greatlearning
Learning for Life

HR ANALYTICS - CTC PREDICTION

USING DATA TO INFORM // TRANSFORM // AND EMPOWER HR DECISIONS

CAPSTONE PROJECT FINAL REPORT

PREPARED BY - RUPESH KUMAR
SUBMISSION DATE: 15/05/2022

Table of Contents

List of Figures

List of Figures

List of Tables

List of Tables

Introduction of the Business Problem

- **Business Problem Statement**

To ensure there is no discrimination between employees, it is imperative for the Human Resources department of Delta Ltd. to maintain a salary range for each employee with similar profiles. Apart from the existing salary, there is a considerable number of factors regarding an employee's experience and other abilities to which they get evaluated in interviews. Given the data related to individuals who applied in Delta Ltd, models can be built that can automatically determine salary which should be offered if the prospective candidate is selected in the company. This model seeks to minimize human judgment with regard to salary to be offered.

- **Executive Summary**

HR Department of Delta Ltd. want to predict a salary range / ctc for applicants with similar profiles apart from the existing salary, there is a considerable number of factors regarding an employee's experience and other abilities to which they get evaluated in interviews.The dataset consists of various attributes of the applicants based their job title/role, Industry, location, years of experience, current ctc , education and skill profile. Based on the different attributes/characteristics the applicants are defined. In this problem statement we will explore the different attributes of applicants like education , total experience , current ctc , role , industry etc and build machine learning model to predict the ctc which company can offer to applicants at time of joining.

The main objective of this problem is to provides salary estimates at time of joining for applicants based on their job title, location, years of experience, and skill profile to minimize human judgment with regard to salary to be offered.

- **Goal & Objective**

Information irregularity amongst employers and employees has become a problem that needs immediate solving. The probable applicants are most often kept blind with regards to the interview procedure and only are aware of it at the end. In the meantime, the employers must be committed to rightly meeting up with the candidate's prospects for making new HR strategies that satisfy the demands of the applicant. Therefore, one must be vigilant enough to not offer too low a salary, which would result in the decline in not just the salary or not offer too high a salary to applicant whose CTC is already as per market range, but also will build more irresponsible, lack-luster individuals with longer untaken positions. Whilst the vice-versa would also be a cause of concern leading to wastage of companies vital resources. Therefore, it is imperative to provide an unbiased salary for an employee which he/she truly deserves, and also has to be appropriate to the market demands.

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis , visualization like univariate , bivariate and multivariate to check the distribution and realtionship between the given attributes. & apply supervised machine learning algorithms i.e. Linear Regression ,XG-Boost Regressor , Decision Tree Regressor , Random Forest Regressor and ANN Regressor to predict the correct salary/ctc range for applicants on the basis of the given information, which will help company to offer correct ctc/salary range to applicants at time of joining plus this model reduced the manual judgement on selecting the ctc/salary range.It is intended to have a robust approach and eliminate any discrimination in salary among similar employee profiles.

Explore the dataset using central tendency and other parameters. The data consists of 25000 different applicants with 29 unique features . Analyse the different attributes of the applicants which can help company in building a machine learning model to predic. the ctc offer to applicant at time of joining.This assignment should help the company to take right judgment with regard to salary to be offered.

Data Description (EDA and Business Implication)

Checking the Records of the Dataset :

Head of the Dataset - First 10 Records of the Dataset.
Tail of the Dataset - Last 10 Records of the Dataset.

	IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	Graduation_Specialization	University_Grad
0	1	22753	0	0	NaN	NaN	NaN	NaN	NaN	PG	Arts	Lucknow
1	2	51087	23	14	HR	Consultant	Analytics	H	HR	Doctorate	Chemistry	Surat
2	3	38413	21	12	Top Management	Consultant	Training	J	NaN	Doctorate	Zoology	Jaipur
3	4	11501	15	8	Banking	Financial Analyst	Aviation	F	HR	Doctorate	Others	Bangalore
4	5	58941	10	5	Sales	Project Manager	Insurance	E	Medical Officer	Grad	Zoology	Mumbai
5	6	30564	16	3	Top Management	Area Sales Manager	Retail	G	Director	Doctorate	Others	Bangalore
6	7	27267	1	1	Engineering	Team Lead	FMCG	L	Marketing Manager	Grad	Chemistry	Delhi
7	8	36521	19	11	Others	Analyst	Others	E	Manager	PG	Sociology	Delhi
8	9	11616	8	7	Analytics/BI	Others	Telecom	L	Marketing Manager	Doctorate	Psychology	Mumbai
9	10	43886	15	15	Analytics/BI	CEO	Telecom	M	Product Manager	Doctorate	Chemistry	Delhi

	IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	Graduation_Specialization	University_Grad
24990	24991	34589	22	1	Top Management	Sr. Business Analyst	Analytics	D	Manager	Under Grad	NaN	NaN
24991	24992	13280	1	1	Sales	Consultant	IT	A	Marketing Manager	Under Grad	NaN	NaN
24992	24993	35325	25	12	Sales	Sales Manager	Automobile	D	Data Analyst	PG	Statistics	Bangalore
24993	24994	31883	15	13	Healthcare	Consultant	Aviation	P	Manager	PG	Statistics	Mangalore
24994	24995	32035	7	3	Top Management	Others	Telecom	A	Data Analyst	Under Grad	NaN	NaN
24995	24996	25550	18	13	Engineering	Project Manager	Automobile	I	Assistant Manager	PG	Psychology	Surat
24996	24997	53442	12	8	HR	Others	Analytics	B	Sr.Manager	Under Grad	NaN	NaN
24997	24998	15777	22	8	Banking	Head	Insurance	D	Software Developer	Under Grad	NaN	NaN
24998	24999	57616	25	8	Marketing	CEO	BFSI	D	Marketing Manager	PG	Economics	Surat
24999	25000	20788	8	0	Banking	Consultant	Automobile	P	Sr.Manager	Grad	Economics	Bangalore

Tab:1 Records of the Dataset Head & Tail

Note: We are dropping the column IDX and Applicant_Id as these columns didn't contribute for analysis and model building exercise , because IDX and Applicant_ID for each applicant is unique and increases the cardinality, variables with high cardinality isn't preferred , hence it is useless for the model.That's why we decided to drop these two columns.

	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	Graduation_Specialization	University_Grad	Passing_Year_Of_Graduation
0	0	0	NaN	NaN	NaN	NaN	NaN	PG	Arts	Lucknow	2020.0
1	23	14	HR	Consultant	Analytics	H	HR	Doctorate	Chemistry	Surat	1988.0
2	21	12	Top Management	Consultant	Training	J	NaN	Doctorate	Zoology	Jaipur	1990.0
3	15	8	Banking	Financial Analyst	Aviation	F	HR	Doctorate	Others	Bangalore	1997.0
4	10	5	Sales	Project Manager	Insurance	E	Medical Officer	Grad	Zoology	Mumbai	2004.0
5	16	3	Top Management	Area Sales Manager	Retail	G	Director	Doctorate	Others	Bangalore	1998.0
6	1	1	Engineering	Team Lead	FMCG	L	Marketing Manager	Grad	Chemistry	Delhi	2011.0
7	19	11	Others	Analyst	Others	E	Manager	PG	Sociology	Delhi	2001.0
8	8	7	Analytics/BI	Others	Telecom	L	Marketing Manager	Doctorate	Psychology	Mumbai	2003.0
9	15	15	Analytics/BI	CEO	Telecom	M	Product Manager	Doctorate	Chemistry	Delhi	1998.0

Tab:2 Records of the Dataset After Dropping Unwanted Columns

Observation:

- Now we have all the columns which are useful for the model.
- We changed name of Education column to Highest Education because it looks for appropriate in terms of readability & clearly give idea about the educational background of the applicant.
- Second we rename the Curent Location column to Current location because there is spelling error.

Data Dictionary for Business Problem Statement.

Target variable: Expected_CTC

Data dictionary:

IDX	Index
Applicant_ID	Application ID
Total_Experience	Total industry experience
Total_Experience_in_field_applied	Total experience in the field applied for (past work experience that is relevant to the job)
Department	Department name of current company
Role	Role in the current company
Industry	Industry name of current field
Organization	Organization name
Designation	Designation in current company
Education	Education
Graduation_Specialization	Specialization subject in graduation
University_Grad	University or college in Graduation
Passing_Year_Of_Graduation	Year of passing Graduation
PG_Specialization	Specialization subject in Post-Graduation
University_PG	University or college in Post-Graduation
Passing_Year_Of_PG	Year of passing Post Graduation
PHD_Specialization	Specialization subject in Post-Graduation
University_PHD	University or college in Post Doctorate
Passing_Year_Of_PHD	Year of passing PHD
Curent_Location	Curent Location
Preferred_location	Preferred location to work in the company applied
Current_CTC	Current CTC
Inhand_Offer	Holding any offer in hand (Y: Yes, N:No)
Last_Appraisal_Rating	Last Appraisal Rating in current company
No_Of_Companies_worked	No. of companies worked till date
Number_of_Publications	Number of papers published
Certifications	Number of relevant certifications completed
International_degree_any	Hold any international degree (1: Yes, 0: No)
Expected_CTC	Expected CTC (Final CTC offered by Delta Ltd.)

Tab:3 Data Dictionary For Business Problem Statement

Checking the Summary of the Dataset :

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Total_Experience	25000.0	NaN	NaN	NaN	12.49308	7.471398	0.0	6.0	12.0	19.0	25.0
Total_Experience_in_field_applied	25000.0	NaN	NaN	NaN	6.2582	5.819513	0.0	1.0	5.0	10.0	25.0
Department	22222	12	Marketing	2379	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Role	24037	24	Others	2248	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Industry	24092	11	Training	2237	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Organization	24092	16	M	1574	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Designation	21871	18	HR	1648	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Highest Education	25000	4	PG	6326	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Graduation_Specialization	18820	11	Chemistry	1785	NaN	NaN	NaN	NaN	NaN	NaN	NaN
University_Grad	18820	13	Bhubaneswar	1510	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Passing_Year_Of_Graduation	18820.0	NaN	NaN	NaN	2002.193624	8.31664	1986.0	1996.0	2002.0	2009.0	2020.0
PG_Specialization	17308	11	Mathematics	1800	NaN	NaN	NaN	NaN	NaN	NaN	NaN
University_PG	17308	13	Bhubaneswar	1377	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Passing_Year_Of_PG	17308.0	NaN	NaN	NaN	2005.153571	9.022963	1988.0	1997.0	2006.0	2012.0	2023.0
PHD_Specialization	13119	11	Others	1545	NaN	NaN	NaN	NaN	NaN	NaN	NaN
University_PHD	13119	13	Kolkata	1069	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Passing_Year_Of_PHD	13119.0	NaN	NaN	NaN	2007.396372	7.493601	1995.0	2001.0	2007.0	2014.0	2020.0
Current_location	25000	15	Bangalore	1742	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Preferred_location	25000	15	Kanpur	1720	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Current_CTC	25000.0	NaN	NaN	NaN	1760945.38388	920212.512479	0.0	1027311.5	1802567.5	2443883.25	3999693.0
Inhand_Offer	25000	2	N	17418	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Last_Appraisal_Rating	24092	5	B	5501	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_Of_Companies_worked	25000.0	NaN	NaN	NaN	3.48204	1.690335	0.0	2.0	3.0	5.0	6.0
Number_of_Publications	25000.0	NaN	NaN	NaN	4.08904	2.606612	0.0	2.0	4.0	6.0	8.0
Certifications	25000.0	NaN	NaN	NaN	0.77368	1.199449	0.0	0.0	0.0	1.0	5.0
International_degree_any	25000.0	NaN	NaN	NaN	0.08172	0.273943	0.0	0.0	0.0	0.0	1.0
Expected_CTC	25000.0	NaN	NaN	NaN	2250154.5104	1160480.144938	203744.0	1306277.5	2252136.5	3051353.75	5599570.0

Tab:4 Summary of the Dataset

Observations

- From the above table we can infer the count,mean, std , 25% , 50% ,75% and min & max values of the all numeric variables present in the dataset.
- From the above table we get the count,unique,top,freq of all the categorical variables present in the dataset.

Checking the Shape of the Dataframe :

No. of Rows	No. of Columns
25000	27

Tab:5 Shape of the Dataframe

Insights -

Shape attribute tells us number of observations and variables we have in the data set. It is used to check the dimension of data. The expected_ctc.csv data set has 25000 observations (rows) and 27 variables (columns) in the dataset.

Checking the Appropriateness of Datatypes & Information of the Dataframe :

The info() function is used to print a concise summary of a DataFrame. This method prints information about a DataFrame including the index d-type and column d-types, non-null values and memory usage.

S.no.	Features / Columns	Non-Null Count	D-Type
0	Total_Experience	25000 non-null	int64
1	Total_Experience_in_field_applied	25000 non-null	int64
2	Department	22222 non-null	object
3	Role	24037 non-null	object
4	Industry	24092 non-null	object
5	Organization	24092 non-null	object
6	Designation	21871 non-null	object
7	Highest Education	25000 non-null	object
8	Graduation_Specialization	18820 non-null	object
9	University_Grad	18820 non-null	object
10	Passing_Year_Of_Graduation	18820 non-null	float64
11	PG_Specialization	17308 non-null	object
12	University_PG	17308 non-null	object
13	Passing_Year_Of_PG	17308 non-null	float64
14	PHD_Specialization	13119 non-null	object
15	University_PHD	13119 non-null	object
16	Passing_Year_Of_PHD	13119 non-null	float64
17	Current_location	25000 non-null	object
18	Preferred_location	25000 non-null	object
19	Current_CTC	25000 non-null	int64
20	Inhand_Offer	25000 non-null	object
21	Last_Appraisal_Rating	24092 non-null	object
22	No_Of_Companies_worked	25000 non-null	int64
23	Number_of_Publications	25000 non-null	int64
24	Certifications	25000 non-null	int64
25	International_degree_any	25000 non-null	int64
26	Expected_CTC	25000 non-null	int64

Tab: 6 Appropriateness of Datatypes & Information of the Dataframe

Insights -

From the above results we can see that there are null values present in most of the columns
like(**Department,Role,Industry,Organization,Designation,Graduation_Specialization ,University_Grad Passing_Year_Of_Graduation , PG_Specialization ,University_PG , Passing_Year_Of_PG and PHD_Specialization , University_PHD and Passing_Year_Of_PHD etc**) of the dataset.Their are total 25000 rows & 27 columns given in this dataset,indexed from 0 to 24999. Out of 27 variables 3 are float64 , 16 variables are object and 8 variable are int64 d-type. Memory used by the dataset: 5.1+ MB.

Skewness of the Dataset :

In statistics, skewness is a measure of asymmetry of the probability distribution about its mean and helps describe the shape of the probability distribution. Basically it measures the level of how much a given distribution is different from a normal distribution (which is symmetric).

Skewness is a well-established statistical concept for continuous and to a lesser extent for discrete quantitative statistical variables.Here we are going to check the skewness of the features which are present in our dataset.

S.no.	Features / Columns	Skewness
1	Total_Experience	0.004109
2	Total_Experience_in_field_applied	0.961951
3	Passing_Year_Of_Graduation	0.061408
4	Passing_Year_Of_PG	-0.066166
5	Passing_Year_Of_PHD	0.014436
6	Current_CTC	0.097643
7	Expected_CTC	0.331972

Tab: 7 Skewness of the Dataset

Insights -

- From the above result, we can check which variable is normally distributed and which is not.
- The variables with skewness > 1 are highly positively skewed.
- The variables with skewness < -1 are highly negatively skewed.
- The variables with 0.5 < skewness < 1 are moderately positively skewed.
- The variables with -0.5 < skewness < -1 such as stroke are moderately negatively skewed.
- And, the variables with -0.5 < skewness < 0.5 are symmetric i.e normally distributed

Checking for Null Values :

S.No.	Features / Columns	Null Count
1	Total_Experience	0
2	Total_Experience_in_field_applied	0
3	Department	2778
4	Role	963
5	Industry	908
6	Organization	908
7	Designation	3129
8	Highest Education	0
9	Graduation_Specialization	6180
10	University_Grad	6180
11	Passing_Year_Of_Graduation	6180
12	PG_Specialization	7692
13	University_PG	7692
14	Passing_Year_Of_PG	7692
15	PHD_Specialization	11881
16	University_PHD	11881
17	Passing_Year_Of_PHD	11881
18	Current_location	0
19	Preferred_location	0
20	Current_CTC	0
21	Inhand_Offer	0
22	Last_Appraisal_Rating	908
23	No_Of_Companies_worked	0
24	Number_of_Publications	0
25	Certifications	0
26	International_degree_any	0
27	Expected_CTC	0

Tab:8 Checking Null Values.

Insights -

From the above output we found that most of the columns have null values. Graduation_Specialization , University_Grad , Passing_Year_Of_Graduation , PG_Specialization ,University_PG , Passing_Year_Of_PG and PHD_Specialization , University_PHD and Passing_Year_Of_PHD have some kind of pattern for null values may be these applicants didn't have UG , PG or PhD degree or data not available will impute these missing values with suitable imputation method down the line.

Checking for Anomalies in the Dataset :

Total_Experience

array([0, 23, 21, 15, 10, 16, 1, 19, 8, 13, 7, 12, 20, 4, 14, 17, 22, 3, 5, 24, 2, 25, 9, 6, 11, 18])

Total_Experience_in_field_applied

array([0, 14, 12, 8, 5, 3, 1, 11, 7, 15, 10, 9, 4, 6, 2, 20, 16,25, 13, 19, 21, 22, 23, 17, 18, 24])

Department

array([nan, 'HR', 'Top Management', 'Banking', 'Sales', 'Engineering', 'Others', 'Analytics/BI', 'Education', 'Marketing', 'Healthcare', 'IT-Software', 'Accounts'], dtype=object)

Role

array([nan, 'Consultant', 'Financial Analyst', 'Project Manager', 'Area Sales Manager', 'Team Lead', 'Analyst', 'Others', 'CEO', 'Business Analyst', 'Sales Manager', 'Bio statistician', 'Scientist', 'Research Scientist', 'Head', 'Associate', 'Senior Researcher', 'Sales Execituve', 'Sr. Business Analyst','Principal Analyst', 'Data scientist', 'Researcher', 'Senior Analyst', 'Professor', 'Lab Executuve'], dtype=object)

Industry

array([nan, 'Analytics', 'Training', 'Aviation', 'Insurance', 'Retail', 'FMCG', 'Others', 'Telecom', 'Automobile', 'IT', 'BFSI'], dtype=object)

Organization

array([nan, 'H', 'J', 'F', 'E', 'G', 'L', 'M', 'O', 'D', 'N', 'A', 'B', 'I', 'K', 'P', 'C'], dtype=object)

Designation

array([nan, 'HR', 'Medical Officer', 'Director', 'Marketing Manager', 'Manager', 'Product Manager', 'Consultant', 'CA','Research Scientist', 'Sr.Manager', 'Data Analyst', 'Assistant Manager', 'Others', 'Web Designer', 'Research Analyst', 'Software Developer', 'Network Engineer', 'Scientist'], dtype=object)

Highest Education

array(['PG', 'Doctorate', 'Grad', 'Under Grad'], dtype=object)

Graduation_Specialization

array(['Arts', 'Chemistry', 'Zoology', 'Others', 'Sociology', 'Psychology', 'Mathematics', nan, 'Engineering', 'Botony', 'Statistics', 'Economics'], dtype=object)

University_Grad

array(['Lucknow', 'Surat', 'Jaipur', 'Bangalore', 'Mumbai', 'Delhi', 'Mangalore', nan, 'Nagpur', 'Kolkata', 'Ahmedabad', 'Guwahati', 'Pune', 'Bhubaneswar'], dtype=object)

Passing_Year_Of_Graduation

array([2020., 1988., 1990., 1997., 2004., 1998., 2011., 2001., 2003.,2000., nan, 2012., 2002., 2016., 2013., 1999., 1993., 2009., 1989., 1991., 2008., 2005., 2018., 1992., 1996., 2010., 2019., 1986., 2007., 2015., 1995., 2006., 2014., 1987., 2017., 1994.])

PG_Specialization

array([nan, 'Others', 'Zoology', 'Chemistry', 'Psychology', 'Mathematics','Engineering', 'Sociology', 'Arts', 'Statistics', 'Economics','Botony'], dtype=object)

University_PG

array([nan, 'Surat', 'Jaipur', 'Bangalore', 'Mumbai', 'Delhi', 'Mangalore', 'Nagpur', 'Kolkata', 'Lucknow', 'Ahmedabad', 'Guwahati', 'Pune', 'Bhubaneswar'], dtype=object)

Passing_Year_Of_PG

array([nan, 1990., 1992., 1999., 2006., 2000., 2013., 2005., 2002., 2014., 2004., 2009., 2017., 2001., 1995., 2011., 1991., 1993.,2003., 2007., 2010., 1994., 2020., 2016., 1998., 2012., 2022.,1988., 2019., 2018., 1997., 2008., 2015., 1989., 2021., 1996.,2023.])

Current_location

array(['Guwahati', 'Bangalore', 'Ahmedabad', 'Kanpur', 'Pune', 'Delhi','Surat', 'Nagpur', 'Jaipur', 'Kolkata', 'Bhubaneswar', 'Mangalore', 'Mumbai', 'Lucknow', 'Chennai'], dtype=object)

Preferred_location

array(['Pune', 'Nagpur', 'Jaipur', 'Kolkata', 'Ahmedabad', 'Bhubaneswar', 'Bangalore', 'Guwahati', 'Mangalore', 'Kanpur', 'Mumbai','Chennai', 'Surat', 'Delhi', 'Lucknow'], dtype=object)

Current_CTC

array([0, 2702664, 2236661, ..., 1681796, 3311090, 935897])

Inhand_Offer

array(['N', 'Y'], dtype=object)

Last_Appraisal_Rating

array([nan, 'Key_Performer', 'C', 'B', 'A', 'D'], dtype=object)

No_Of_Companies_worked

array([0, 2, 5, 3, 6, 4, 1])

Number_of_Publications

array([0, 4, 3, 1, 6, 8, 2, 7, 5])

Certifications

array([0, 1, 5, 2, 4, 3])

International_degree_any

array([0, 1])

Expected_CTC

array([384551, 3783729, 3131325, ..., 1934065, 4370638, 1216666])

Tab:9 Checking Anomalies for Variables in the Dataset

Insights

There is no Anomalies present in the dataset , but have nan values in many of columns.

