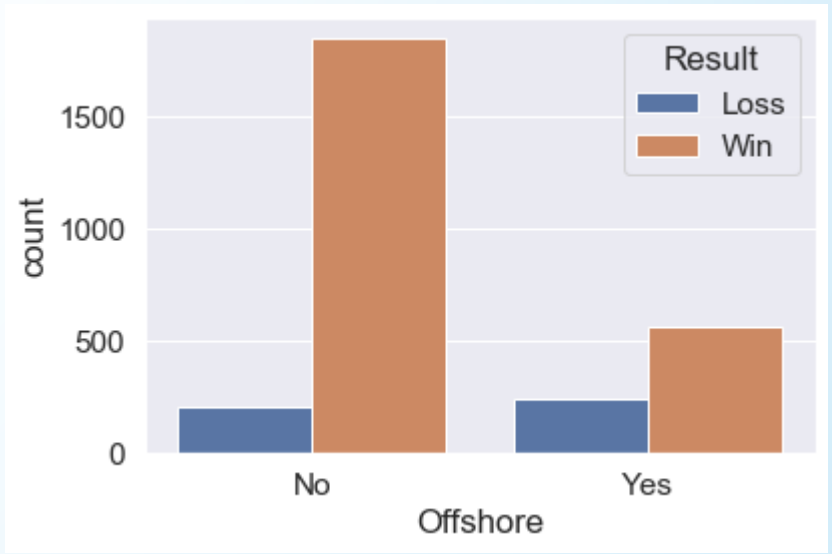




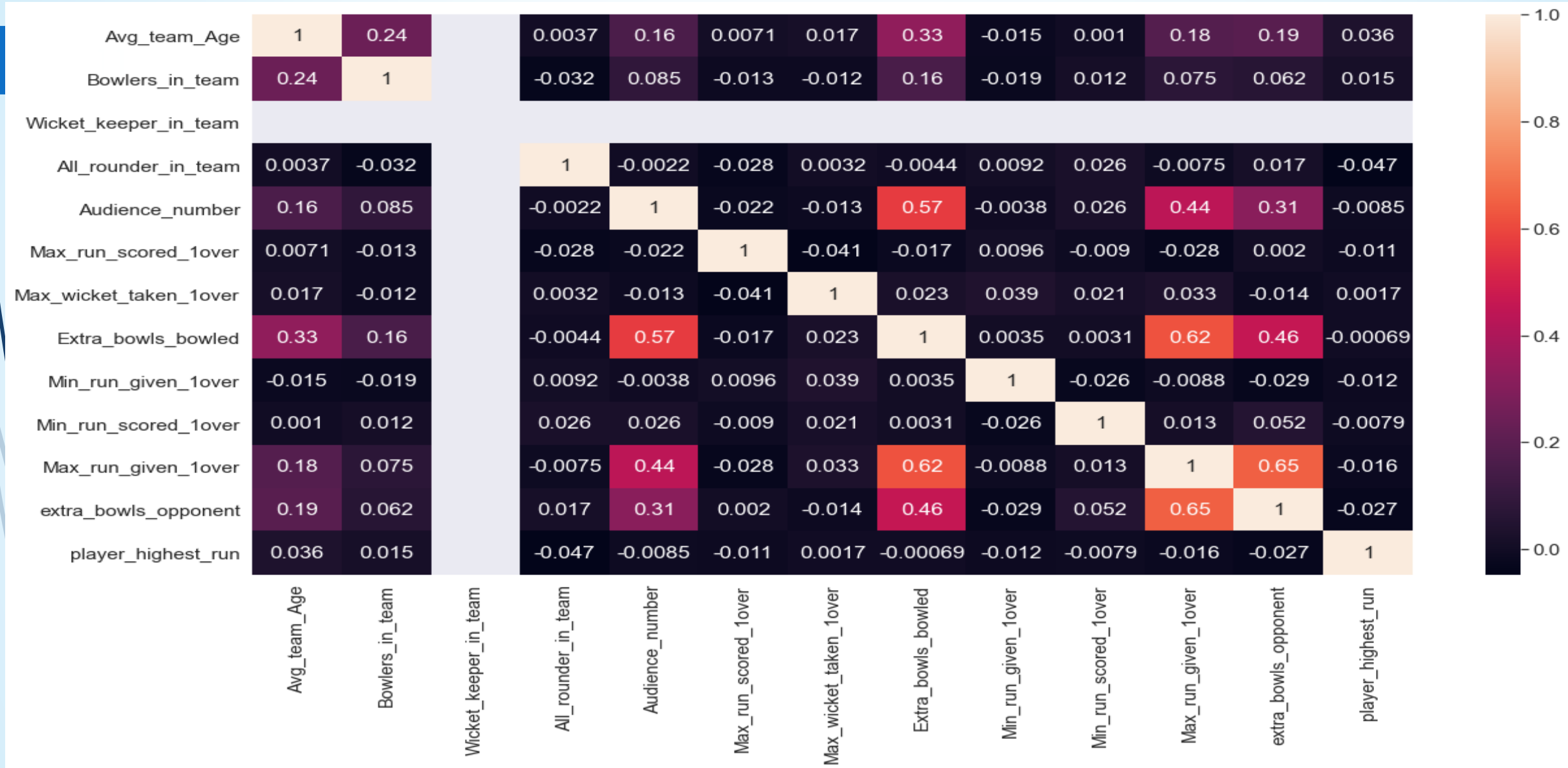
BIVARIATE ANALYSIS OF DATA







CORRELATION OF CONTINUOUS VARIABLES