Checking Duplicate Values :

Observation -There is no duplicates rows present in the dataset.

Checking the Value counts on all the Categorical Column.

S.no.	Department	Value_Counts
1	Marketing	2379
2	Analytics/BI	2096
3	Healthcare	2062
4	Others	2041
5	Sales	1991
6	HR	1988
7	Banking	1952
8	Education	1948
9	Engineering	1937
10	Top Management	1632
11	Accounts	1118
12	IT-Software	1078

Tab:10 Value Counts for Categorical Feature (Department)

Insights -

- There are 12 types of Department present in the data named as 'Marketing' , 'Analytics/BI' , 'Healthcare' , 'Others' , 'Sales' , 'HR' ,'Banking' , 'Education' ,'Engineering' , 'Top Management' ,'Accounts' and 'IT-Software'.
- Majority of applicants are of Marketing Department (2379) ,Analytics/BI (2096) , Health-care (2062) and others (2041).
- Least applicants belongs to IT-Software Department(1078).
- There is fine distribution of applicants in each department.

S.no.	Role	Value_Counts
1	Others	2248
2	Bio statistician	1913
3	Analyst	1892
4	Project Manager	1850
5	Team Lead	1833
6	Consultant	1780
7	Business Analyst	1711
8	Sales Execituve	1574
9	Sales Manager	1427
10	Senior Researcher	1236
11	Financial Analyst	1182
12	CEO	1149
13	Scientist	1139
14	Head	1108
15	Associate	767
16	Data scientist	363
17	Principal Analyst	275
18	Area Sales Manager	134
19	Senior Analyst	128
20	Researcher	123
21	Sr. Business Analyst	114
22	Professor	33
23	Research Scientist	33
24	Lab Executuve	25

Tab:11 Value Counts for Categorical Feature (Role)

Insights -

- There are 24 types of Roles present in the data named as 'Others' , 'Bio-statistician' ,'Analyst' , 'Project Manager' , 'Team Lead' ,'Consultant' ,'Business Analyst ' ,'Sales Execituve' , 'Sales Manager' , 'Senior Researcher' , 'Financial Analyst' , 'CEO' , 'Scientist' , 'Head' , 'Associate' ,'Data scientist' ,'Principal Analyst' , 'Area Sales Manager' ,'Senior Analyst' ,'Researcher' , 'Sr. Business Analyst' , 'Professor' ,'Research Scientist' and 'Lab Executuve'.
- Majority of applicants worked as others(Role) i.e (2248).
- Minority of the applicants worked as Lab executive (Role) i.e.(25).

S.no.	Industry	Value_Counts
1	Training	2237
2	IT	2228
3	Insurance	2219
4	BFSI	2207
5	Automobile	2202
6	Analytics	2201
7	Retail	2195
8	Telecom	2190
9	Aviation	2183
10	FMCG	2180
11	Others	2050

Tab:12 Value Counts for Categorical Feature (Industry)

Insights -

- There are 11 types of Industries present in the data named as 'Training' , 'IT' ,'Insurance' ,'BFSI' ,'Automobile' , 'Analytics' , 'Retail' ,'Telecom' ,'Aviation' ,'FMCG' and 'Others'.
- Majority of applicants worked in Training Industry.
- There is not much variation in the Industry column have a fair distribution for all applicants.

S.no.	Organization	Value_Counts
1	M	1574
2	J	1555
3	P	1542
4	H	1532
5	A	1526
6	F	1505
7	G	1504
8	K	1503
9	I	1489
10	E	1488
11	B	1488
12	L	1484
13	C	1482
14	N	1476
15	D	1474
16	O	1470

Tab:13 Value Counts for Categorical Feature (Organisation)

Insights -

- There are 16 types of Organization name present in the dataset named as 'M' ,'J' ,'P' ,'H' , 'A ' , 'F ' ,'G' ,'K' ,'I' ,'E' ,'B' ,'L' ,'C' ,'N' , 'D' and 'O'.
- There is not too much variations in the Organization columns equivalent number of applicants worked in 16 different organization.

S.no.	Designation	Value_Counts
1	HR	1648
2	Others	1647
3	Manager	1628
4	Product Manager	1626
5	Sr.Manager	1617
6	Consultant	1606
7	Marketing Manager	1590
8	Assistant Manager	1590
9	Data Analyst	1575
10	Research Analyst	1563
11	Medical Officer	1047
12	Software Developer	914
13	Web Designer	882
14	Network Engineer	862
15	Director	772
16	CA	715
17	Research Scientist	537
18	Scientist	52

Tab:14 Value Counts for Categorical Feature (Designation)

Insights -

- There are 18 types of Designation present in the dataset named as 'HR' , 'Others' ,'Manager' , 'Product Manager' ,'Sr.Manager' ,'Consultant' ,'Marketing Manager' ,'Assistant Manager' , 'Data Analyst' ,'Research Analyst','Medical Officer' ,'Software Developer' , 'Web Designer' .'Network Engineer' , 'Director' ,'CA' ,'Research Scientist' and 'Scientist'.
- Majority of applicants worked as HR i.e.(1648). Only 52 applicants worked as Scientist.
- There is fine distribution of applicants across various Designation

S.no.	Highest Education	Value_Counts
1	Under Grad	6326
2	Grad	6285
3	PG	6209
4	Doctorate	6180

Tab:15 Value Counts for Categorical Feature (Highest Education)

Insights -

- There are 4 types of Education labels present in the dataset named as 'Under Grad' ,'Grad' ,'PG' and Doctorate.
- 6180 applicants are Under Graduate.
- 6209 applicants are Graduate.
- 6326 applicants are Post Graduate
- 6285 applicants are Doctorate.

S.no.	Graduation_Specialization	Value_Counts
1	Chemistry	1785
2	Economics	1774
3	Mathematics	1770
4	Zoology	1730
5	Arts	1721
6	Psychology	1705
7	Sociology	1697
8	Botony	1674
9	Engineering	1661
10	Others	1660
11	Statistics	1643

Insights -

- There are 11 types of Graduation_Specialization labels present in the dataset named as 'Chemistry' , 'Economics ' , 'Mathematics' ,'Zoology' ,'Arts' ,'Psychology' , 'Sociology' , 'Botony' ,'Engineering' , 'Others' and 'Statistics'.
- Majority of applicants did their graduation specialization in chemistry.
- There is fine distribution of applicants in each Graduation_Specialization label.

Tab:16 Value Counts for Categorical Feature (Graduation_Specialization)

S.no.	University_Grad	Value_Counts
1	Bhubaneswar	1510
2	Delhi	1492
3	Manglore	1490
4	Mumbai	1488
5	Jaipur	1478
6	Lucknow	1457
7	Guwahati	1449
8	Pune	1428
9	Kollkata	1426
10	Surat	1424
11	Nagpur	1420
12	Bangalore	1394
13	Ahmedabad	1364

Insights -

- There are 13 types of University_Grad labels present in the dataset named as 'Bhubaneswar' ,'Delhi' ,'Mangalore' ,'Mumbai' ,'Jaipur' ,'Lucknow' ,'Guwahati' ,'Pune' ,'Kolkata' ,'Surat' ,'Nagpur' ,'Bangalore' and 'Ahmedabad'.
- Majority of applicants did their graduation from Bhubaneswar University (1510) and Delhi University(1492).
- There is equivalent number of graduates from all universities.

Tab:17 Value Counts for Categorical Feature (University_Grad)

S.no.	PG_Specialization	Value_Counts
1	Mathematics	1800
2	Chemistry	1796
3	Economics	1755
4	Engineering	1674
5	Statistics	1639
6	Others	1629
7	Psychology	1425
8	Zoology	1424
9	Arts	1410
10	Sociology	1385
11	Botony	1371

Insights -

- There are 11 types of PG_Specialization labels present in the dataset named as 'Mathematics' , 'Chemistry' ,'Economics' ,'Engineering' ,'Statistics' , 'Others' , 'Psychology' , 'Zoology' , 'Arts' , 'Sociology' and 'Botony'.
- Majority of applicants did their post-graduation specialization in mathematics (1800) and chemistry (1796).
- There is fine distribution of applicants in each PG_Specialization label.

Tab:18 Value Counts for Categorical Feature(PG_Specialization)

S.no.	University_PG	Value_Counts
1	Bhubaneswar	1377
2	Delhi	1368
3	Manglore	1367
4	Mumbai	1366
5	Jaipur	1359
6	Guwahati	1340
7	Surat	1329
8	Lucknow	1328
9	Pune	1314
10	Nagpur	1313
11	Kolkata	1306
12	Bangalore	1287
13	Ahmedabad	1254

Tab:19 Value Counts for Categorical Feature (University_PG)

Insights -

- There are 13 types of University_PG labels present in the dataset named as 'Bhubaneswar' ,'Delhi' ,'Mangalore' ,'Mumbai' ,'Jaipur' ,'Lucknow' ,'Guwahati' ,'Pune' ,'Kolkata' ,'Surat' ,'Nagpur' ,'Bangalore' and 'Ahmedabad'.
- Majority of applicants did their post-graduation from Bhubaneswar University (1377) and Delhi University(1368).
- There is equivalent number of post-graduates from all universities.Distribution /Frequency is nearly same.

S.no.	PHD_Specialization	Value_Counts
1	Others	1545
2	Chemistry	1458
3	Mathematics	1378
4	Economics	1343
5	Engineering	1259
6	Statistics	1236
7	Zoology	1011
8	Sociology	989
9	Psychology	986
10	Botony	976
11	Arts	938

Tab:20 Value Counts for Categorical Feature (PHD_Specialization)

Insights -

- There are 11 types of PHD_Specialization labels present in the dataset named as 'Others' ,'Chemistry','Mathematics','Economics' ,'Engineering' ,'Statistics' ,'Zoology','Sociology' ,'Psychology ' ,'Botony' and 'Arts'.
- Majority of applicants did their PhD specialization in others (1545) and chemistry (1458).

S.no.	University_PHD	Value_Counts
1	Kolkata	1069
2	Delhi	1064
3	Mumbai	1046
4	Guwahati	1030
5	Pune	1011
6	Surat	1009
7	Jaipur	998
8	Lucknow	995
9	Bangalore	994
10	Bhubaneswar	992
11	Manglore	988
12	Nagpur	964
13	Ahmedabad	959

Tab:21 Value Counts for Categorical Feature (University_PHD)

Insights -

- There are 13 types of University_PHD labels present in the dataset named as 'Bhubaneswar' ,'Delhi' ,'Mangalore' ,'Mumbai' ,'Jaipur' ,'Lucknow' ,'Guwahati' ,'Pune' ,'Kolkata' ,'Surat' ,'Nagpur' ,'Bangalore' and 'Ahmedabad' .
- Majority of applicants did their PhD from Kolkata University (1069) and Delhi University(1064).
- There is equivalent number of PhD applicants from all universities.

S.no.	Current_location	Value_Counts
1	Bangalore	1742
2	Jaipur	1706
3	Bhubaneswar	1704
4	Manglore	1697
5	Delhi	1680
6	Ahmedabad	1677
7	Guwahati	1672
8	Chennai	1669
9	Kanpur	1664
10	Nagpur	1663
11	Mumbai	1658
12	Lucknow	1637
13	Pune	1622
14	Kolkata	1620
15	Surat	1589

Tab:22 Value Counts for Categorical Feature (Current_location)

Insights -

- There are 15 types of Curent_Location labels present in the dataset named as 'Bangalore' , 'Jaipur' ,'Bhubaneswar' , 'Mangalore' ,'Delhi' ,'Ahmedabad' ,'Guwahati' ,'Chennai' ,'Kanpur' ,'Nagpur' ,'Mumbai' ,'Lucknow' ,'Pune' ,'Kolkata' and 'Surat'.
- Majority of applicant's current location is Bangalore i.e.(1742).
- Rest there is fair distribution of frequency for applicant's current location for all locations

S.no.	Preferred_location	Value_Counts
1	Kanpur	1720
2	Ahmedabad	1715
3	Guwahati	1695
4	Manglore	1694
5	Surat	1693
6	Delhi	1683
7	Chennai	1680
8	Kolkata	1669
9	Jaipur	1659
10	Pune	1654
11	Bhubaneswar	1653
12	Nagpur	1650
13	Mumbai	1617
14	Lucknow	1612
15	Bangalore	1606

Tab:23 Value Counts for Categorical Feature (Preferred_location)

Insights -

- There are 15 types of Preferred_location labels present in the dataset named as 'Bangalore' , 'Jaipur' ,'Bhubaneswar' , 'Mangalore' ,'Delhi' ,'Ahmedabad' ,'Guwahati' ,'Chennai' ,'Kanpur' ,'Nagpur' ,'Mumbai' ,'Lucknow' ,'Pune' ,'Kolkata' and 'Surat'.
- Majority of applicant's preferred location is Kanpur i.e.(1720).
- Rest there is fair distribution of frequency for applicant's preferred location for all locations.

S.no.	Inhand_Offer	Value_Counts
1	N	17418
2	Y	7582

Tab:24 Value Counts for Categorical Feature (Inhand_Offer)

Insights -

- There are 2 types of Inhand_Offer labels present in the dataset named as 'Y'(Yes) and 'N' (No).
- 17418 applicants don't have In-hand job offer while 7582 applicants have In-hand job offer.

S.no.	Last_Appraisal_Rating	Value_Counts
1	B	5501
2	D	4917
3	C	4812
4	A	4671
5	Key_Performer	4191

Tab:25 Value Counts for Categorical Feature (Last_Appraisal_Rating)

Insights -

- There are 5 types of Last_Appraisal_Rating labels present in the dataset named as 'Key_Performer' , 'A' ,'B' ,'C' and 'D'.
- 4191 applicants Last_Appraisal_Rating is Key_Performer.
- 4671 applicants Last_Appraisal_Rating is A.
- 5501 applicants Last_Appraisal_Rating is B.
- 4812 applicants Last_Appraisal_Rating is C.
- 4917 applicants Last_Appraisal_Rating is D.

Data Visualization

Univariate Analysis of continuous Numerical Variables.

A histogram takes as input a numeric variable only. The variable is cut into several bins, and the number of observation per bin is represented by the height of the bar. It is possible to represent the distribution of several variable on the same axis using this technique.

A box-plot gives a nice summary of one or several numeric variables. The line that divides the box into 2 parts represents the median of the data. The end of the box shows the upper and lower quartiles. The extreme lines show the highest and lowest value excluding outliers.

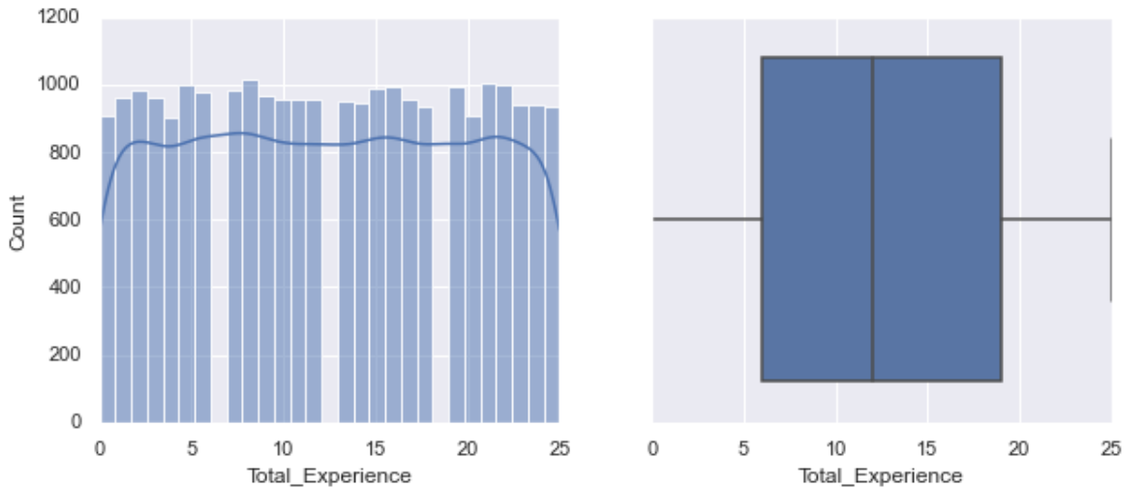


Fig: 1 Histogram & Box-Plot of Total_Experience

Statistical Summary	Values
Count	25000
Mean	12.49
Std	7.47
Min	0.0
25%	6.0
50%	12.0
75%	19.0
Max	25.0

Tab: 26 Statistical Description of Total_Experience

Insights -

- Total_Experience - Total industry experience ranges from a minimum of 0 to maximum of 25.
- Average Total_Experience is around 12.49.
- The standard deviation of Total_Experience is 7.47.
- 25% , 50% (median) and 75 % of Total_Experience are 6 , 12 and 19.
- Total_Experience don't have outliers.

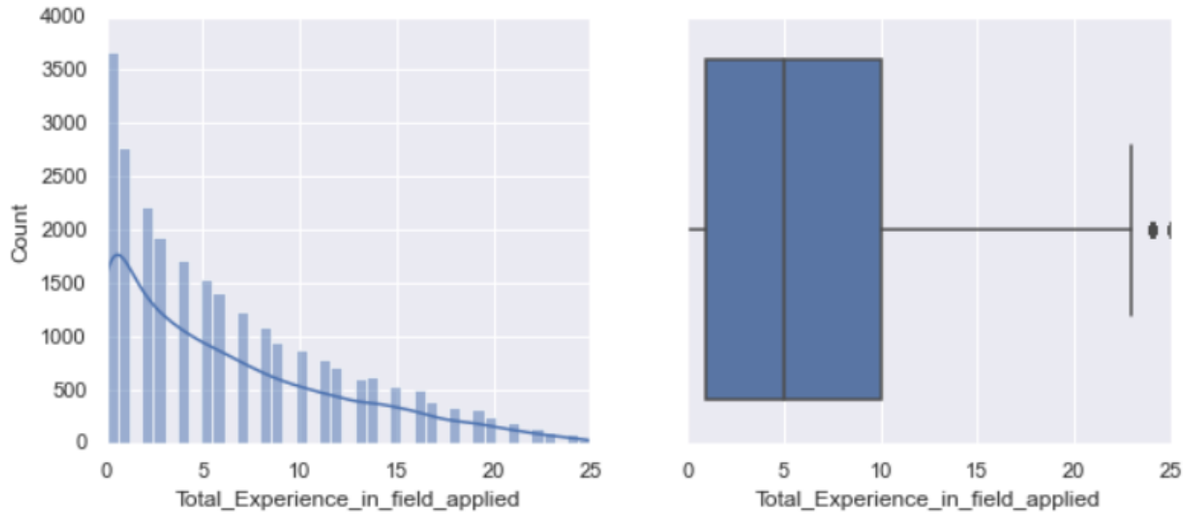


Fig: 2 Histogram & Box-Plot of Total_Experience_in_field_applied

Statistical Summary	Values
Count	25000
Mean	6.25
Std	5.81
Min	0.0
25%	1.0
50%	5.0
75%	10.0
Max	25.0

Tab: 27 Statistical Description of Total_Experience_in_field_applied

Insights -

- Total_Experience_in_field_applied - Total experience in the field applied for (past work experience that is relevant to the job) ranges from a minimum of 0 to maximum of 25.
- Average Total_Experience_in_field_applied is around 6.25.
- The standard deviation of Total_Experience_in_field_applied is 5.81.
- 25% , 50% (median) and 75 % of Total_Experience_in_field_applied are 1 , 5 and 10.
- Total_Experience_in_field_applied have a few outliers.

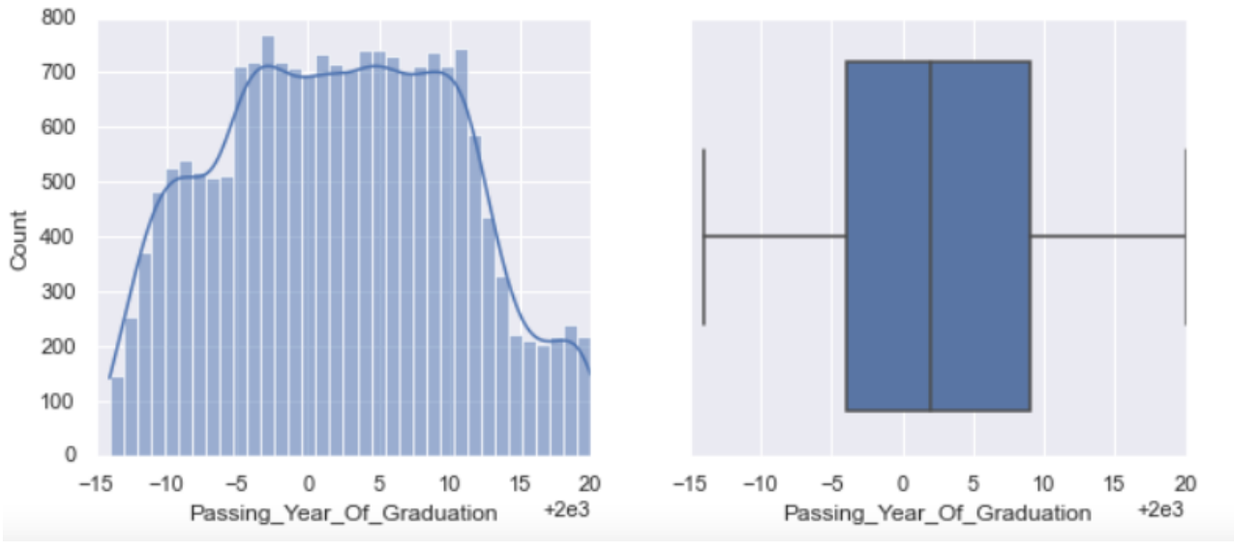


Fig: 3 Histogram & Box-Plot of Passing_Year_Of_Graduation

Statistical Summary	Values
Count	18820
Mean	2020.19
Std	8.31
Min	1986
25%	1996
50%	2002
75%	2009
Max	2020

Tab: 28 Statistical Description of Passing_Year_Of_Graduation

Insights -

- Passing_Year_Of_Graduation - Year of passing Graduation ranges from a minimum of 1986 to maximum of 2020.
- Average Passing_Year_Of_Graduation is around 2002.
- The standard deviation of Passing_Year_Of_Graduation is 8.3.
- 25% , 50% (median) and 75 % of Passing_Year_Of_Graduation are 1996 , 2002 and 2009.
- Passing_Year_Of_Graduation don't have outliers.