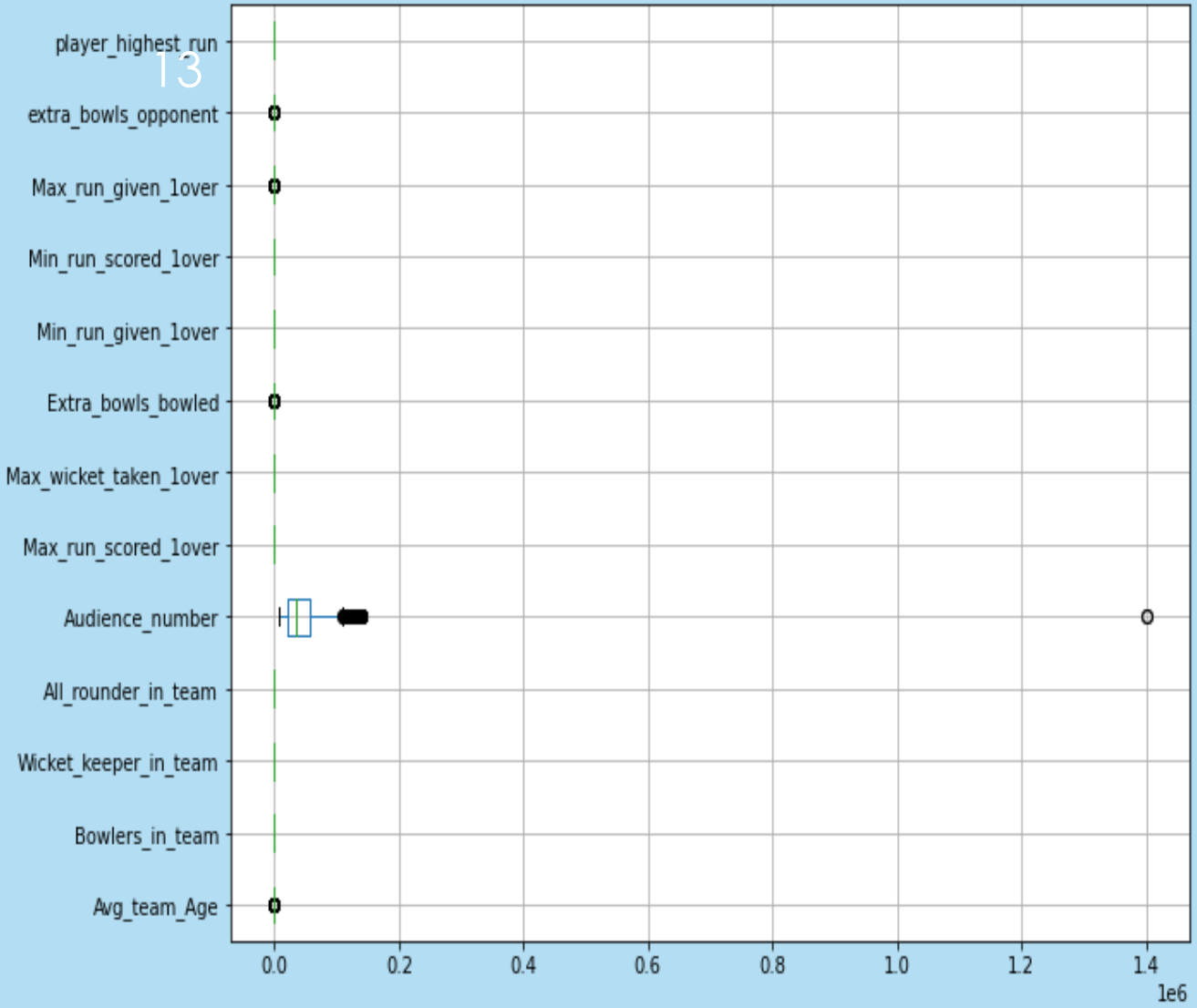


CORRELATION IS MAXIMUM FOR EXTRA_BOWLS_OPPONENT AND MAX_RUN_GIVEN_1OVER (0.65)

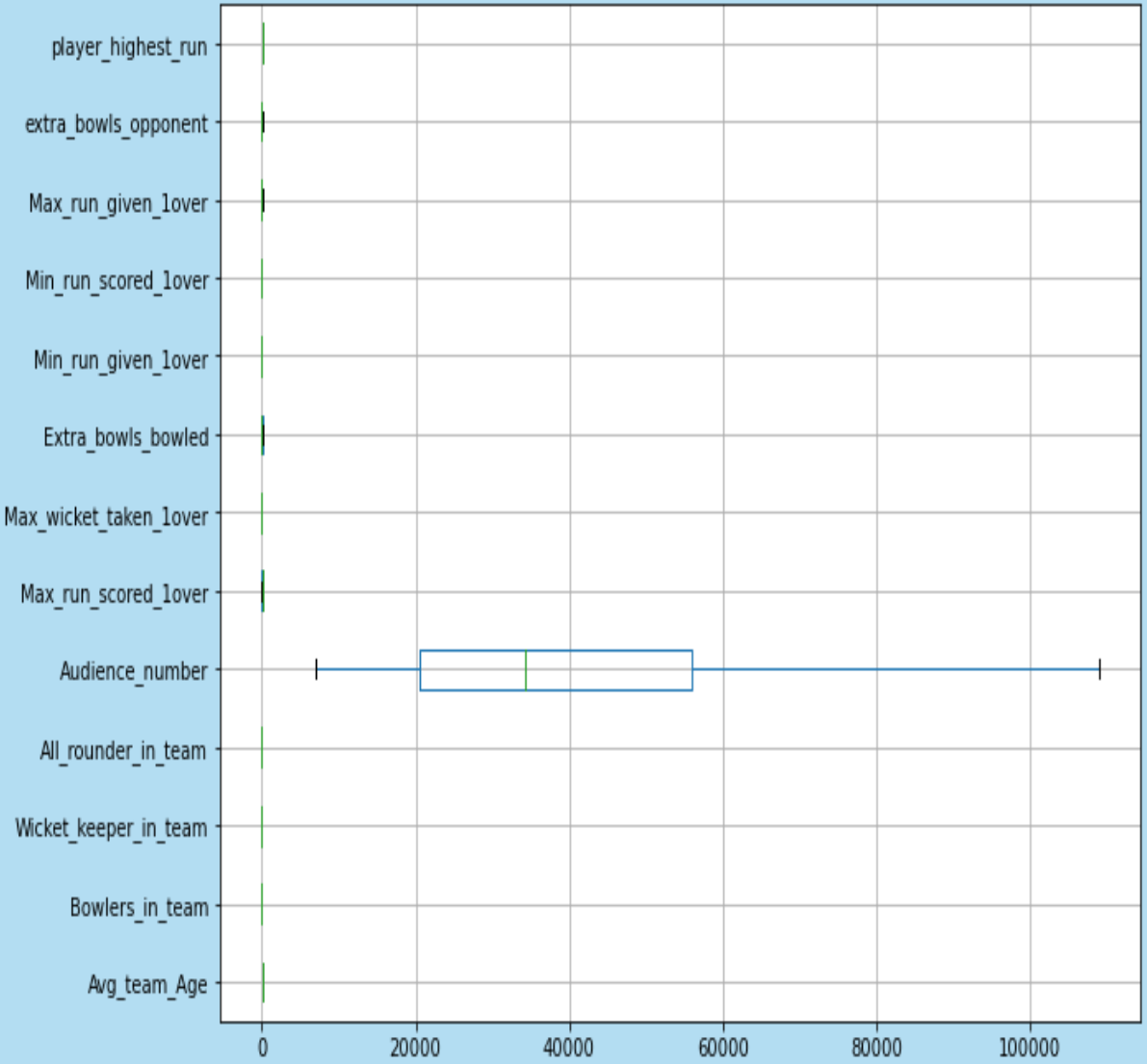
MISSING VALUES

Variables	Before imputaion	After imputaion
Result	0	0
Avg_team_Age	97	0
Match_light_type	52	0
Match_format	70	0
Bowlers_in_team	82	0
Wicket_keeper_in_team	0	0
All_rounder_in_team	40	0
First_selection	59	0
Opponent	36	0
Season	62	0
Audience_number	81	0
Offshore	64	0
Max_run_scored_1over	28	0
Max_wicket_taken_1over	0	0
Extra_bowls_bowled	29	0
Min_run_given_1over	0	0
Min_run_scored_1over	27	0
Max_run_given_1over	34	0
extra_bowls_opponent	0	0
player_highest_run	28	0
Players_scored_zero	0	0
player_highest_wicket	0	0
Total missing values	**789**	**0**

BEFORE OUTLIER TREATMENT



AFTER OUTLIER TREATMENT



AVERAGE TEAM AGE, AUDIENCE NUMBER, EXTRA BOWLS BOWLED, MAXIMUM RUN GIVEN 1OVER, EXTRA BOWLS OPPONENT HAS OUTLIERS WHICH WERE REMOVED.