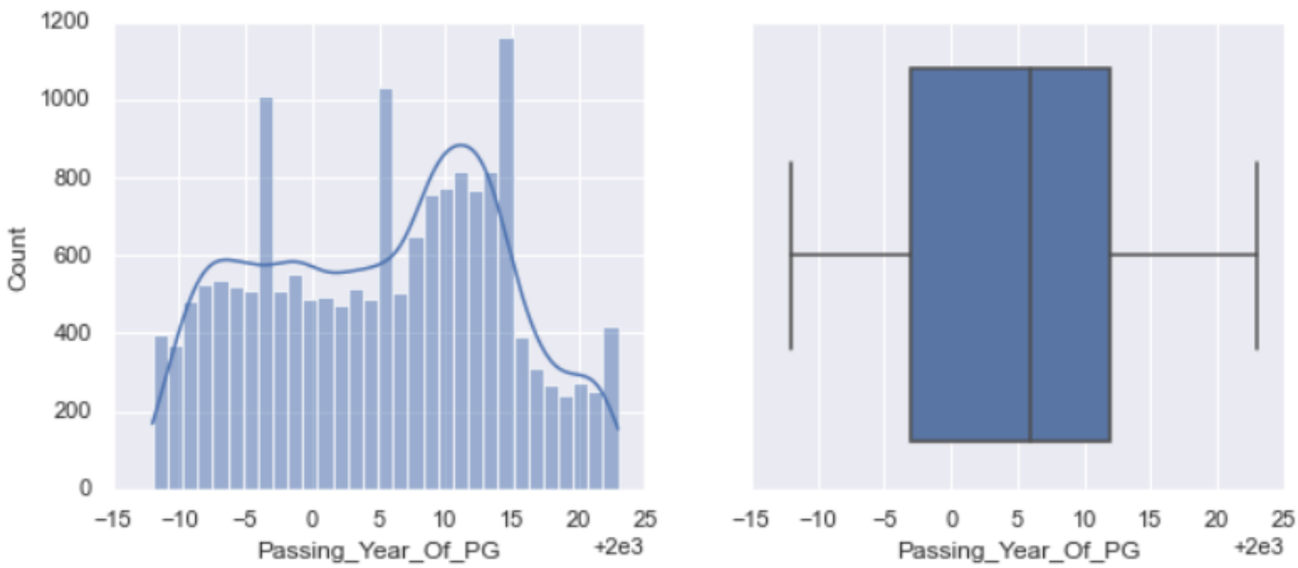


Fig: 4 Histogram & Box-Plot of Passing_Year_Of_PG

Statistical Summary	Values
Count	17308
Mean	2005.15
Std	9.02
Min	1988
25%	1997
50%	2006
75%	2012
Max	2023

Tab: 29 Statistical Description of Passing_Year_Of_PG

Insights -

- Passing_Year_Of_PG - Year of passing Post-Graduation ranges from a minimum of 1988 to maximum of 2023.
- Average Passing_Year_Of_PG is around 2005.
- The standard deviation of Passing_Year_Of_PG is 9.0.
- 25% , 50% (median) and 75 % of Passing_Year_Of_PG are 1997 , 2006 and 2012.
- Passing_Year_Of_PG don't have outliers.

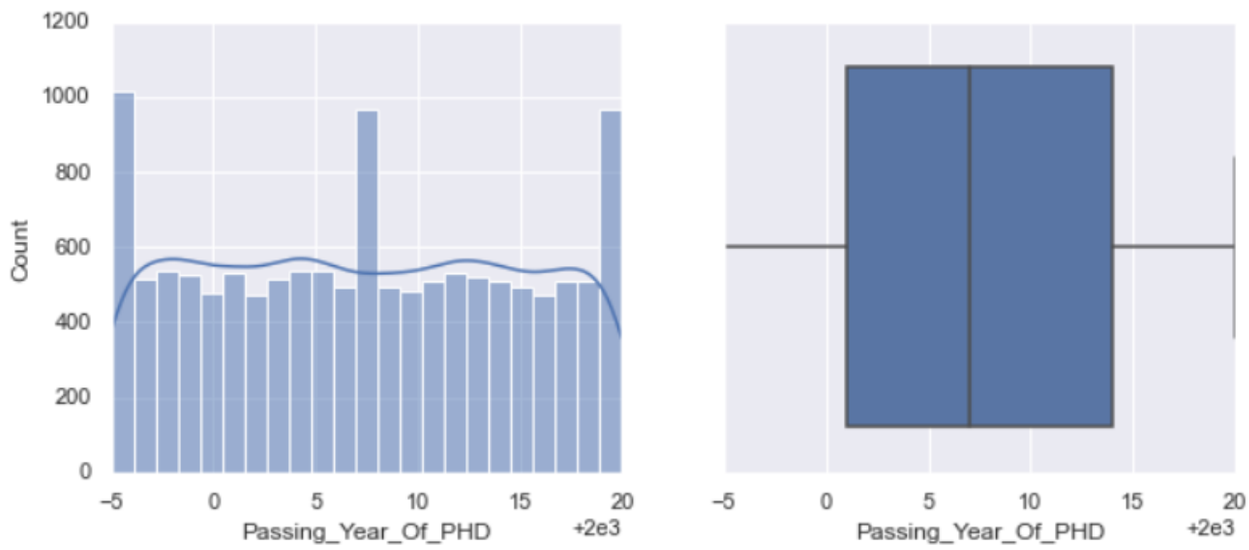


Fig: 5 Histogram & Box-Plot of Passing_Year_Of_PHD

Statistical Summary	Values
Count	13119
Mean	2007.39
Std	7.49
Min	1995
25%	2001
50%	2007
75%	2014
Max	2020

Tab: 30 Statistical Description of Passing_Year_Of_PHD

Insights -

- Passing_Year_Of_PHD - Year of passing PHD ranges from a minimum of 1995 to maximum of 2020.
- Average Passing_Year_Of_PHD is around 2007.
- The standard deviation of Passing_Year_Of_PHD is 7.
- 25% , 50% (median) and 75 % of Passing_Year_Of_PHD are 2001 , 2007 and 2014.
- Passing_Year_Of_PHD don't have outliers.

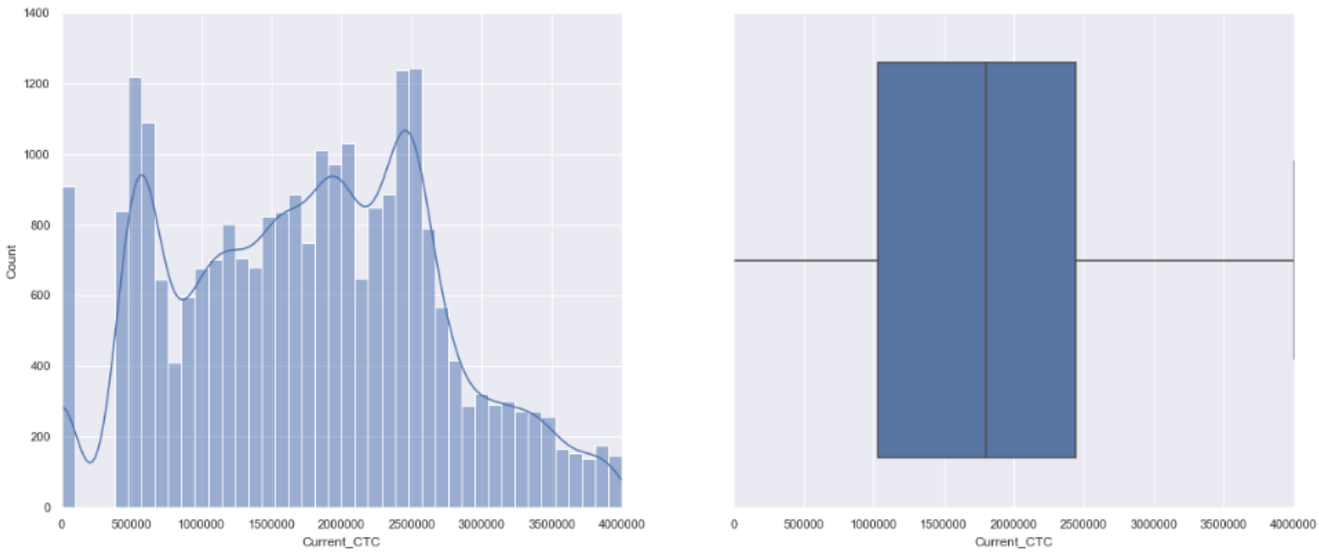


Fig: 6 Histogram & Box-Plot of Current_CTC

Statistical Summary	Values
Count	25000
Mean	1.760945E+06
Std	9.202125E+05
Min	0.0
25%	1.027312E+06
50%	1.802568E+06
75%	2.443883E+06
Max	3.999693E+06

Tab: 31 Statistical Description of Current_CTC

Insights -

- Current CTC ranges from a minimum of 0 lac to maximum of 3999693 lac.
- Average Current CTC is around 1760945 lac.
- The standard deviation of Current CTC is 920212.5 lac.
- 25% , 50% (median) and 75 % of Current CTC are 1027311.5 lac , 1802567.5 lac and 2443883.25 lac.
- Current CTC don't have outliers.

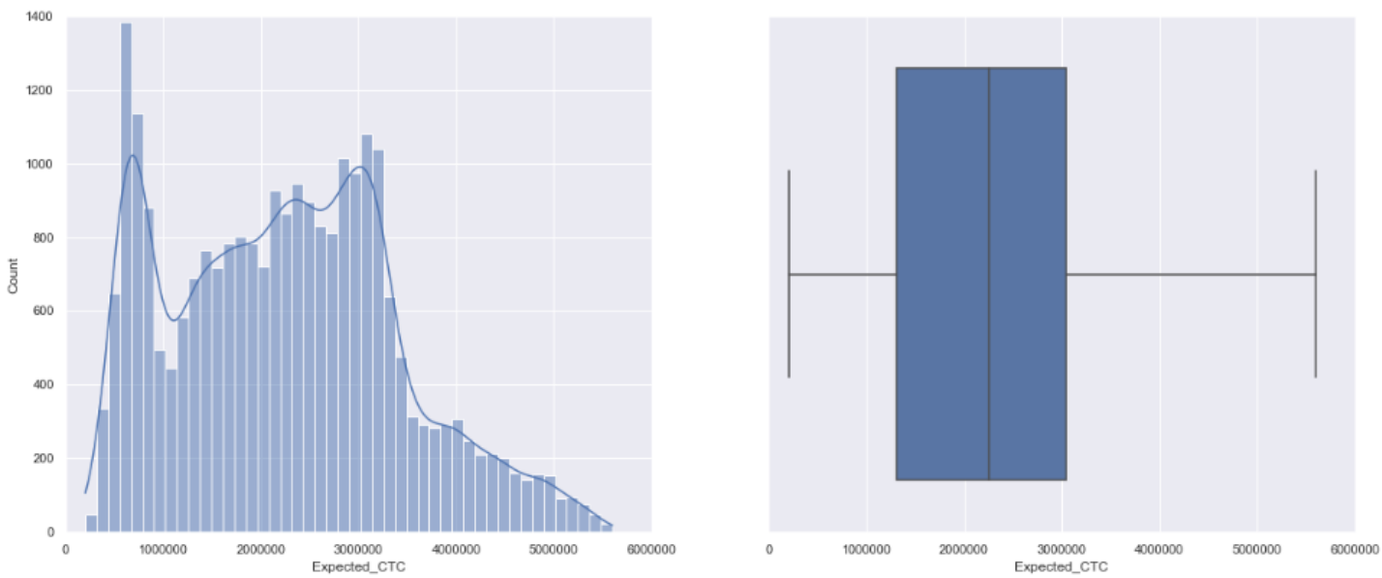


Fig: 7 Histogram & Box-Plot of Exprected_CTC

Statistical Summary	Values
Count	25000
Mean	2.250155E+06
Std	1.160480E+06
Min	2.037440E+05
25%	1.306278E+06
50%	2.252136E+06
75%	3.051354E+06
Max	5.599570E+06

Tab: 32 Statistical Description of Expected_CTC

Insights -

- Expected CTC (Final CTC offered by Delta Ltd.) ranges from a minimum of 203744 lac to maximum of 5599570 lac.
- Average Expected CTC (Final CTC offered by Delta Ltd.) is around 2250155 lac.
- The standard deviation of Expected CTC (Final CTC offered by Delta Ltd.) is 1160480 lac.
- 25% , 50% (median) and 75 % of Expected CTC (Final CTC offered by Delta Ltd.) are 1306277.5 lac ,2252136.5 lac and 3051353.75 lac.
- Expected CTC (Final CTC offered by Delta Ltd.) don't have outliers.

Univariate Analysis of Categorical Variables :

PieChart :

A pie chart is a circle divided into sectors that each represent a proportion of the whole. It is often used to show proportion, where the sum of the sectors equal 100%.

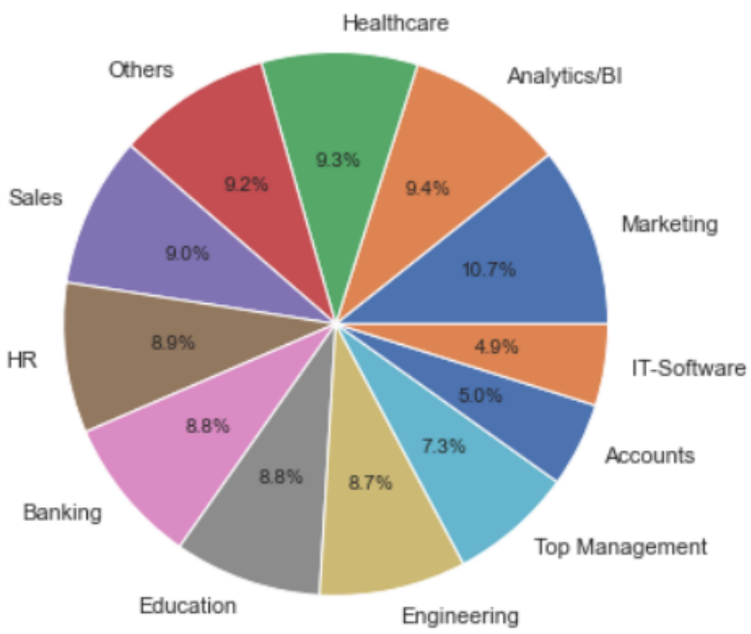


Fig:8 Pie-Plot of Department

Insights -

- There are 12 types of Department present in the data named as 'Marketing' , 'Analytics/BI' , 'Healthcare' , 'Others' , 'Sales' , 'HR' ,'Banking' , 'Education' ,'Engineering' , 'Top Management' ,'Accounts' and 'IT-Software'.
- Majority of applicants are of Marketing Department (10.70% - applicants) ,Analytics/BI (9.43% - applicants) , Health-care (9.27% - applicants) and others (9.18% - applicants).
- Least applicants belongs to IT-Software Department(4.8% - applicants).
- There is fine distribution of applicants in each department.

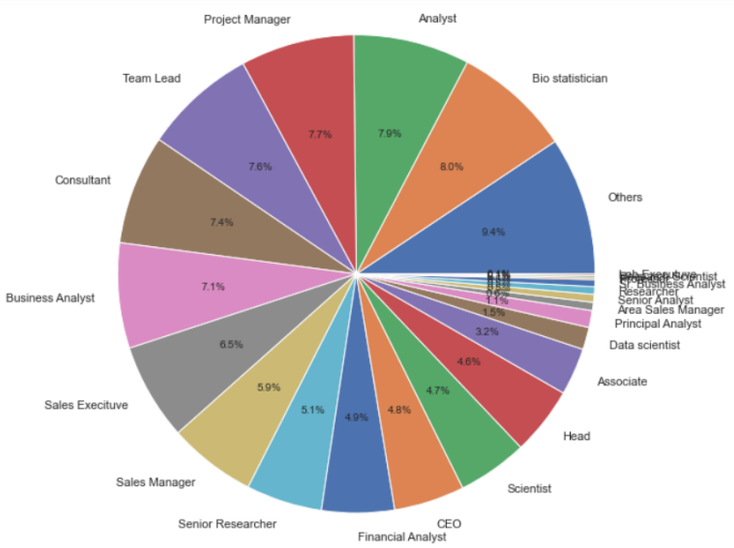


Fig:9 Pie-Plot of Role

Insights -

- There are 24 types of Roles present in the data named as 'Others' , 'Bio-statistician' ,'Analyst' , 'Project Manager' , 'Team Lead' ,'Consultant' ,'Business Analyst ' ,'Sales Execituve' , 'Sales Manager' , 'Senior Researcher' , 'Financial Analyst' , 'CEO' , 'Scientist' , 'Head' , 'Associate' ,'Data scientist' ,'Principal Analyst' , 'Area Sales Manager' ,'Senior Analyst' ,'Researcher' , 'Sr. Business Analyst' , 'Professor' ,'Research Scientist' and 'Lab Executuve'.
- Majority of applicants worked as others(Role) i.e (9.4 %).
- Only 0.1% applicants worked as Lab executive.

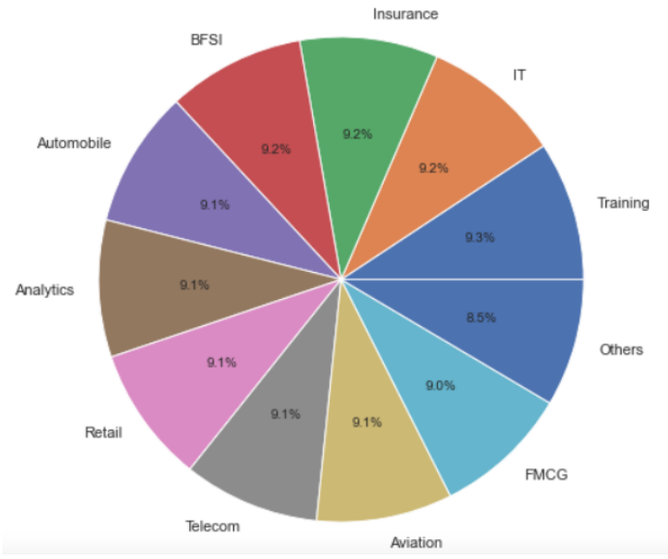


Fig:10 Pie-Plot of Industry

Insights -

- There are 11 types of Industries present in the data named as 'Training' , 'IT' ,'Insurance' ,'BFSI' ,'Automobile' , 'Analytics' , 'Retail' ,'Telecom' ,'Aviation' ,'FMCG' and 'Others'.
- Most of applicants worked in Training Industry i.e.(9.3%).
- There is not much variation in the Industry column have a fair distribution for all applicants.

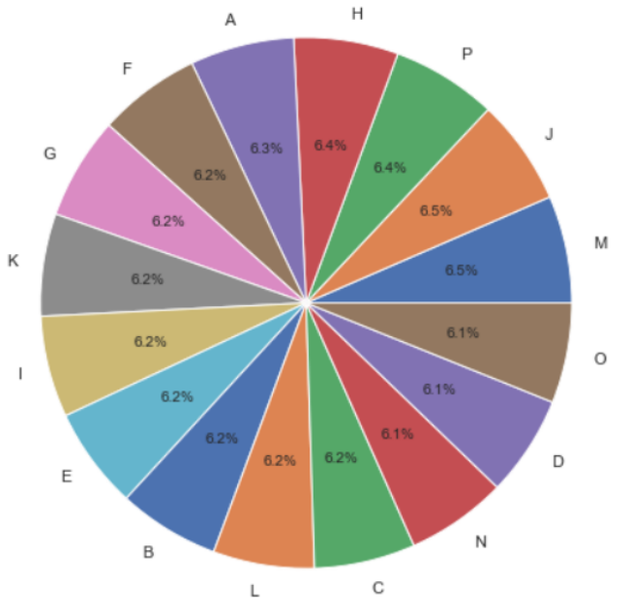


Fig:11 Pie-Plot of Organization

Insights -

- There are 16 types of Organization name present in the dataset named as 'M' ,'J' ,'P' ,'H' , 'A ' , 'F ' ,'G' ,'K' ,'I' ,'E' ,'B' ,'L' ,'C' ,'N' , 'D' and 'O'.
- There is not too much variations in the Organization columns equivalent number of applicants worked in 16 different organization.

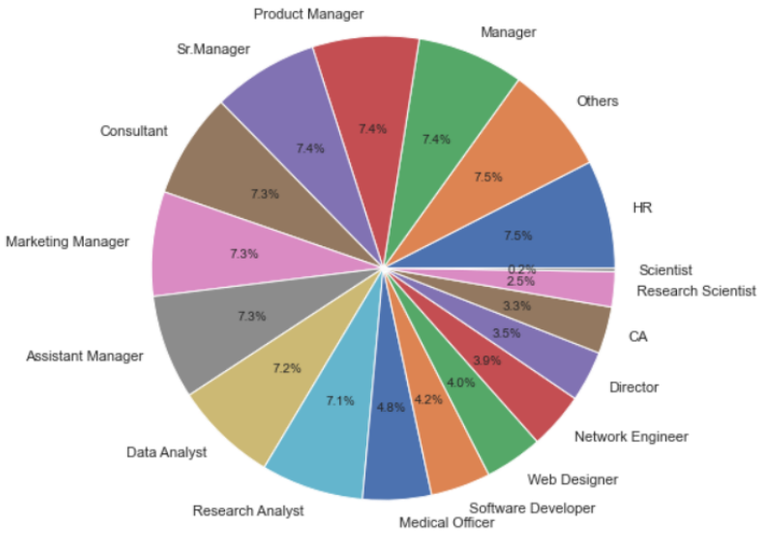


Fig:12 Pie-Plot of Designation

Insights -

- There are 18 types of Designation present in the dataset named as 'HR' , 'Others' ,'Manager' , 'Product Manager' ,'Sr.Manager' ,'Consultant' ,'Marketing Manager' ,'Assistant Manager' , 'Data Analyst' ,'Research Analyst','Medical Officer' ,'Software Developer' , 'Web Designer' .'Network Engineer' , 'Director' ,'CA' ,'Research Scientist' and 'Scientist'.
- Most of applicants worked as HR and others i.e.(7.5%).
- Only 0.2% applicants worked as Scientist.
- There is fine distribution of applicants across various Designation.

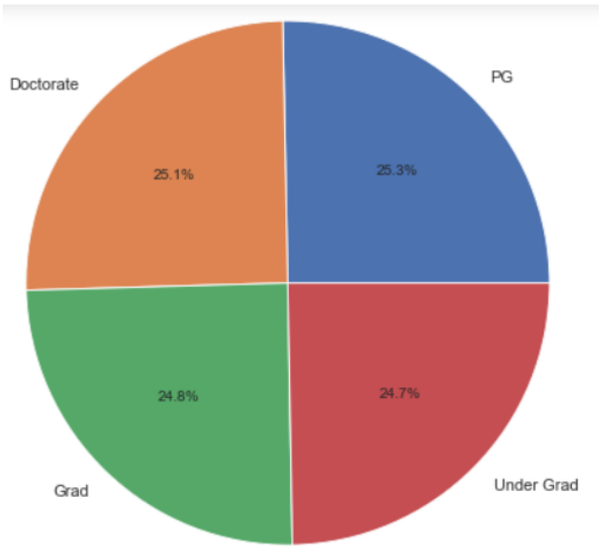


Fig:13 Pie-Plot of Highest Education

Insights -

- There are 4 types of Education labels present in the dataset named as 'Under Grad' ,'Grad' ,'PG' and Doctorate.
- 24.72% applicants are Under Graduate.
- 24.83% applicants are Graduate.
- 25.30% applicants are Post Graduate.
- 25.14% applicants are Doctorate.
- There is fine distribution of applicants in each education label.

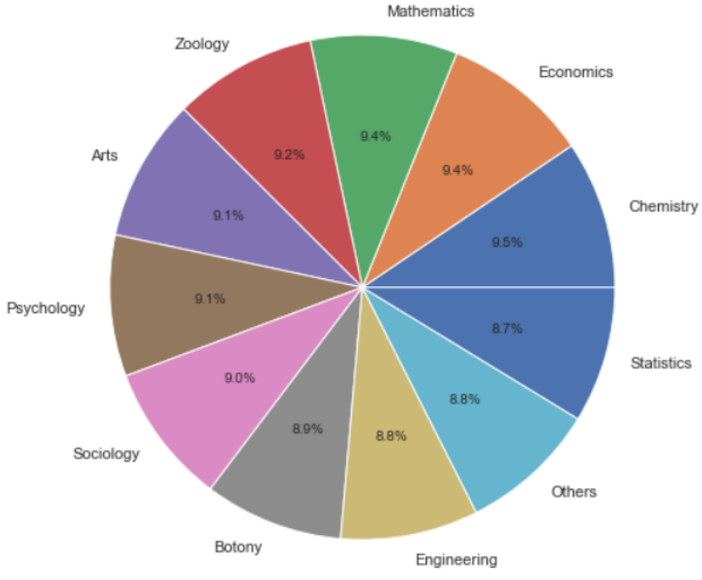


Fig:14 Pie-Plot of Graduation_Specialization

Insights -

- There are 11 types of Graduation_Specialization labels present in the dataset named as 'Chemistry' , 'Economics ' , 'Mathematics' ,'Zoology' ,'Arts' ,'Psychology' , 'Sociology' , 'Botony' ,'Engineering' , 'Others' and 'Statistics'.
- Most of applicants did their graduation specialization in chemistry ,i.e.(9.5%).
- There is fine distribution of applicants in each Graduation_Specialization label.

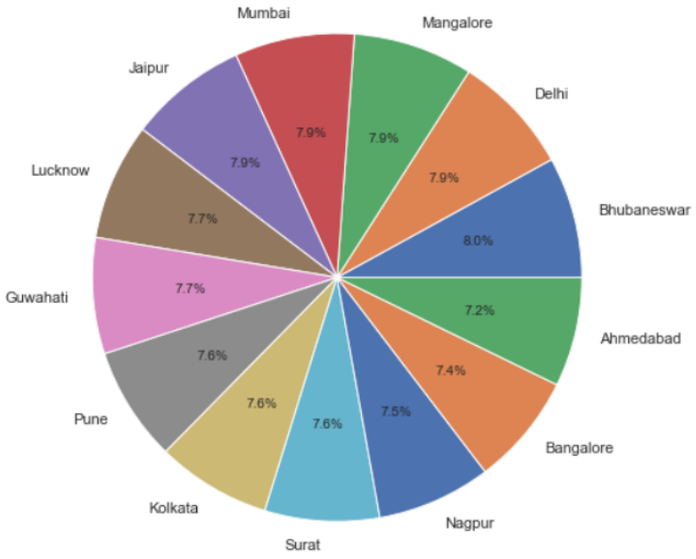


Fig:15 Pie-Plot of University_Grad

Insights -

- There are 13 types of University_Grad labels present in the dataset named as 'Bhubaneswar' ,'Delhi' ,'Mangalore' ,'Mumbai' ,'Jaipur' ,'Lucknow' ,'Guwahati' ,'Pune' ,'Kolkata' ,'Surat' ,'Nagpur' ,'Bangalore' and 'Ahmedabad'.
- Most of applicants did their graduation from Bhubaneswar University (8%) and Delhi University(7.9%).
- There is equivalent number of graduates from all universities.

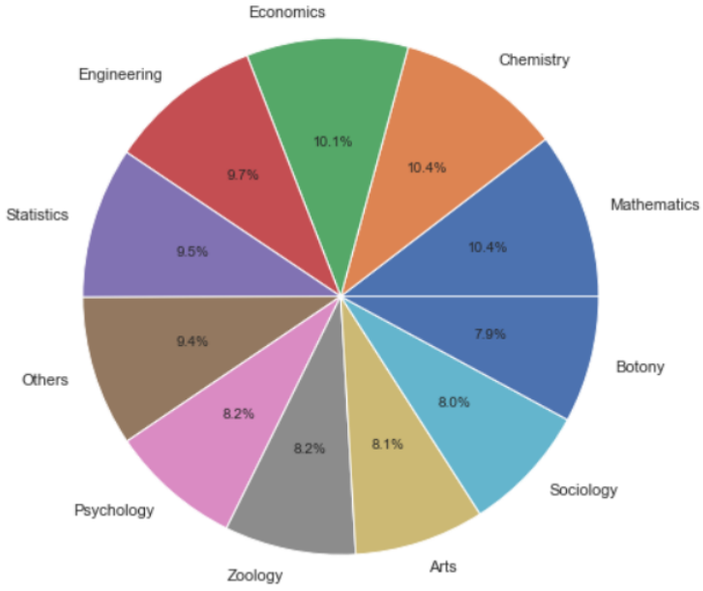


Fig:16 Pie-Plot of PG_Specialization

Insights -

- There are 11 types of PG_Specialization labels present in the dataset named as 'Mathematics' , 'Chemistry' ,'Economics' ,'Engineering' ,'Statistics' , 'Others' , 'Psychology' , 'Zoology' , 'Arts' , 'Sociology' and 'Botony'.
- Most of applicants did their post-graduation specialization in mathematics (10.4%) and chemistry (10.4%).
- There is fine distribution of applicants in each PG_Specialization label.

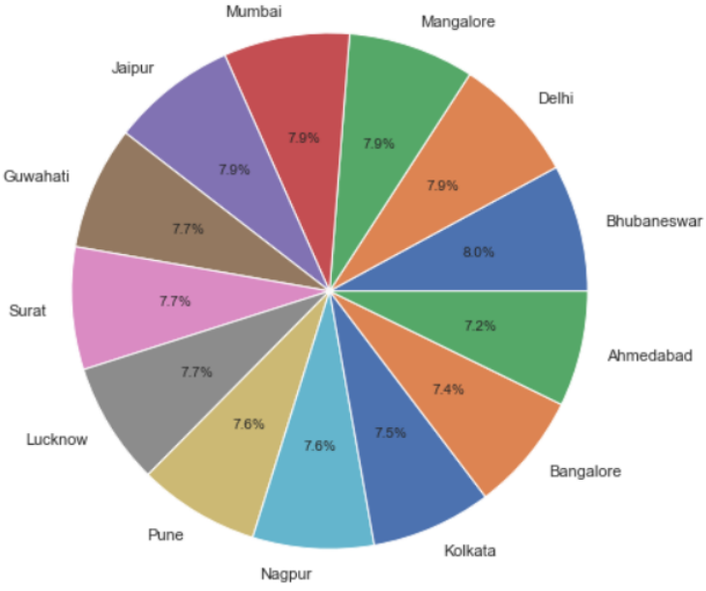


Fig:17 Pie-Plot of University_PG

Insights -

- There are 13 types of University_PG labels present in the dataset named as 'Bhubaneswar' ,'Delhi' ,'Mangalore' ,'Mumbai' ,'Jaipur' ,'Lucknow' ,'Guwahati' ,'Pune' ,'Kolkata' ,'Surat' ,'Nagpur' ,'Bangalore' and 'Ahmedabad'.
- Most of applicants did their graduation from Bhubaneswar University (8%) and Delhi University(7.9%).
- There is equivalent number of post-graduates from all universities.

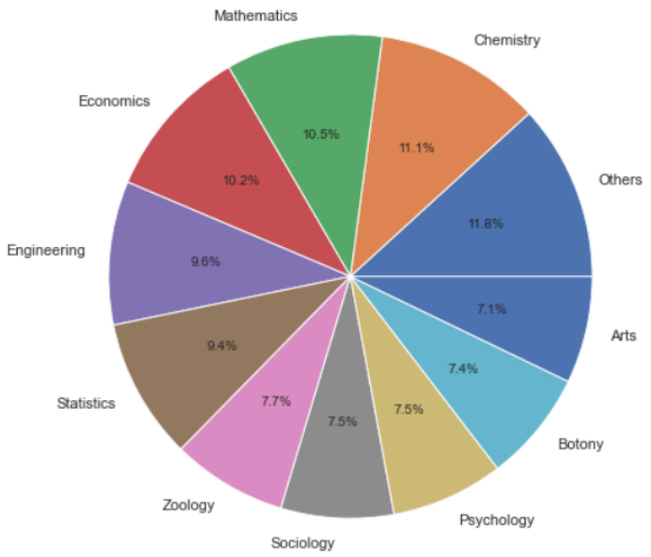


Fig:18 Pie-Plot of PHD_Specialization

Insights -

- There are 11 types of PHD_Specialization labels present in the dataset named as 'Others' ,'Chemistry','Mathematics','Economics' ,'Engineering' ,'Statistics' ,'Zoology','Sociology' ,'Psychology ' ,'Botony' and 'Arts'.
- Most of applicants did their PhD specialization in others (11.8%) and chemistry (11.1%).
- Rest there is well distribution of applicants in PHD_Specialization labels.

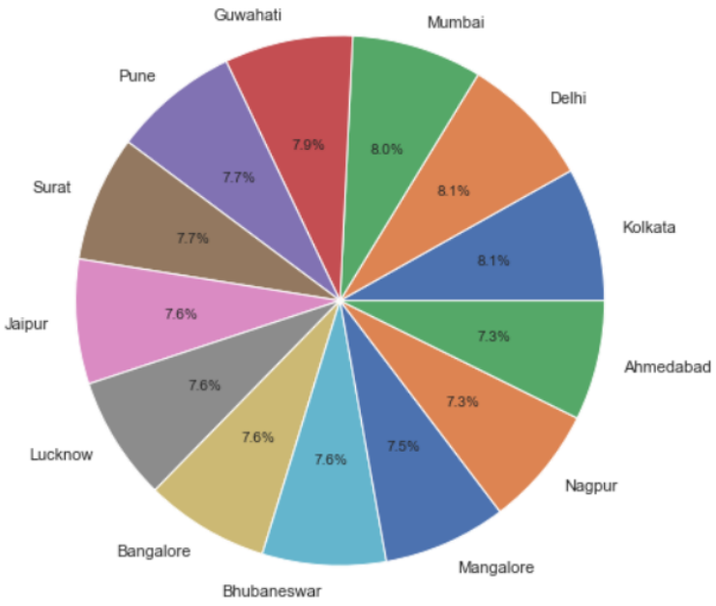


Fig:19 Pie-Plot of University_PHD

Insights -

- There are 13 types of University_PHD labels present in the dataset named as 'Bhubaneswar' ,'Delhi' ,'Mangalore' ,'Mumbai' ,'Jaipur' ,'Lucknow' ,'Guwahati' ,'Pune' ,'Kolkata' ,'Surat' ,'Nagpur' ,'Bangalore' and 'Ahmedabad'
- Most of applicants did their PhD from Kolkata University (8.1%) and Delhi University(8.1%).
- There is almost equivalent number of PhD applicants from all universities.

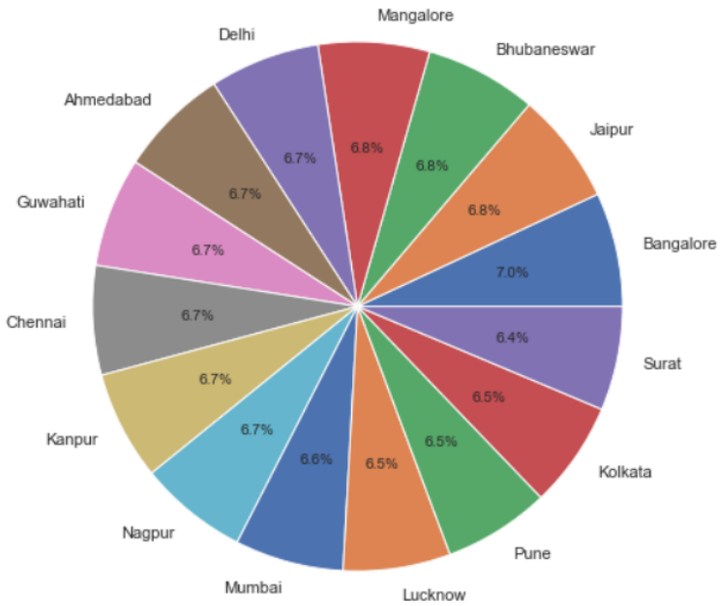


Fig:20 Pie-Plot of Current_location

Insights -

- There are 15 types of Curent_Location labels present in the dataset named as 'Bangalore' , 'Jaipur' ,'Bhubaneswar' , 'Mangalore' ,'Delhi' ,'Ahmedabad' ,'Guwahati' ,'Chennai' ,'Kanpur' ,'Nagpur' ,'Mumbai' ,'Lucknow' ,'Pune' ,'Kolkata' and 'Surat'.
- Most of applicants Curent_Location is Bangalore (7.0%) , Jaipur , Bhubaneswar and Mangalore (6.8%).
- Rest there is fair distribution of frequency for applicant's current location.

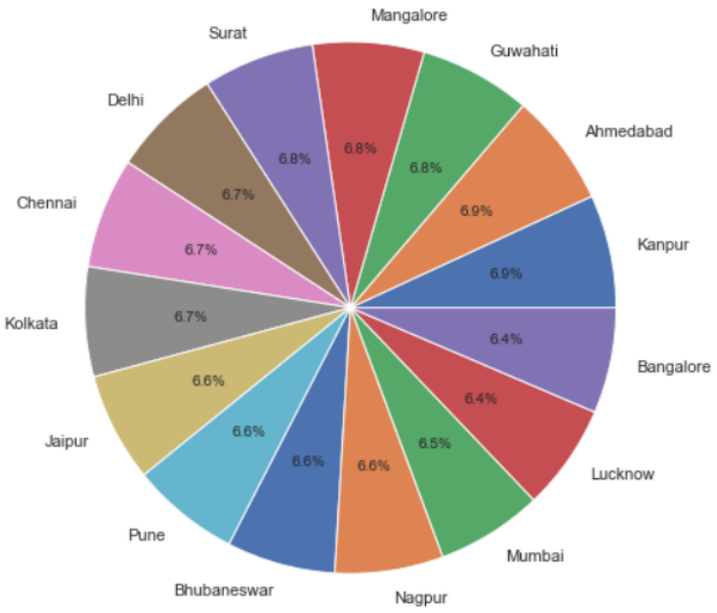


Fig:21 Pie-Plot of Preferred_location

Insights -

- There are 15 types of Preferred_location labels present in the dataset named as 'Bangalore' , 'Jaipur' ,'Bhubaneswar' , 'Mangalore' ,'Delhi' ,'Ahmedabad' ,'Guwahati' ,'Chennai' ,'Kanpur' ,'Nagpur' ,'Mumbai' ,'Lucknow' ,'Pune' ,'Kolkata' and 'Surat'.
- Most of applicant's preferred location is Kanpur , Ahmedabad i.e (6.9%).
- Rest there is fair distribution of frequency for applicant's preferred location.

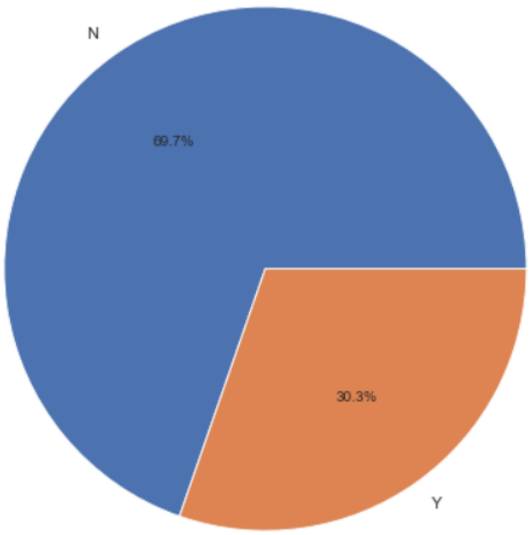


Fig:22 Pie-Plot of Inhand_Offer

Insights -

- There are 2 types of Inhand_Offer labels present in the dataset named as 'Y'(Yes) and 'N' (No).
- Around 69.7% applicants don't holds any offer in their hand.
- Only 30.3% applicants holds any offer in their hand.

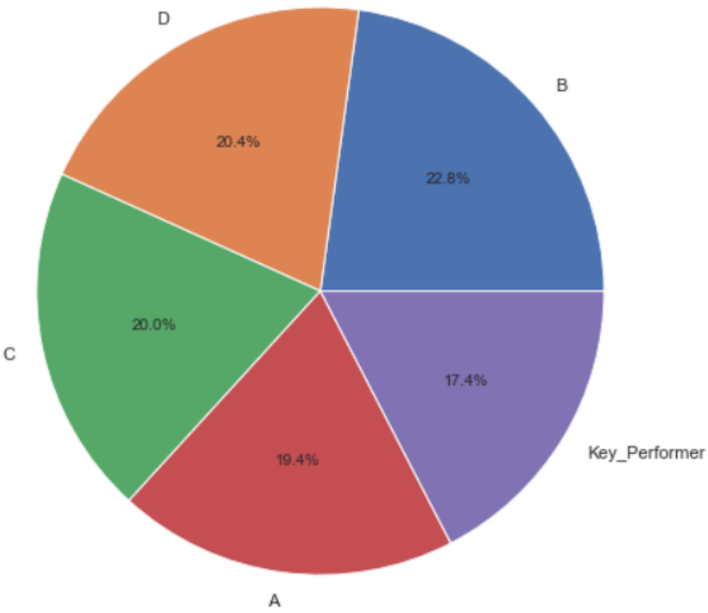


Fig:23 Pie-Plot of Last_Appraisal_Rating

Insights -

- There are 5 types of Last_Appraisal_Rating labels present in the dataset named as 'Key_Performer' , 'A' ,'B' ,'C' and 'D'.
- Around 17.4% applicants Last_Appraisal_Rating is Key_Performer.
- 19.4% applicants Last_Appraisal_Rating is A.
- 22.8% applicants Last_Appraisal_Rating is B.
- 20% applicants Last_Appraisal_Rating is C.
- 20.4% applicants Last_Appraisal_Rating is D.