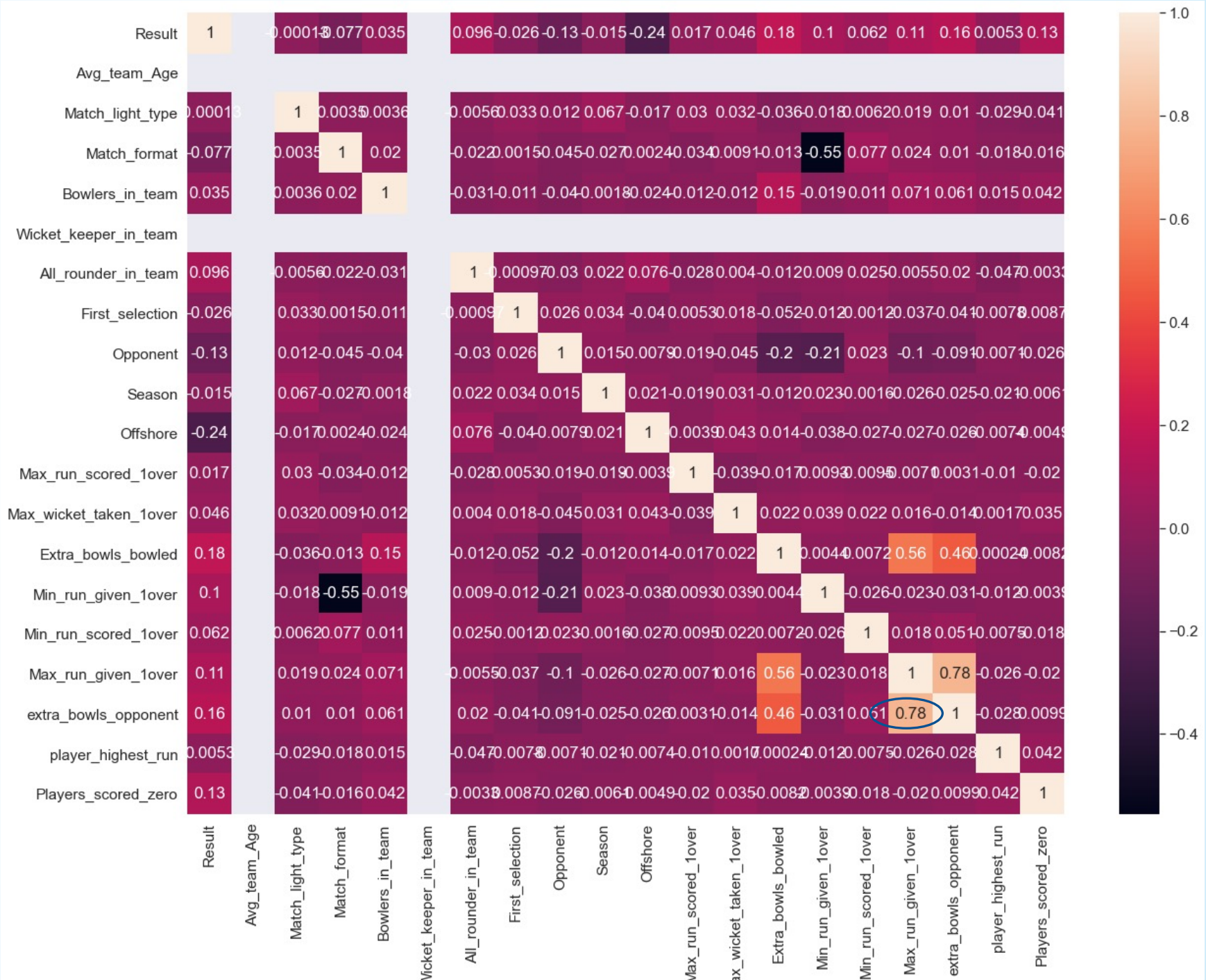
VARIANCE INFLATION FACTOR

	feature	VIF
0	Result	1.184994
1	Avg_team_Age	0
2	Match_light_type	1.014881
3	Match_format	1.545418
4	Bowlers_in_team	1.028235
5	All_rounder_in_team	1.030228
6	Wicket_keeper_in_team	0
7	First_selection	1.008614
8	Opponent	1.165731
9	Season	1.010222
10	Audience_number	7.097595
11	Offshore	1.090127
12	Max_run_scored_1over	1.008721
13	Max_wicket_taken_1over	1.016987
14	Extra_bowls_bowled	3.084353
15	Min_run_given_1over	1.616754
16	Min_run_scored_1over	1.019664
17	Max_run_given_1over	3.036599
18	extra_bowls_opponent	2.663806
19	player_highest_run	1.010529
20	Players_scored_zero	1.029654
21	player_highest_wicket	8.096399