Bivariate Analysis :

Scatter Plot :

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

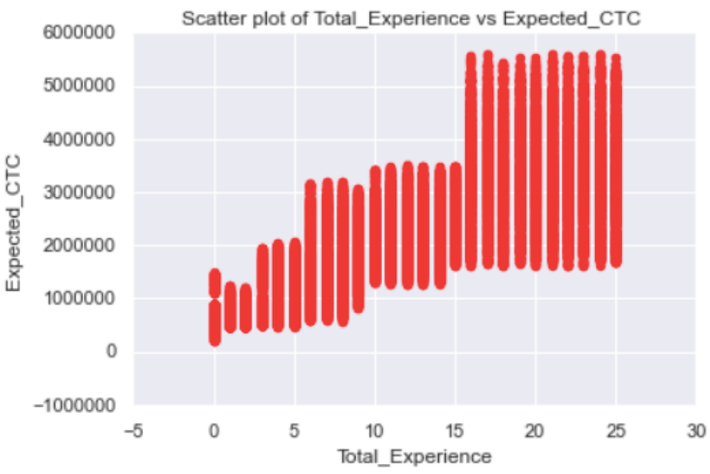


Fig: 24 Scatter plot of Total_Experience vs Expected_CTC

Insights -

- From the above plot we see that the Total_Experience and the Expected_CTC is showing a strong relationship,with increase in Total_Experience(Independent Variable),Expected_CTC (Target Variable)is also increases.
- Applicants with higher Total_Experience have higher Expected_CTC.
- This will be a good feature for predicting the target variable ("Expected_CTC).

Fig: 25 Scatter plot of Total_Experience_in_field_appliedvs Expected_CTC

Insights -

- From the above plot we see that the Total_Experience_in_field_applied and the Expected_CTC is showing a positive relationship,with increase in Total_Experience_in_field_applied(Independent Variable),Expected_CTC (Target Variable)is slightly increases.
- Applicants with higher Total_Experience_in_field_applied have high Expected_CTC.
- This may be a good feature for predicting the target variable ("Expected_CTC).

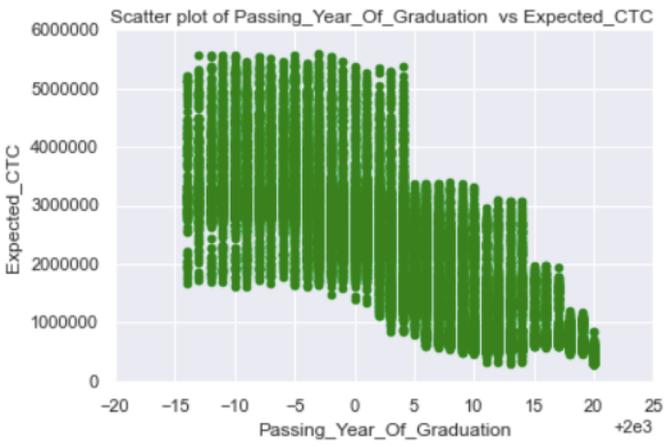


Fig: 26 Scatter plot of Passing_Year_Of_Graduation vs Expected_CTC

Insights -

- From the above plot we see that the Passing_Year_Of_Graduation and the Expected_CTC is showing a negative relationship, as the Passing_Year_Of_Graduation increases the Expected_CTC goes on decreases.
- Recent graduate applicants have lower expected ctc as compared to other applicants.

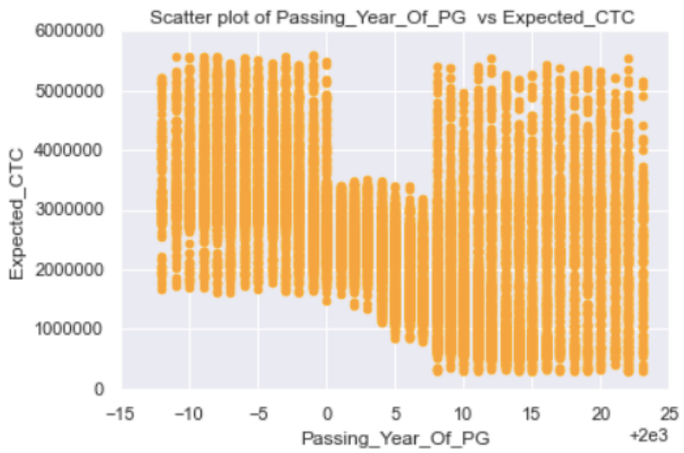


Fig: 27 Scatter plot of Passing_Year_Of_PG vs Expected_CTC

Insights -

- There is no specific relation between Passing_Year_Of_PG and Expected_CTC.
- Recently Post-Grad applicants have lower as well as high expected_ctc.

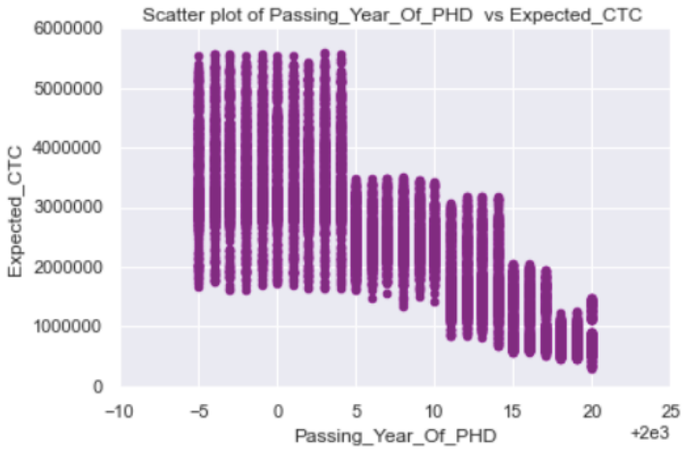


Fig: 28 Scatter plot of Passing_Year_Of_PHD vs Expected_CTC

Insights -

- From the above plot we see that the Passing_Year_Of_PHD and the Expected_CTC is showing a negative relationship, as the Passing_Year_Of_PHD increases the Expected_CTC goes on decreases.
- Recently passed PhD applicants have lower expected ctc as compared to other applicants.

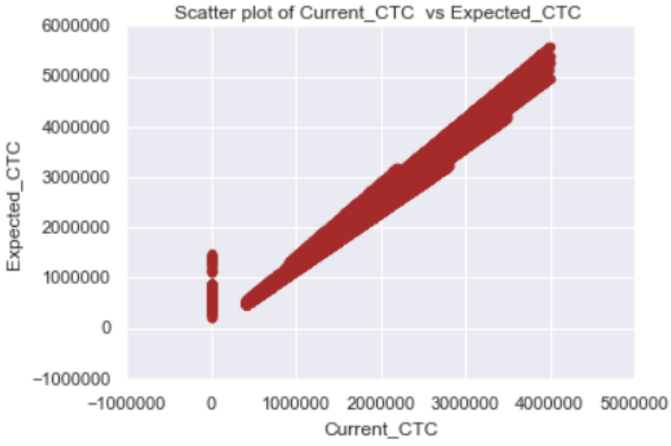


Fig: 29 Scatter plot of Current_CTC vs Expected_CTC

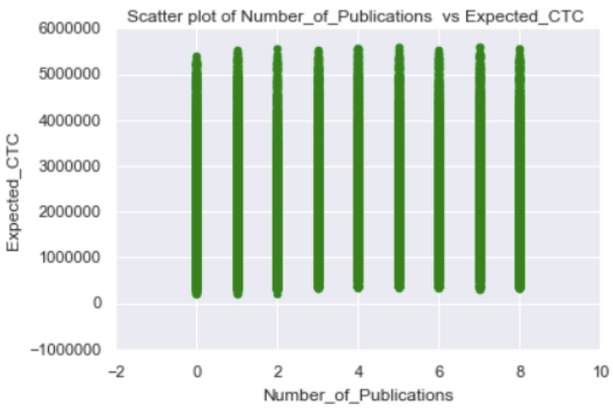
Insights -

- From the above plot we see that the Current_CTC and the Expected_CTC is showing a positive relationship, as the Current_CTC increases the Expected_CTC goes on increases.
- Applicants with Current ctc zero have lower expected ctc.
- Current_CTC will be a good predictor to predict the Expected_CTC.

Fig: 30 Scatter plot of No_Of_Companies_worked vs Expected_CTC

Insights -

From the above plot we infer that there is some sort of relation between No_Of_Companies_worked and Expected_CTC.As No_Of_Companies_worked increase there is also some increase in Expected_CTC.



Insights -

- There is no such relationship between Number_of_Publications and Expected_CTC.From the above visual we infer there is not any kind of impact of Number_of_Publications on Expected_CTC.Expected_CTC is somehow equivalent for all who have less or more Number_of_Publications.

Fig: 31 Scatter plot of Number_of_Publications vs Expected_CTC

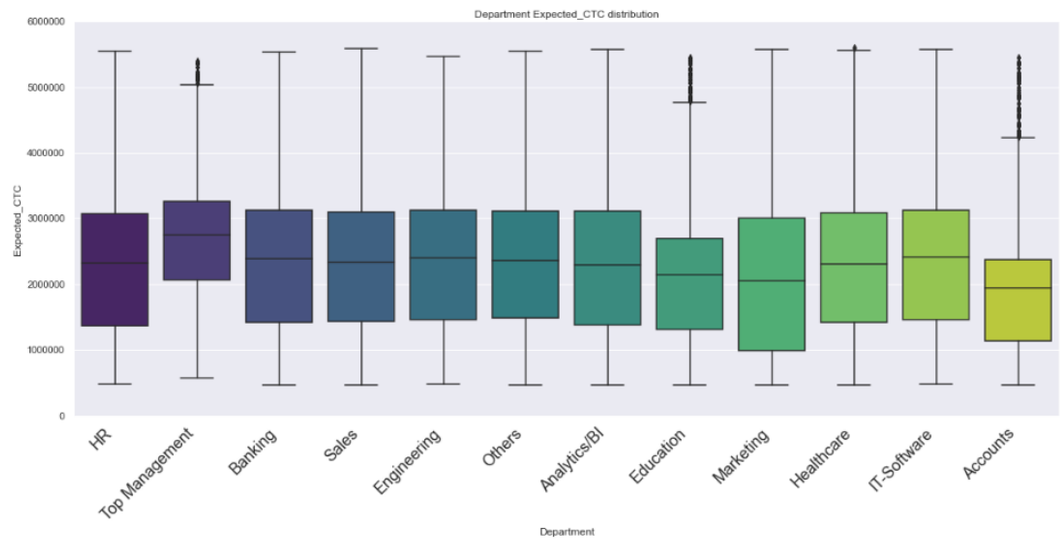


Fig: 32 Box- Plot of Department Vs Expected_CTC

Insights -

- Expected_CTC does vary based on the Department as expected. This conclusion can only be drawn through the graphical plots.
- Applicants for Top Management have higher median value than others for Expected_CTC Distribution for Expected_CTC is bigger for Marketing Department applicants.
- Banking , Sales , Engineering and Others have median almost equivalent to each other and have almost similar kind of distribution too.

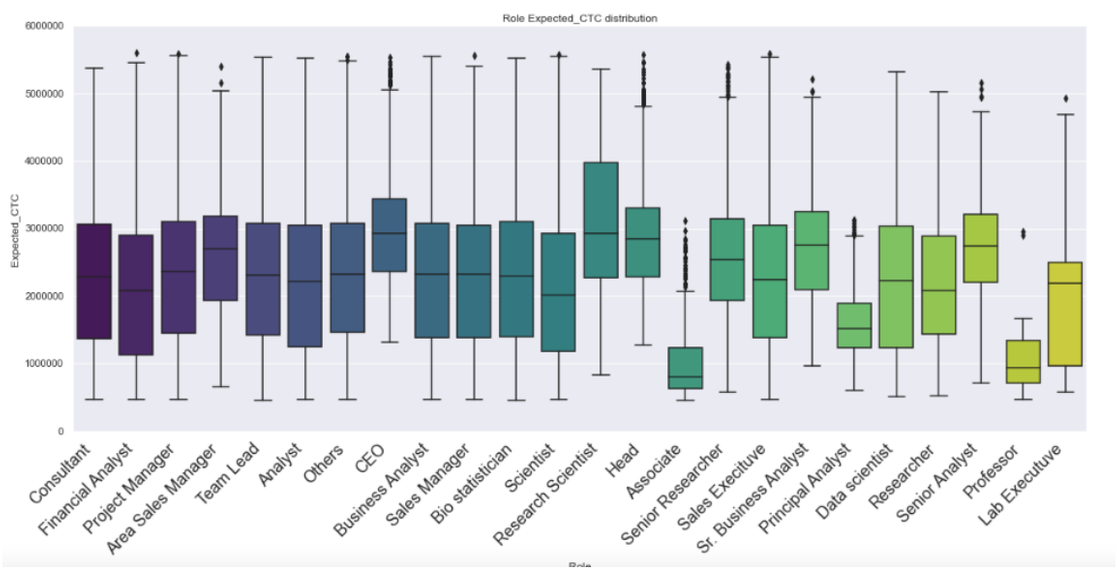


Fig: 33 Box- Plot of Role Vs Expected_CTC

Insights -

- Median values of CEO and Research Scientists for Expected_CTC are quite high as compared to others but distribution is wider for Research Scientists.
- Median values for Expected_CTC of Business Analyst , Sales Manager and Bio-Statistician are almost equivalent to each other.
- Professors also have low Expected_CTC than others.
- Associate have least Expected_CTC compared to others roles.

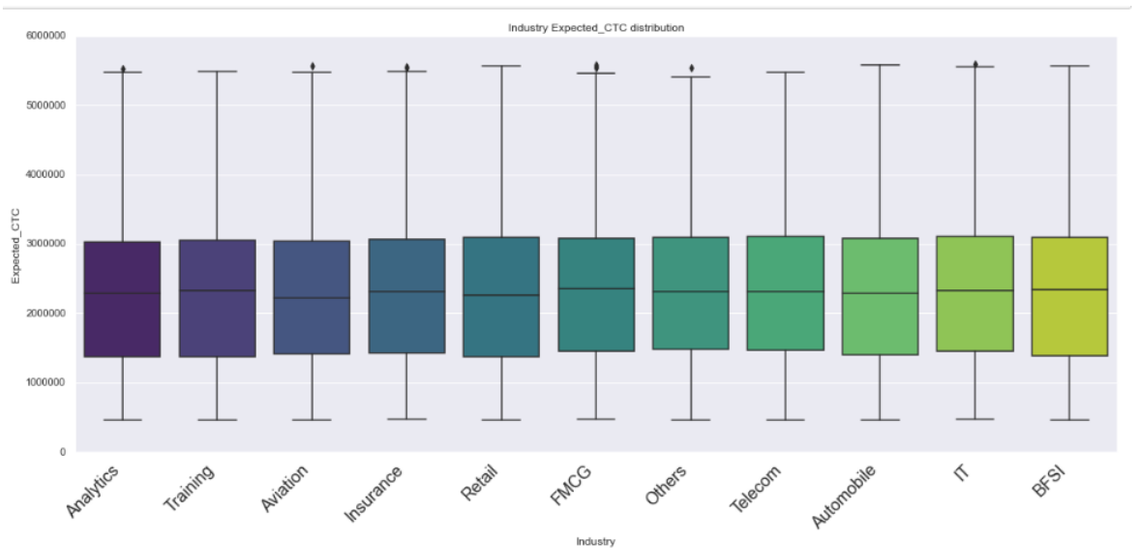


Fig: 34 Box- Plot of Industry Vs Expected_CTC

Insights -

- There is not any variation in the distribution of Expected_CTC w.r.t. Industry , looks almost similar plus median values are also almost equivalent for all distribution not varies too much.

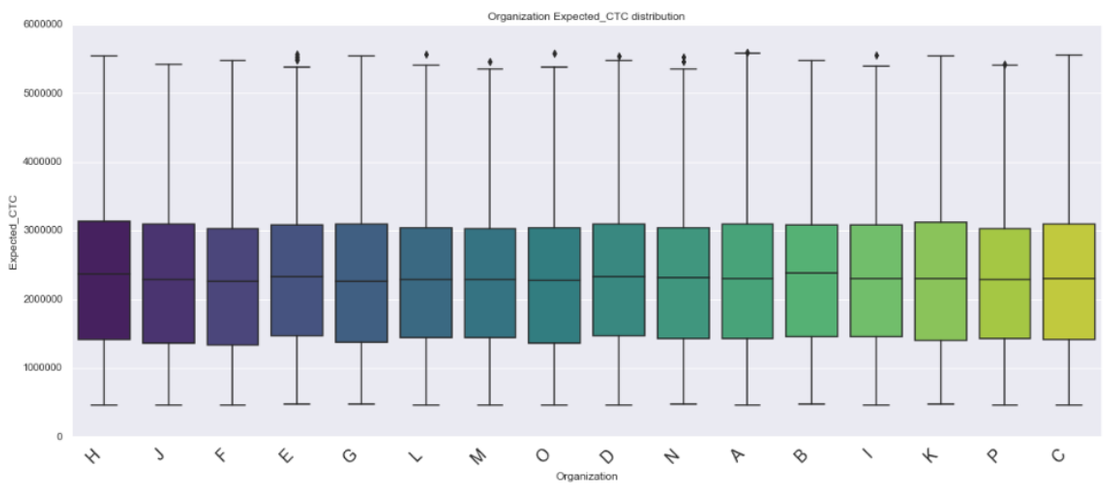


Fig: 35 Box- Plot of Organisation Vs Expected_CTC

Insights -

- There is not any variation in the distribution of Expected_CTC w.r.t. Organization , looks almost similar plus median values are also almost equivalent for all distribution not varies too much.

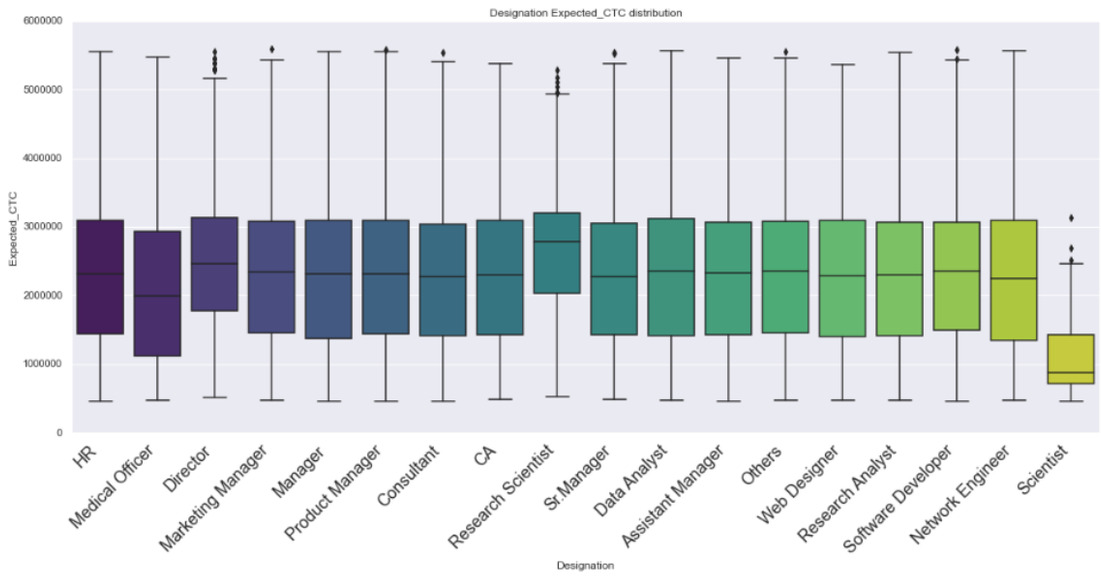


Fig: 36 Box- Plot of Designation Vs Expected_CTC

Insights -

- Median values of Research Scientists for Expected_CTC are quite high as compared to others.
- Marketing Manger , Manager ,Product Manager and HR almost have equivalent median values for Expected_CTC.
- Similarly Data Analyst , Assistant Manger , Others , Web Designers and Research Analyst have equivalent median values for Expected_CTC.

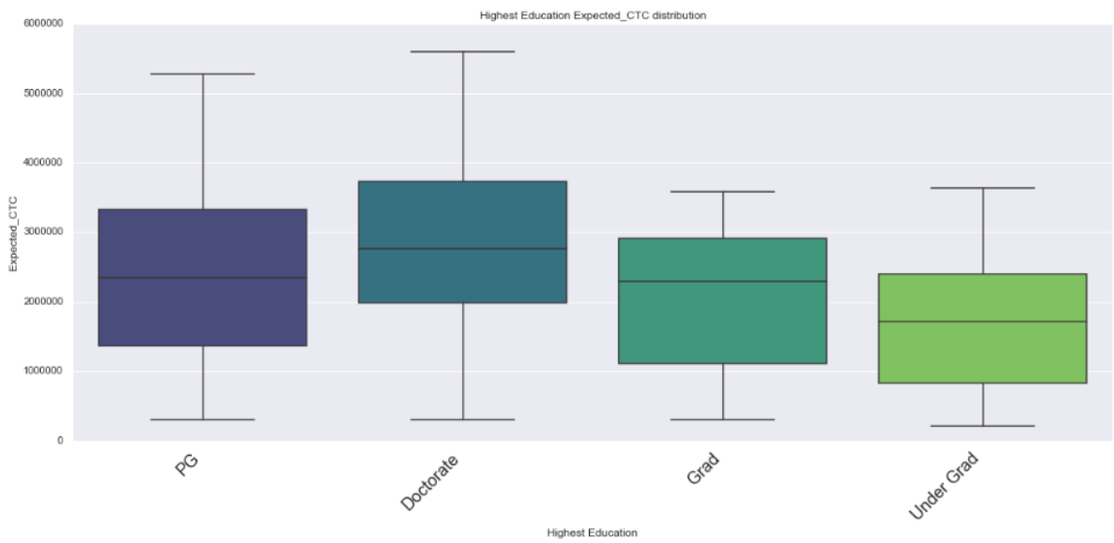


Fig: 37 Box- Plot of Highest Education Vs Expected_CTC

Insights -

- Box-plot of Doctorate have higher median values for Expected_CTC as compared to others.
- Under Grad Box-plot have lowest median values for Expected_CTC.
- PG Box-plot is almost normal distributed.

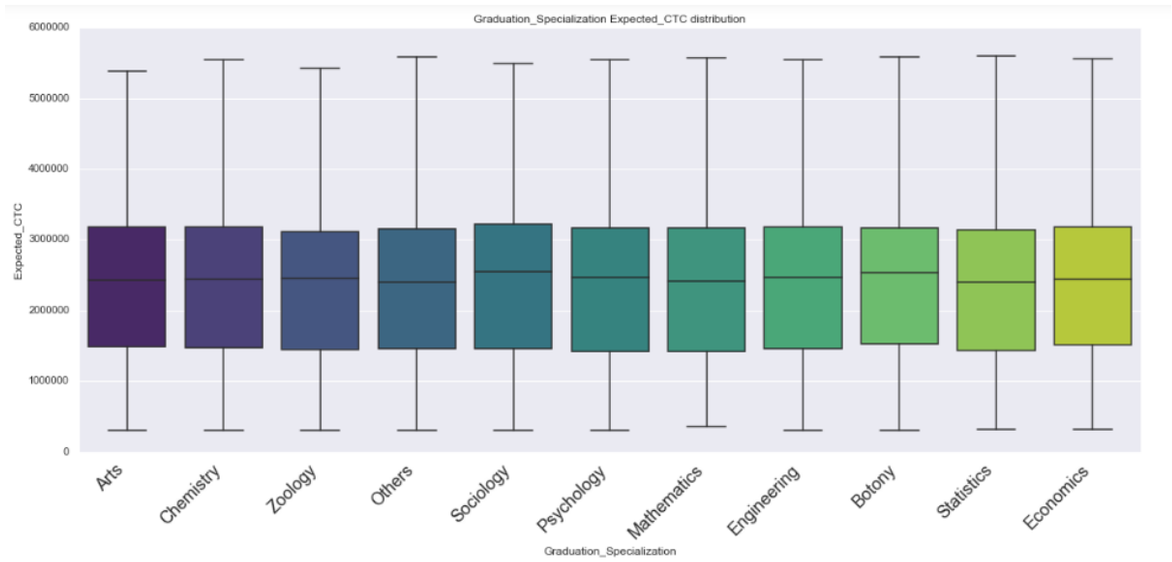


Fig: 38 Box- Plot of Graduation_Specialization Vs Expected_CTC

Insights -

There is not any variation in the distribution of Expected_CTC w.r.t. Graduation_Specialization , looks almost similar plus median values are also almost equivalent for all not varies too much.

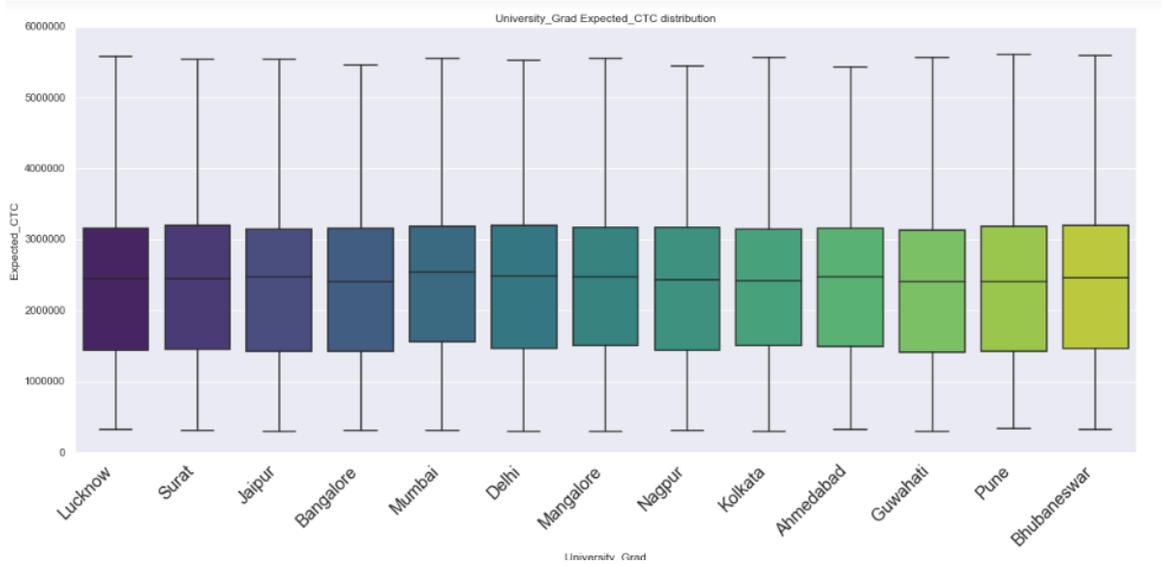


Fig: 39 Box- Plot of University_Grad Vs Expected_CTC

Insights -

There is not any variation in the distribution of Expected_CTC w.r.t. University_Grad , looks almost similar plus median values are also almost equivalent for all not varies too much.

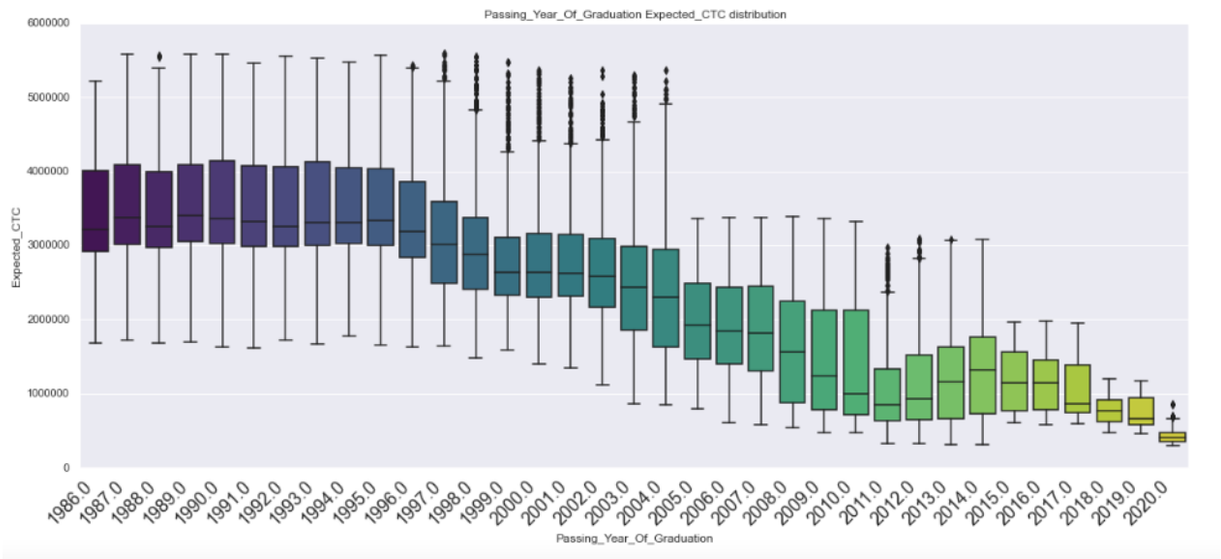


Fig: 40 Box- Plot of Passing_Year_Of_Graduation Vs Expected_CTC

Insights -

- Expected_CTC does vary based on the Passing_Year_Of_Graduation as expected. This conclusion can only be drawn through the above plot.
- We infer that Expected_CTC for recently graduated applicants is least as compared to others.
- There is variation in distribution of Expected_CTC w.r.t Passing_Year_Of_Graduation former graduate applicants have high median values for Expected_CTC than recently graduates.

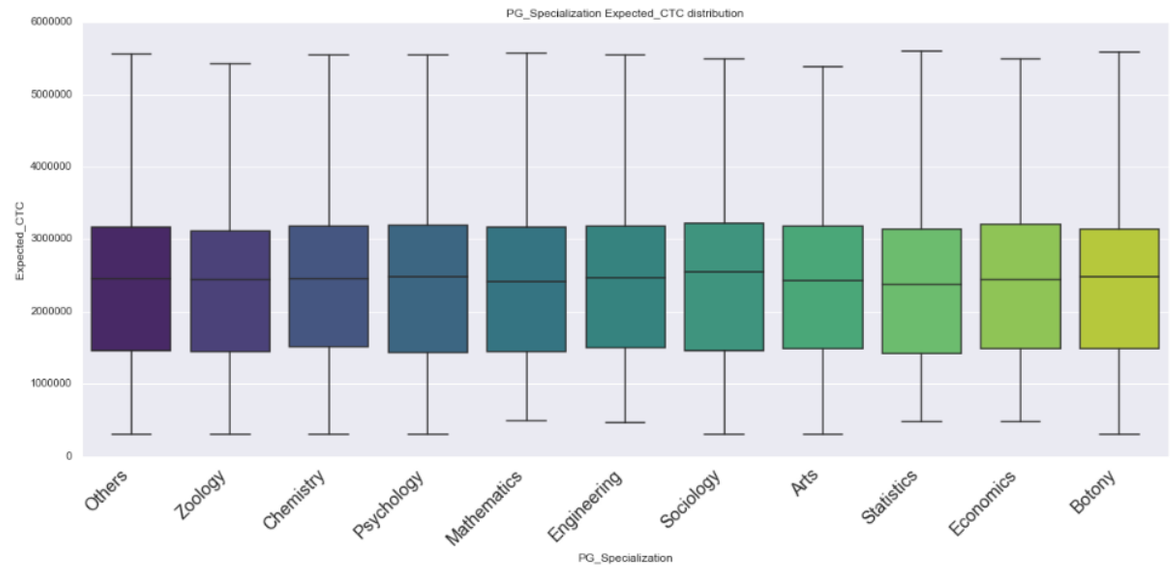


Fig: 41 Box- Plot of PG_Specialization Vs Expected_CTC

Insights -

There is not any variation in the distribution of Expected_CTC w.r.t. PG_Specialization , looks almost similar plus median values are also almost equivalent for all not varies too much.

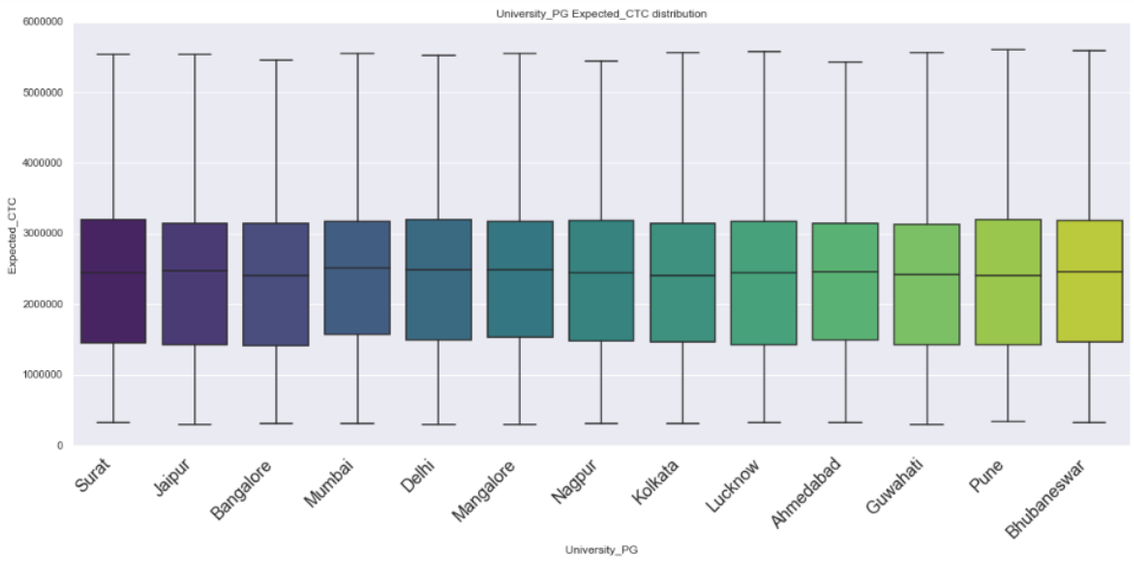


Fig: 42 Box- Plot of University_PG Vs Expected_CTC

Insights -

There is not any variation in the distribution of Expected_CTC w.r.t. PG_Specialization , looks almost similar plus median values are also almost equivalent for all not varies too much.

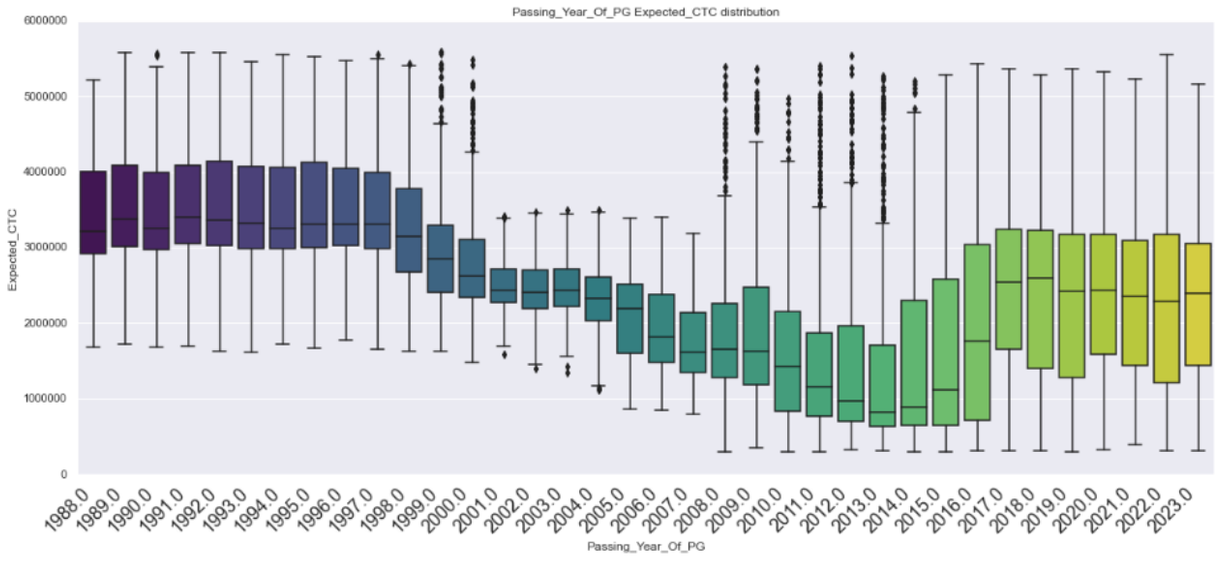


Fig: 43 Box- Plot of Passing_Year_Of_PG Vs Expected_CTC

Insights -

- Expected_CTC does vary based on the Passing_Year_Of_PG as expected. This conclusion can only be drawn through the above plot.
- We infer that Expected_CTC for recently post-graduated applicants is more than applicants who completed post-graduation in early 1990s.
- There is variation in distribution of Expected_CTC w.r.t Passing_Year_Of_PG , early 1990s applicants have high median for Expected_CTC , then in 20s there is some fall which keeps on increasing by each year passed , this variation may be caused as some of them unable to complete their PG in specific 2 year span or unable to complete their PG by any reasons.

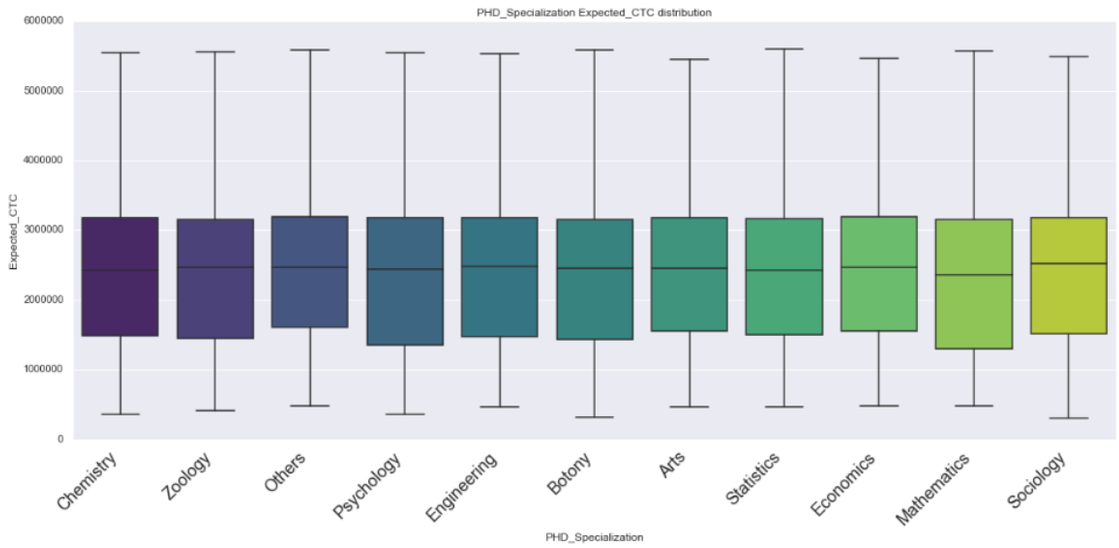


Fig: 44 Box- Plot of PHD_Specialization Vs Expected_CTC

Insights -

- Expected_CTC doesn't vary based on the PHD_Specialization as expected. This conclusion can only be drawn through the above plot.
- Psychology and Mathematics have wider distribution of box-plot.

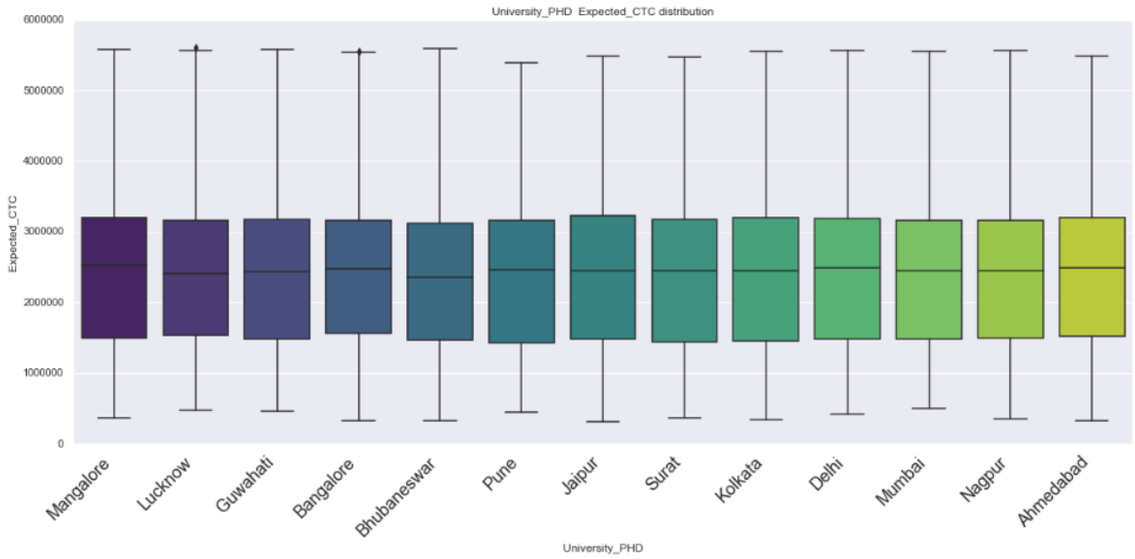


Fig: 45 Box- Plot of University_PHD Vs Expected_CTC

Insights -

There is not any variation in the distribution of Expected_CTC w.r.t. University_PHD , looks almost similar plus median values are also almost equivalent for all not varies too much.

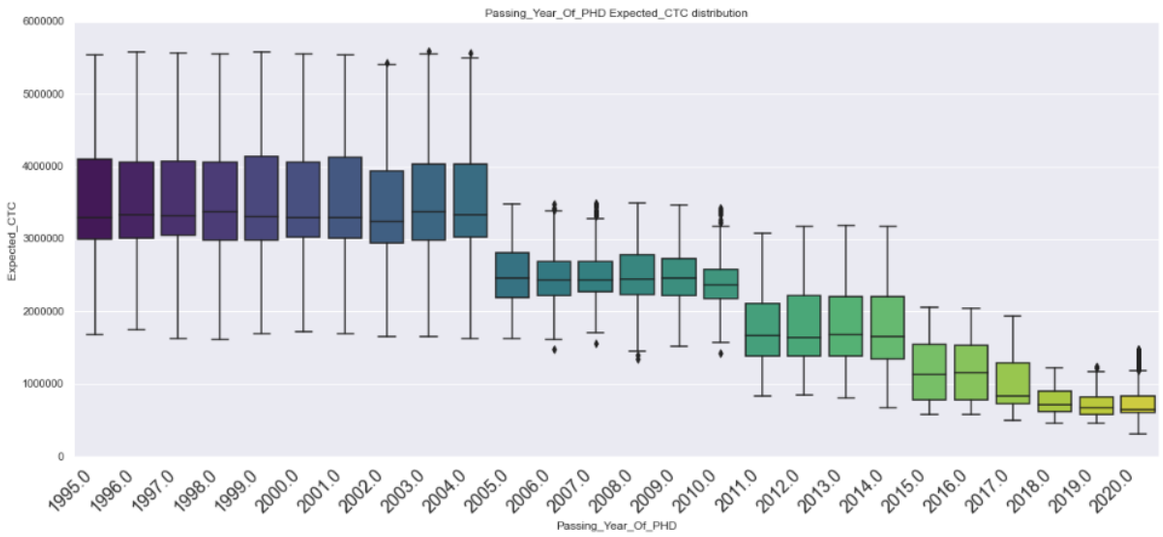


Fig: 46 Box- Plot of Passing_Year_Of_PHD Vs Expected_CTC

Insights -

- Expected_CTC does vary based on the Passing_Year_Of_PHD as expected. This conclusion can only be drawn through the above plot.
- We infer that Expected_CTC for recently PhD passed applicants is less than applicants who completed PhD in early 1990s and 2000s.