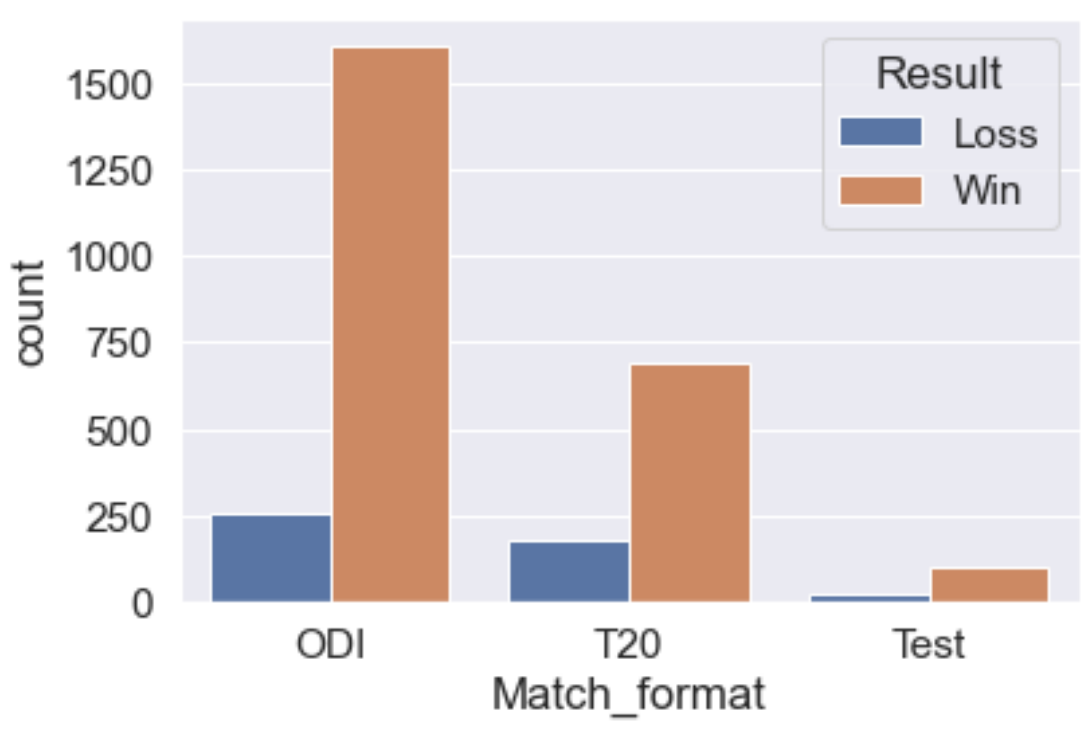
AUDIENCE_NUMBER AND PLAYER_HIGHEST_WICKET HAS HIGHEST VIF >5.0 AND CAN BE DROPPED

CORRELATION OF DATA AFTER VARIABLE TRANSFORMATION

CORRELATION IS MAXIMUM FOR EXTRA_BOWLS_OPPONENT AND MAX_RUN_GIVEN_1OVER (0.78)



The data of the target variable is unbalanced as it depicts target variable=1 for 83.8%(win) of the cases and 16.2% (loss)for target variable=0.