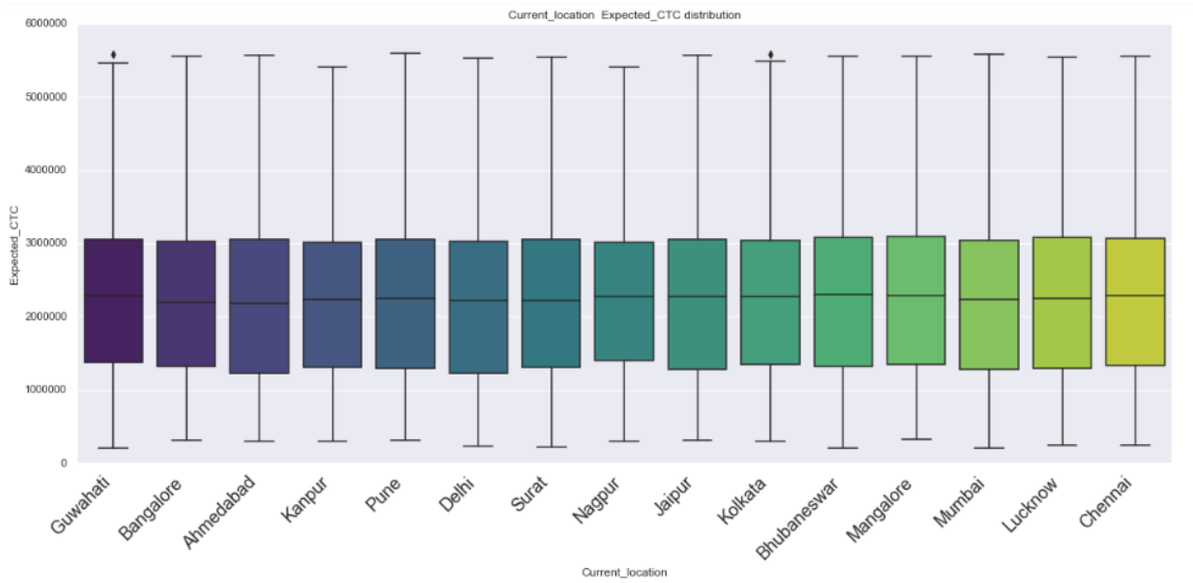


Fig: 47 Box- Plot of Current_location Vs Expected_CTC

Insights -

There is not any variation in the distribution of Expected_CTC w.r.t.Current_Location , looks almost similar plus median values are also almost equivalent for all not varies too much.

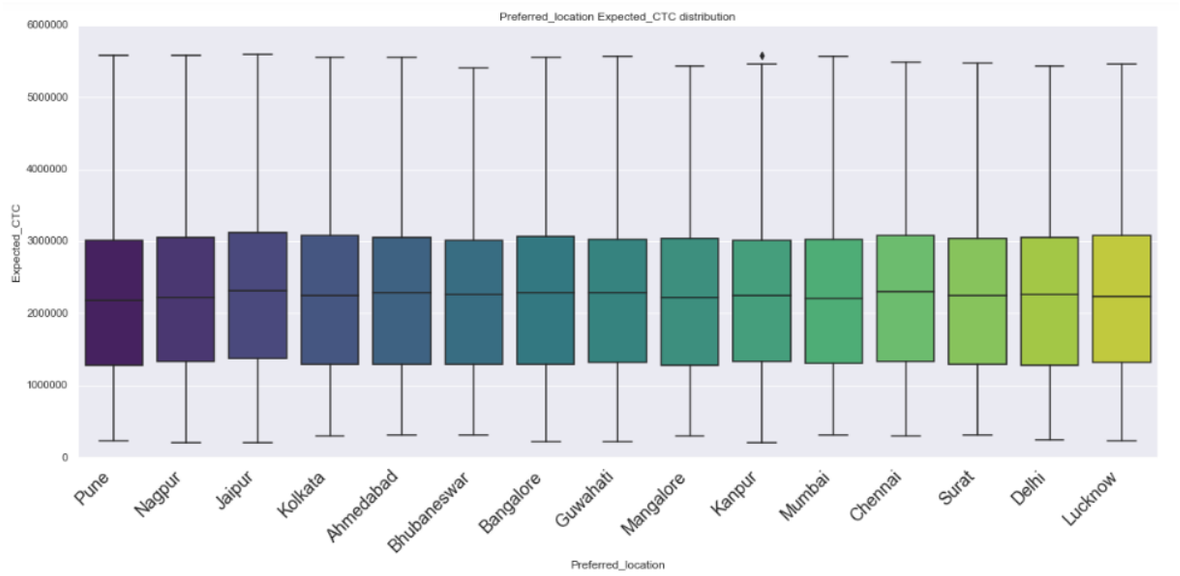


Fig: 48 Box- Plot of Preferred_location Vs Expected_CTC

Insights -

There is not any variation in the distribution of Expected_CTC w.r.t. Preferred_location , looks almost similar plus median values are also almost equivalent for all not varies too much.

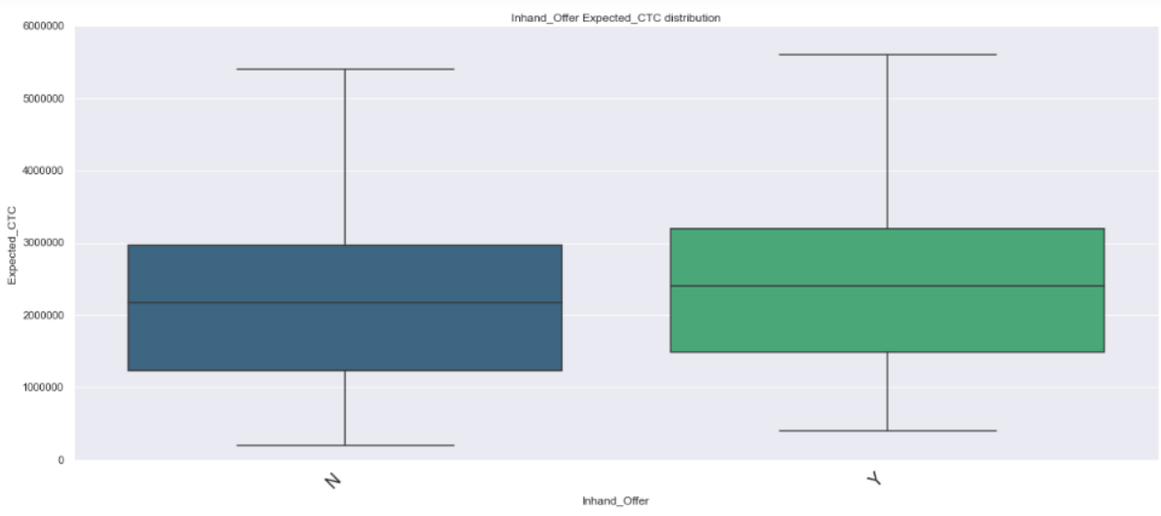


Fig: 49 Box- Plot of Inhand_Offer Vs Expected_CTC

Insights -

Distribution of Expected_CTC for applicants who have offer in hand or who don't have offer in hand is almost similar , but median value for applicants who have in hand offer is slightly high.

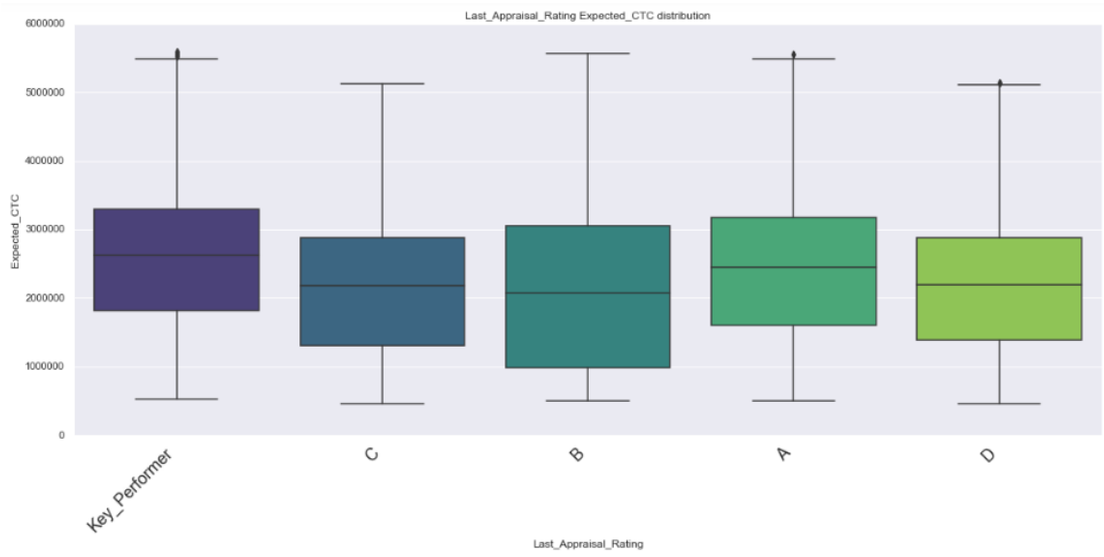


Fig: 50 Box- Plot of Last_Appraisal_Rating Vs Expected_CTC

Insights -

- Median values for Key_Performers are higher than others.
- There is larger distribution for applicants who have B as Last_Appraisal_Rating.
- Median values for applicants who have B and D as Last_Appraisal_Rating are almost equivalent.

Multivariate Analysis :

Heat-map :

A correlation heat-map uses coloured cells, typically in a monochromatic scale, to show a 2D correlation matrix (table) between two discrete dimensions or event types.Correlation heat-maps are ideal for comparing the measurement for each pair of dimension values.Darker Shades have higher Correlation , while lighter shades have smaller values of Correlation as compared to darker shades values.Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

	Total_Experience	Total_Experience_in_field_applied	Passing_Year_Of_Graduation	Passing_Year_Of_PG	Passing_Year_Of_PHD	Current_CTC	No_Of_Companies_worked	I
Total_Experience	1.000000	0.645135	-0.902931	-0.634718	-1.000000	0.846476	0.398135	
Total_Experience_in_field_applied	0.645135	1.000000	-0.581495	-0.410642	-0.648457	0.548017	0.249045	
Passing_Year_Of_Graduation	-0.902931	-0.581495	1.000000	0.841074	0.989101	-0.778366	-0.362545	
Passing_Year_Of_PG	-0.634718	-0.410642	0.841074	1.000000	0.989101	-0.544691	-0.255205	
Passing_Year_Of_PHD	-1.000000	-0.648457	0.989101	0.989101	1.000000	-0.863459	-0.402878	
Current_CTC	0.846476	0.548017	-0.778366	-0.544691	-0.863459	1.000000	0.379740	
No_Of_Companies_worked	0.398135	0.249045	-0.362545	-0.255205	-0.402878	0.379740	1.000000	
Number_of_Publications	-0.000494	-0.010663	-0.336380	-0.491231	0.015752	-0.006399	0.000608	
Certifications	-0.001130	-0.002814	-0.030236	-0.026095	-0.015784	-0.143402	0.012990	
International_degree_any	0.084072	0.043070	-0.085648	-0.066140	-0.083883	0.078774	0.047270	
Expected_CTC	0.816593	0.529115	-0.758694	-0.530964	-0.834222	0.986718	0.343150	
Percentage_Relevant_Exp_in_Field	0.006853	0.656567	-0.003310	-0.000663	-0.014550	0.006166	-0.003106	

	Number_of_Publications	Certifications	International_degree_any	Expected_CTC	Percentage_Relevant_Exp_in_Field
Total_Experience	-0.000494	-0.001130	0.084072	0.816593	0.006853
Total_Experience_in_field_applied	-0.010663	-0.002814	0.043070	0.529115	0.656567
Passing_Year_Of_Graduation	-0.336380	-0.030236	-0.085648	-0.758694	-0.003310
Passing_Year_Of_PG	-0.491231	-0.026095	-0.066140	-0.530964	-0.000663
Passing_Year_Of_PHD	0.015752	-0.015784	-0.083883	-0.834222	-0.014550
Current_CTC	-0.006399	-0.143402	0.078774	0.986718	0.006166
No_Of_Companies_worked	0.000608	0.012990	0.047270	0.343150	-0.003106
Number_of_Publications	1.000000	0.018549	0.016419	0.001518	-0.011520
Certifications	0.018549	1.000000	0.009298	-0.173992	-0.000963
International_degree_any	0.016419	0.009298	1.000000	0.074557	-0.010811
Expected_CTC	0.001518	-0.173992	0.074557	1.000000	0.006404
Percentage_Relevant_Exp_in_Field	-0.011520	-0.000963	-0.010811	0.006404	1.000000

Tab: 33 Correlation Table

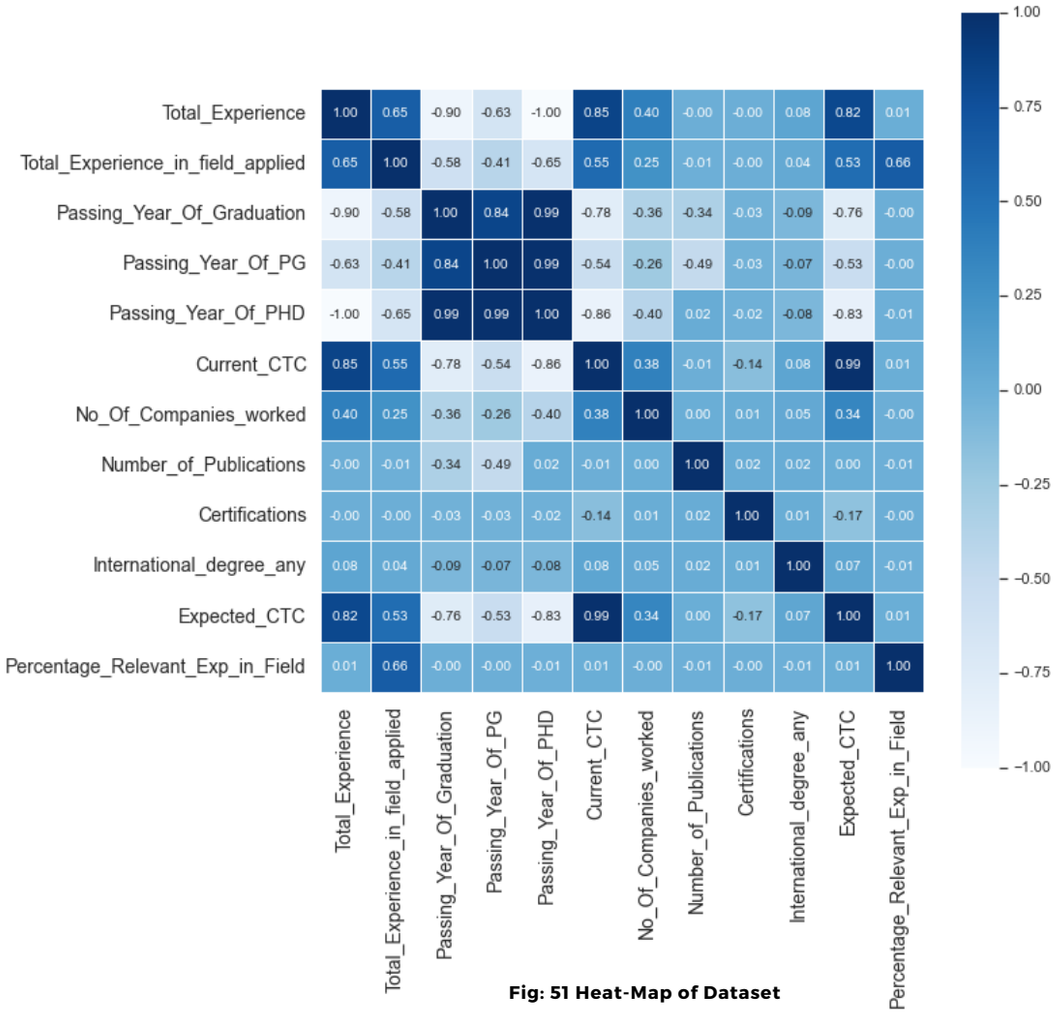


Fig: 51 Heat-Map of Dataset

Insights -

- Total_Experience with Current_CTC , Expected_CTC have strong correlation i.e. (0.85 and 0.82).
- Total_Experience with Total_Experience_in_field_applied have correlation i.e. (0.65).
- Total_Experience with Passing_Year_Of_Graduation shows negative correlation i.e.(-0.90).
- Total_Experience with Passing_Year_Of_PG shows negative correlation i.e.(-0.63).
- Total_Experience with No_Of_Companies_worked shows weak correlation i.e.(0.40).
- Total_Experience_in_field_applied with Expected_CTC shows some correlation i.e. (0.53).
- Passing_Year_Of_Graduation with Expected_CTC shows negative correlation i.e.(-0.76).
- Passing_Year_Of_PG with Expected_CTC shows negative correlation i.e.(-0.53).
- Passing_Year_Of_PHD with Expected_CTC shows negative correlation i.e.(-0.83).
- Current_CTC with Expected_CTC shows very strong correlation i.e (0.99).
- No_Of_Companies_worked with Expected_CTC shows correlation i.e. (0.34).
- Number_of_Publications with Expected_CTC shows no correlation i.e.(0).
- Certifications with Expected_CTC shows negative correlation i.e. (- 0.17).
- International_degree_any with Expected_CTC shows very weak correlation i.e. (0.07).
- Percentage_Relevant_Exp_in_Field with Expected_CTC shows very weak correlation i.e. - (0.01).
- Total_Experience_in_field_applied with Current_CTC shows correlation i.e.(0.55).

- Passing_Year_Of_Graduation with Passing_Year_Of_PHD shows strong correlation i.e.(0.99).
- Passing_Year_Of_Graduation with Passing_Year_Of_PG shows strong correlation i.e. (0.84).
- Passing_Year_Of_Graduation with Current_CTC and Expected_CTC shows negative correlation i.e. (-0.78 and -0.76).
- Rest as there is no issues of strong multi-collinearity only few features have strong correlation with each other we can select out them which suits best as per domain.

Pairplot :

Pairplot shows the relationship between the variables in the form of scatter-plot and the distribution of the variable in the form of histogram.

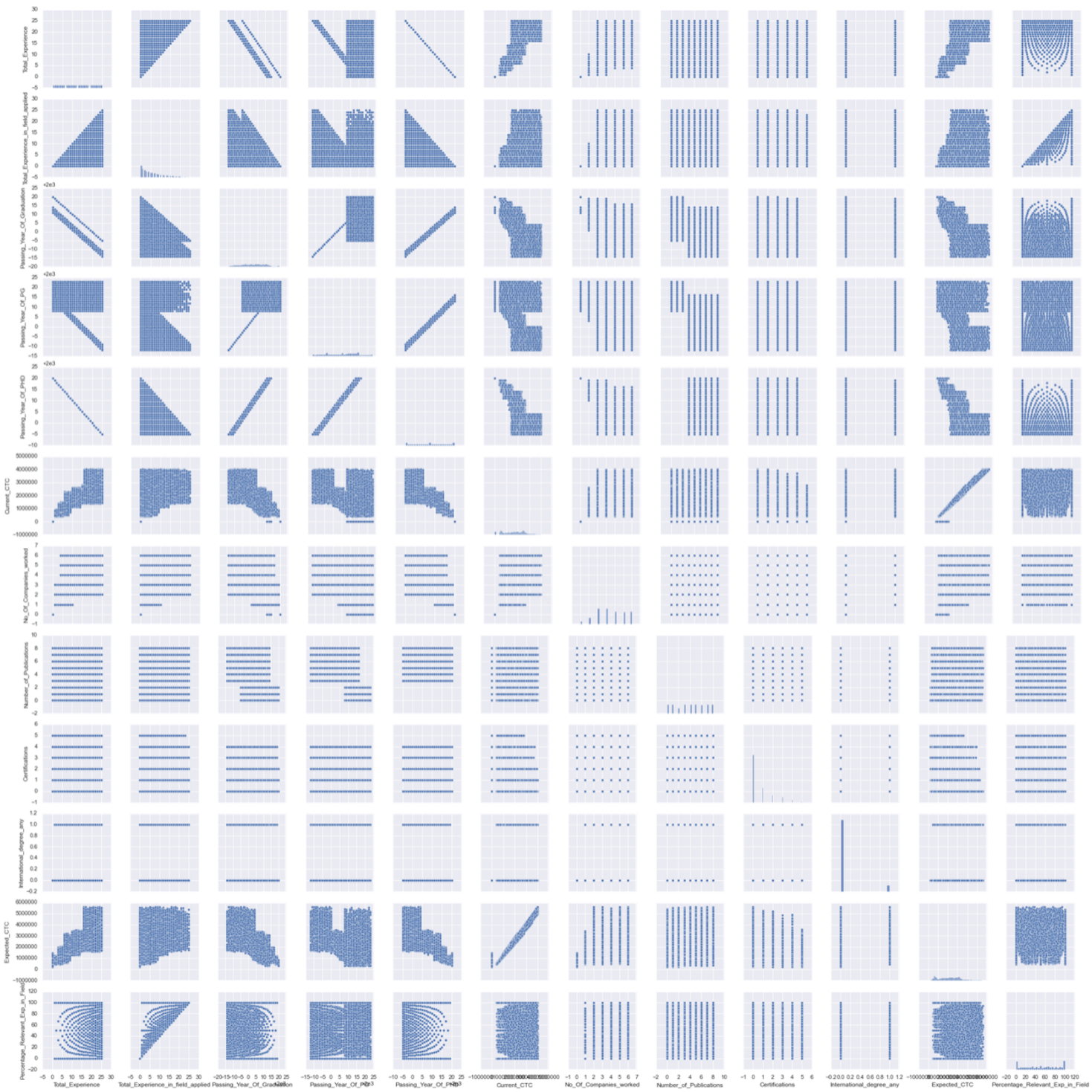
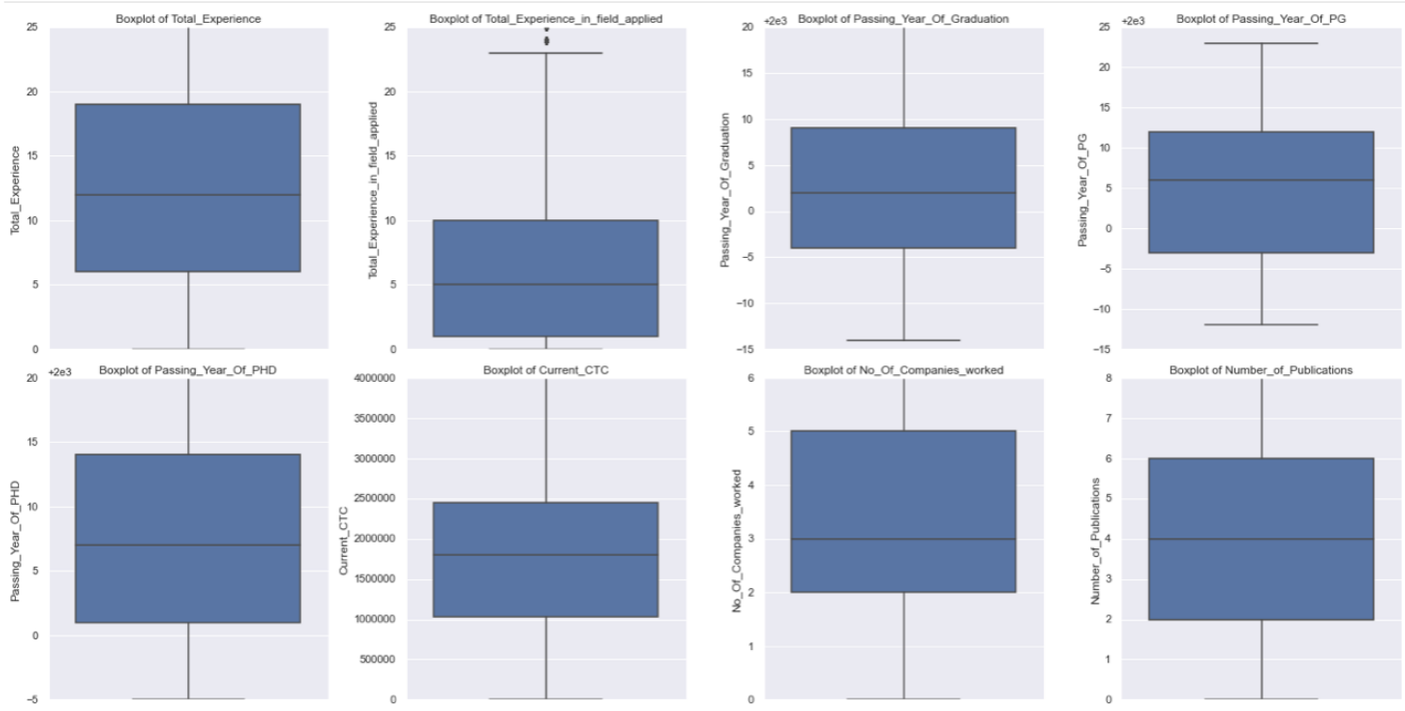