CLUSTERING

CLUSTERING WAS DONE AND THREE DIFFERENT MATCH FORMATS OF ODI, T20 AND TEST ARE CONSIDERED TO DEVELOP MODELS

MODEL BUILDING

FOR OVERALL DATA SET DIFFERENT MODELS MENTIONED BELOW ARE BUILD

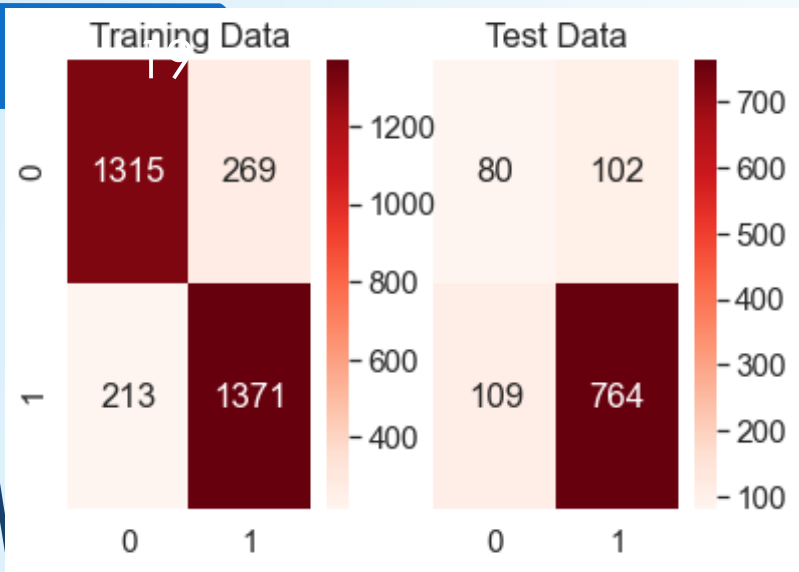
- DECISION TREE CLASSIFIER (DTC)
- RANDOM FOREST CLASSIFIER (RFC)
- NEURAL NETWORK CLASSIFIER (NNC)
- LOGISTIC REGRESSION (LOG REG)
- LINEAR DISCRIMINANT ANALYSIS (LDA)
- NAÏVE BAYES WITH SMORE (NBS)
- KNN WITH SMOTE (KMMS)

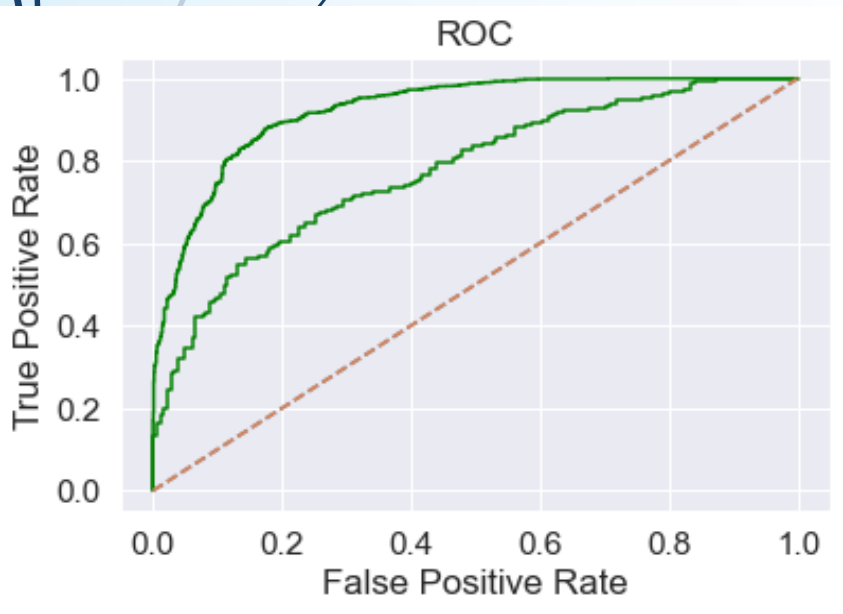
MATCH FORMAT	MODELS BUILD
ODI MATCH	➢ RANDOM FOREST CLASSIFIER (RFC) ➢ NEURAL NETWORK CLASSIFIER (NNC)
T20 MATCH	➢ RANDOM FOREST CLASSIFIER (RFC) ➢ NEURAL NETWORK CLASSIFIER (NNC) ➢ KNN WITH SMOTE (KMMS)
TEST MATCH	➢ LOGISTIC REGRESSION (LOG REG) ➢ NEURAL NETWORK CLASSIFIER (NNC)

PERFORMANCE METRICS OF DIFFERENT MODELS BUILD FOR OVERALL DATA

		PERFORMANCE METRICS					
MODEL BUILDING	DATA	PRECISION	ACCURACY	RECALL	F1 SCORE	AUC	Win %
DTC	TRAIN SET	0.77	0.77	0.77	0.77	0.86	77.82
DTC	TEST SET	0.88	0.72	0.76	0.82	0.732	75.21
RFC	TRAIN SET	0.84	0.84	0.87	0.85	0.924	86.38
RFC	TEST SET	0.88	0.8	0.8	0.88	0.776	85.34
NNC	TRAIN SET	0.73	0.77	0.85	0.79	0.863	84.08
NNC	TEST SET	0.91	0.80	0.85	0.88	0.789	83.5
LOG REG	TRAIN SET	0.77	0.77	0.76	0.76	0.829	73.6
LOG REG	TEST SET	0.91	0.73	0.75	0.82	0.752	74.59
LDA	TRAIN SET	0.77	0.76	0.76	0.76	0.828	76.56
LDA	TEST SET	0.91	0.74	0.75	0.83	0.754	74.63
NBS	TRAIN SET	0.74	0.72	0.68	0.71	0.719	67.93
NBS	TEST SET	0.9	0.66	0.67	0.77	0.657	65.73
KNNS	TRAIN SET	0.99	0.9	0.8	0.89	0.899	84.08
KNNS	TEST SET	0.95	0.74	0.72	0.82	0.764	83.5

MODEL EVALUATION OF DIFFERENT MODELS BUILD FOR OVERALL DATA





VARIABLE	IMPORTANCE
Players_scored_zero	0.192431
extra_bowls_opponent	0.119914
All_rounder_in_team	0.111122
Min_run_given_1over	0.105068
Extra_bowls_bowled	0.104505
Min_run_scored_1over	0.091722
Season	0.080861
Bowlers_in_team	0.060956
player_highest_run	0.028348
Max_wicket_taken_1over	0.022994
Max_run_scored_1over	0.01843
Offshore	0.016262
Opponent	0.012108
Match_format	0.012077
Max_run_given_1over	0.011116
First_selection	0.00706
Match_light_type	0.005025
Wicket_keeper_in_team	0
Avg_team_Age	0

INTERPREATION:

RANDOM FOREST CLASSIFIER IS THE BEST MODEL BUILD ON ENTIRE DATASET.

VARIABLE IMPORTANCE OF 0.1 CAN BE CONSIDERED FOR DEFINING STRATEGIES TO WIN

PERFORMANCE METRICS OF DIFFERENT MODELS BUILD FOR DIFFERENT CLUSTERS

			PERFORMANCE METRICS					
	MODEL BUILDING	DATA	PRECISION	ACCURACY	RECALL	F1 SCORE	AUC	Win %
ODI-MATCH	RFC	TRAIN SET	0.85	0.86	0.87	0.86	0.941	86.84
ODI-MATCH	RFC	TEST SET	0.92	0.81	0.86	0.89	0.722	83.89
ODI-MATCH	NNC	TRAIN SET	0.83	0.82	0.8	0.81	0.888	79.76
ODI-MATCH	NNC	TEST SET	0.94	0.77	0.79	0.86	0.737	77.89

	MODEL BUILDING	DATA	PRECISION	ACCURACY	RECALL	F1 SCORE	AUC	WIN %
T20 MATCH	RFC	TRAIN SET	0.8	0.79	0.79	0.79	0.877	83.44
T20 MATCH	RFC	TEST SET	0.86	0..70	0.74	0.79	0.738	76.92
T20 MATCH	NNC	TRAIN SET	0.89	0.76	0.59	0.71	0.854	80.8
T20 MATCH	NNC	TEST SET	0.9	0.57	0.51	0.65	0.716	76.58
T20 MATCH	KNNS	TRAIN SET	0.99	0.89	0.79	0.88	0.891	85.41
T20 MATCH	KNNS	TEST SET	0.91	0.67	0.66	0.66	0.703	82.64

	MODEL BUILDING	DATA	PRECISION	ACCURACY	RECALL	F1 SCORE	AUC	WIN %
TEST MATCH	LOG REG	TRAIN SET	0.85	0.85	0.85	0.85	0.958	84.85
TEST MATCH	LOG REG	TEST SET	0.91	0.84	0.89	0.9	0.857	87.48
TEST MATCH	NNC	TRAIN SET	0.78	0.8	0.85	0.81	0.893	84.4
TEST MATCH	NNC	TEST SET	0.9	0.75	0.77	0.83	0.825	76.7

PERFORMANCE METRICS OF DIFFERENT MODELS AFTER MODEL TUNING FOR DIFFERENT CLUSTERS

			PERFORMANCE METRICS					
	MODEL BUILDING	DATA	PRECISION	ACCURACY	RECALL	F1 SCORE	AUC	Win %
ODI MATCH	BAGGING	TRAIN SET	1	1	1	1	0.998	99.90
ODI MATCH	BAGGING	TEST SET	0.96	0.95	0.98	0.97	0.824	96.64
ODI MATCH	E RF	TRAIN SET	1	1	1	1	1	100.00
ODI MATCH	E RF	TEST SET	0.96	0.95	0.98	0.97	0.838	96.91
T20 MATCH	BAGGING	TRAIN SET	1	1	1	1	1	100.00
T20 MATCH	BAGGING	TEST SET	0.94	0.95	0.98	0.96	0.868	96.80
T20 MATCH	E RF	TRAIN SET	1	1	1	1	1	100.00
T20 MATCH	E RF	TEST SET	0.95	0.93	0.96	0.95	0.883	95.44
TEST MATCH	GB	TRAIN SET	1	1	1	1	1	100.00
TEST MATCH	GB	TEST SET	0.89	0.86	0.94	0.92	0.749	92.35
TEST MATCH	AB	TRAIN SET	1	1	1	1	1	100.00
TEST MATCH	AB	TEST SET	0.89	0.86	0.94	0.92	0.749	92.35

STRATEGY FOR WINNING TEST MATCH

odds_ratio	variable
1.30E+07	Opponent
9.82E+00	First_selection
3.66E+00	Players_scored_zero
3.24E+00	Min_run_given_1over
1.56E+00	extra_bowls_opponent
1.50E+00	Max_run_scored_1over
1.45E+00	Avg_team_Age
1.32E+00	Extra_bowls_bowled
1.26E+00	Offshore
1.04E+00	Result
1.00E+00	Match_light_type
1.00E+00	All_rounder_in_team
9.69E-01	player_highest_run
7.62E-01	Min_run_scored_1over
7.05E-01	Wicket_keeper_in_team
6.63E-01	Max_wicket_taken_1over
4.22E-01	Max_run_given_1over
2.32E-01	Season
2.19E-01	Bowlers_in_team

BUSINESS RECOMMENDATIONS FOR WINNING TEST MATCH

RECOMMENDATION-1

- India has to opt to bowl first on the offshore match at England if it wins toss. Inning extras must be reduced during the bowling half by the Indian team. Bowlers with little economy rate need to selected as they provide less runs per over to the opponent team.
- If India bats first, based on rainy climate, as there is a chance of DLS, maintaining good runs per over (r. p. o) is necessary. Also, after rains, the second batting becomes tougher. As England pitches are bouncy, selecting more (4 allrounders) and minimising duck-outs i.e., by sending opening batsmen who can stand for a long time at least 20 overs. Maintaining partnerships (at least 2-3) of 100- 120 runs and average of 4-5 runs per over is highly appreciated in the match.
- Practice matches are also preferred as it helps to acclamatise to the offshore climatic conditions at England since no test matches were played with Europe in rainy season based on the dataset.

STRATEGY FOR WINNING T20 MATCH

Variable	Importance
Players_scored_zero	0.220497
Extra_bowls_bowled	0.122009
Opponent	0.108232
Min_run_scored_1over	0.090526
extra_bowls_opponent	0.074193
All_rounder_in_team	0.066362
Bowlers_in_team	0.065253
Max_wicket_taken_1over	0.064699
Season	0.060594
player_highest_run	0.036164
Offshore	0.029334
Max_run_scored_1over	0.028892
Max_run_given_1over	0.021726
First_selection	0.008177
Match_light_type	0.003341
Min_run_given_1over	0
Wicket_keeper_in_team	0
Match_format	0
Avg_team_Age	0

BUSINESS RECOMMENDATIONS FOR WINNING T20 MATCH

RECOMMENDATION-1

- The players must score minimum of 20 runs and no duck-out is mandatory for winning. So, the captain must deploy scoring batsmen as openers, select 3-4 allrounders in team and one all-rounder in the top 4 batting line-up. The runs per over by the team must be maintained around 8-9 at least if batting first.
- The match at times due to fog in winter, the ball becomes slippery and can't slide on ground. So, it is advisable to bat first as Australia is a tough team.

RECOMMENDATION-2

Consider the bowling strategy, it is necessary to minimise the inning extras and pressurise the opponent team by taking key wickets, especially in P1 and P2. The mid overs are to be controlled by spinners and allrounders. More allrounders are generally preferred as in T20 matches, scores fly high.

STRATEGY FOR WINNING ODI MATCH

Variable	Importance
extra_bowls_opponent	0.177686
Extra_bowls_bowled	0.141615
Max_wicket_taken_1over	0.127468
All_rounder_in_team	0.111439
Min_run_scored_1over	0.109006
Players_scored_zero	0.056168
Season	0.052319
Bowlers_in_team	0.04747
Min_run_given_1over	0.041813
player_highest_run	0.04043
Max_run_given_1over	0.034556
Max_run_scored_1over	0.022341
Offshore	0.017851
Opponent	0.014673
First_selection	0.003822
Match_light_type	0.001343
Wicket_keeper_in_team	0
Match_format	0
Avg_team_Age	0

BUSINESS RECOMMENDATIONS FOR WINNING ODI MATCH

RECOMMENDATION 1

- **Provide less bowling extras during the bowling innings. Fast bowlers during the P1 and P3 with good swing are to be selected in the team.**
- **Minimising the opponent runs per over by taking more wickets by selecting excellent spinners with less economy rate.**

RECOMMENDATION 2

- **Selecting 2-3 all-rounders in the team and opt to bat first, if India wins toss due to hard pitch, foggy conditions in the second innings.**
- **Maximum runs must be scored per over if India bats first.**

THANK YOU