FIG: 52 Pair Plot of Dataset

Insights -

- From the above plot we see that the Total_Experience and the Expected_CTC is showing a strong relationship,with increase in Total_Experience(Independent Variable),Expected_CTC (Target Variable)is also increases.
- From the above plot we see that the Total_Experience_in_field_applied and the Expected_CTC is showing a some relationship,with increase in Total_Experience_in_field_applied(Independent Variable),Expected_CTC (Target Variable)is slightly increases.
- From the above plot we see that the Passing_Year_Of_Graduation and the Expected_CTC is showing a some relationship, as the Passing_Year_Of_Graduation increases the Expected_CTC goes on decreases.
- There is no specific relation between Passing_Year_Of_PG and Expected_CTC.
- From the above plot we see that the Passing_Year_Of_PHD and the Expected_CTC is showing a negative relationship, as the Passing_Year_Of_PHD increases the Expected_CTC goes on decreases.
- From the above plot we see that the Current_CTC and the Expected_CTC is showing a strong relationship, as the Current_CTC increases the Expected_CTC goes on increases.
- From the above plot we infer that there is some sort of relation between No_Of_Companies_worked and Expected_CTC.As No_Of_Companies_worked increase there is also some increase in Expected_CTC.
- There is no such relationship between Number_of_Publications and Expected_CTC.From the above visual we infer there is not any kind of impact of Number_of_Publications on Expected_CTC.Expected_CTC is somehow equivalent for all who have less or more Number_of_Publications.
- From the above plot we see that the Percentage_Relevant_Exp_in_Field and the Expected_CTC is showing a no relationship as all the data-points are scatter over plane.

Checking for Outliers in the dataset -



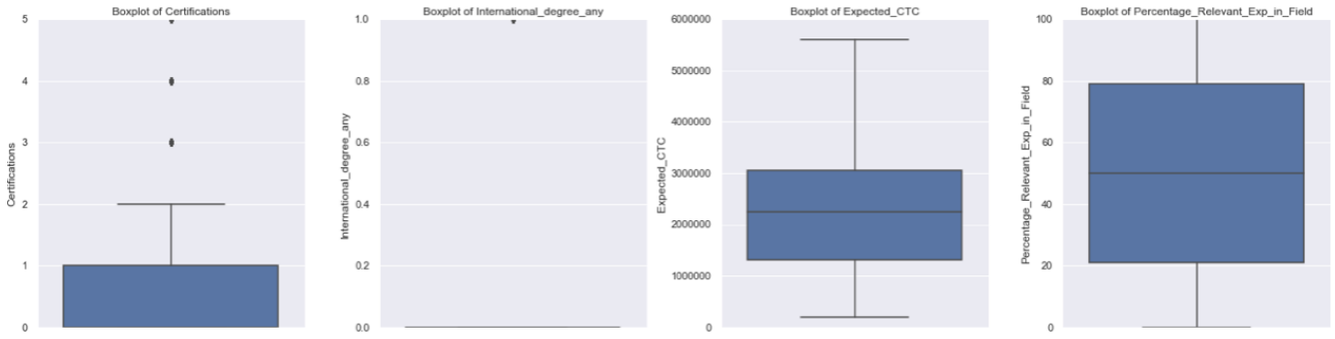


FIG: 53 Outlier Detection

Insights -

Looking at the box plot, it seems that the only Total_Experience_in_field_applied , Certifications and International_degree_any variables have afew outliers , others don't have outliers.

Data Cleaning and Pre-Processing

Removal of unwanted variables

We are dropping the column **IDX and Applicant_Id** as these columns didn't contribute for analysis and model building exercise , because IDX and Applicant_ID for each applicant is unique hence it is useless for the model.That's why we decided to drop these two columns.

	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	Graduation_Specialization	University_Grad	Passing_Year_Of_Graduation
0	0	0	NaN	NaN	NaN	NaN	NaN	PG	Arts	Lucknow	2020.0
1	23	14	HR	Consultant	Analytics	H	HR	Doctorate	Chemistry	Surat	1988.0
2	21	12	Top Management	Consultant	Training	J	NaN	Doctorate	Zoology	Jaipur	1990.0
3	15	8	Banking	Financial Analyst	Aviation	F	HR	Doctorate	Others	Bangalore	1997.0
4	10	5	Sales	Project Manager	Insurance	E	Medical Officer	Grad	Zoology	Mumbai	2004.0
5	16	3	Top Management	Area Sales Manager	Retail	G	Director	Doctorate	Others	Bangalore	1998.0
6	1	1	Engineering	Team Lead	FMCG	L	Marketing Manager	Grad	Chemistry	Delhi	2011.0
7	19	11	Others	Analyst	Others	E	Manager	PG	Sociology	Delhi	2001.0
8	8	7	Analytics/BI	Others	Telecom	L	Marketing Manager	Doctorate	Psychology	Mumbai	2003.0
9	15	15	Analytics/BI	CEO	Telecom	M	Product Manager	Doctorate	Chemistry	Delhi	1998.0

Tab:34 Dataset After Dropping Unwanted Columns

Observation: Now we have all the columns which are useful for the model.

Missing Value Treatment

S.No.	Features / Columns	Null Count
1	Total_Experience	0
2	Total_Experience_in_field_applied	0
3	Department	2778
4	Role	963
5	Industry	908
6	Organization	908
7	Designation	3129
8	Highest Education	0
9	Graduation_Specialization	6180
10	University_Grad	6180
11	Passing_Year_Of_Graduation	6180
12	PG_Specialization	7692
13	University_PG	7692
14	Passing_Year_Of_PG	7692
15	PHD_Specialization	11881
16	University_PHD	11881
17	Passing_Year_Of_PHD	11881
18	Current_location	0
19	Preferred_location	0
20	Current_CTC	0
21	Inhand_Offer	0
22	Last_Appraisal_Rating	908
23	No_Of_Companies_worked	0
24	Number_of_Publications	0
25	Certifications	0
26	International_degree_any	0
27	Expected_CTC	0

Tab:35 Checking Null Values.

Observation -

- By looking at the above results we found that Graduation_Specialization , University_Grad and Passing_Year_Of_Graduation have same number of missing values (6180) which indicates that there is data is missed.
- By looking at the above results we found that PG_Specialization , University_PG and Passing_Year_Of_PG have same number of missing values (7692) which indicates that there is data is missed or these applicants didn't have PG education.
- By looking at the above results we found that PHD_Specialization , University_PHD and Passing_Year_Of_PHD have same number of missing values (11881) which indicates that there is data is missed or these applicants didn't have PG education.
- By looking at the above results we found that Industry and Organization have same number of missing values (908) which indicates that the for these applicants in terms of Industry and Organization is data is unknown.
- Role & Department also have null values which indicates that for these applicants data is also unknown.

Practice -

For numerical features we will going impute Passing_Year_Of_Graduation with meadian and for Passing_Year_Of_PG and Passing_Year_Of_PHD we impute the missing values with 0 by using fillna( ) function as these applicants might not have PG / PHD education.

For categorical features we are going to use fillna ( ) function and impute unknown label in place of null values as data is given because here we can cannot impute with mode because most records is a pattern of missing values and data is missing, if we impute with mode so model will not perform well ,so it be good practicesper the problem .Because in note 2/milestone when we encode them for model building it will be easy to club them or encode them by using target encoding or mean encoding method.

S.No.	Features / Columns	Null Count
1	Total_Experience	0
2	Total_Experience_in_field_applied	0
3	Department	0
4	Role	0
5	Industry	0
6	Organization	0
7	Designation	0
8	Highest Education	0
9	Graduation_Specialization	0
10	University_Grad	0
11	Passing_Year_Of_Graduation	0
12	PG_Specialization	0
13	University_PG	0
14	Passing_Year_Of_PG	0
15	PHD_Specialization	0
16	University_PHD	0
17	Passing_Year_Of_PHD	0
18	Current_location	0
19	Preferred_location	0
20	Current_CTC	0
21	Inhand_Offer	0
22	Last_Appraisal_Rating	0
23	No_Of_Companies_worked	0
24	Number_of_Publications	0
25	Certifications	0
26	International_degree_any	0
27	Expected_CTC	0

Tab:36 Checking Null Values After Imputation

Conclusion -

We successfully imputed all the null values present in the dataset with suitable values as per the context of the business problem. Now we don't have any null values in the dataset.

Feature Selection- Based on Correlation - For Numerical Feature

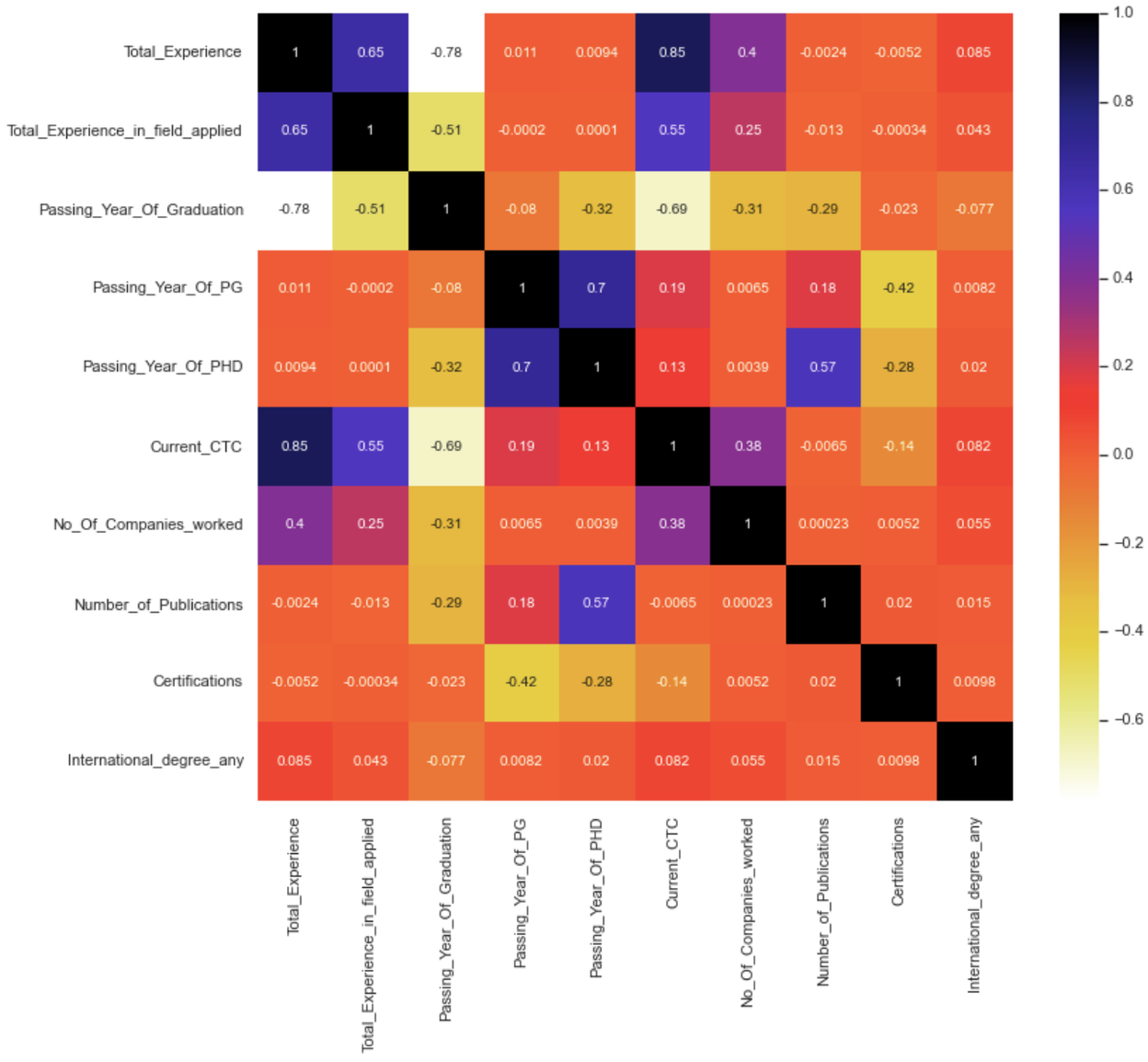


FIG: 54 Check Collinearity Among Features

Observations-

On the basis of Pearson's Correlation Feature Selection method , we come know that these 4 features
{'Current_CTC',
'Passing_Year_Of_Graduation',
'Passing_Year_Of_PHD',
'Total_Experience_in_field_applied'}
are correlated with each other so we can remove them or compare with the variables with which they are correlated and drop as per the context of business / domain knowledge.By this method we come know that about the Multicollinearity also , so we can drop those feature which are highly correlated.This is our basic approach for feature selection rest in next milestone when we going to built linear regression model then by using p-value of OLS summary we can pick features also or we can use another sk-learn automatic functions for feature selection for model building . Till now we used this as we did not built our base model yet beacuse in this excercise we just need to explore data do pre-processing , do some treatment of missing values and visualization for getting the insights.

Feature Selection - For Categorical Feature (Dropping unusual Categorical Columns)

From the above box-plot visuals we saw that many of the categorical feature like Industry , Organization , Graduation_Specialization ,University_Grad ,PG_Specialization , University_PG ,PHD_Specialization , University_PHD ,Current_location , and Preferred_location any of these variables is not showing any variation with the target variable (Expected_CTC) and there is no specific relation between them and target so we decided to drop these variables because they will not possess any impact on the model even it is good practice to remove such variables which are not in relationship with the target variable or helps us to predict the dependent variable.By doing this we also reduces the dimensionality of the dataset as these features don't have any impact on target feature.Now we left 6 categorical features like - Department , Role , Designation , Highest Education , Inhand_Offer and Last_Appraisal_Rating which we are going to use for model building.

Outlier treatment -

An observation is considered to be an outlier if that particular has been mistakenly captured in the data set. Treating outliers sometimes results in the models having better performance but the models lose out on the generalization. So, a good way to approach this would be to build models with and without treating outliers and then report the results. So we are only check the outliers but not treat them as per context of the problem given.

Variable transformation -

As we saw in the skewness table that our target column Expected_CTC columns have skewness value - 0.33 and by looking at the histogram and boxplot we also found it is almost normal distributed.So right now we make an assumption that data is normal. Once we build the first model and check its performance if anything is needed then we go for transformation.

Featuring Engineering Addition of new variables -

Yes , we made a new feature also named as Percentage_Relevant_Exp_in_Field (df_1["Percentage_Relevant_Exp_in_Field"]=round(df_1["Total_Experience_in_field_applied"]/df_1["Total_Experience"]*100)) but this feature is not impacting too much as per heatmap / correaltion.So,when we build our linear regression model check its p-value if it is not worthy then we can drop it.

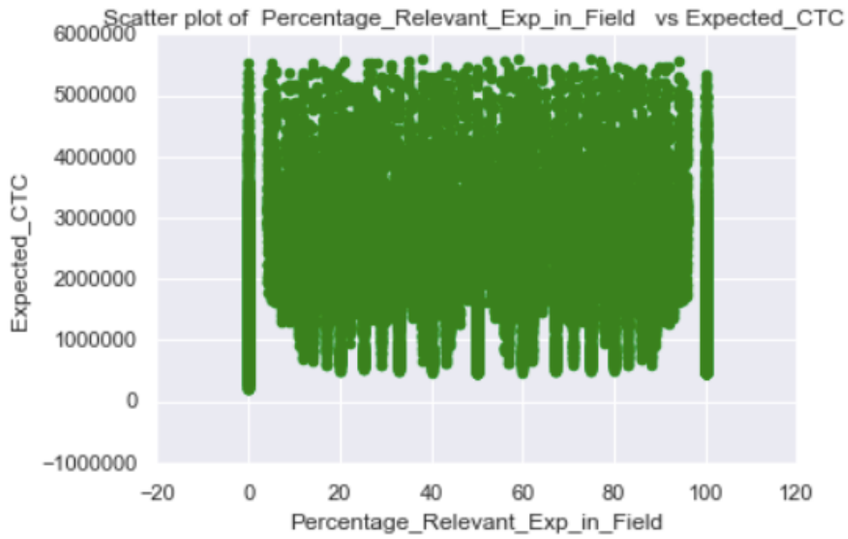


FIG: 55 Scatter plot of Percentage_Relevant_Exp_in_Field vs Expected_CTC

Insights -

From the above plot we see that the Percentage_Relevant_Exp_in_Field and the Expected_CTC is showing a no relationship as all the data-points are scatter over plane.

Is the data unbalanced? If so, what can be done? Please explain in the context of the business -

No , there is no data unbalanced problem as we predicting the expected_ctc for applicants being a regression problem our target column is continous doesnot have class so in the this probelm we donot have any data unblaced problem.

Encoding & Combining of the Sublevels/lables for the Categorical Variables.

In the given data set we left with 6 categorical features named as Department , Role, Designation, Highest Education , Inhand_Offer and Last_Appraisal_Rating. Now we are trying to combine the sub-levels of the categorical features as they have large number of labels and then after combining the labels we do the label encoding on them .

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning methods.

"Department"

S.no.	Department	Value_Counts
1	unknown	2778
2	Marketing	2379
3	Analytics/BI	2096
4	Healthcare	2062
5	Others	2041
6	Sales	1991
7	HR	1988
8	Banking	1952
9	Education	1948
10	Engineering	1937
11	Top Management	1632
12	Accounts	1118
13	IT-Software	1078

Tab: 37 Value Counts for Categorical Feature (Department)

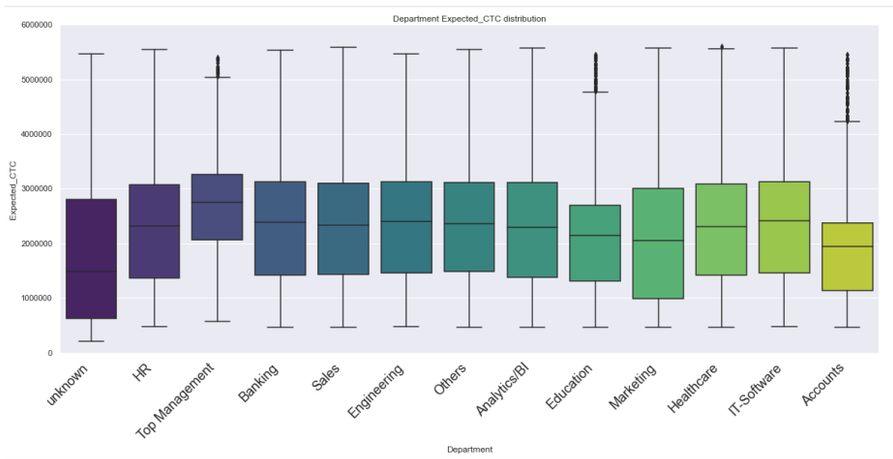


Fig: 56 Box- Plot of Department Vs Expected_CTC

Note -

As we saw in the above figure median values for expected_ctc of HR, Banking , Sales , Engineering, Others,Anlaytics/BI, Healthcare and IT-Software and Marketing is almost same / nearby so we can combine them into one and named them as mid_level_dept. As we saw in the plot Top Managemnet department applicants have higher median values than other so we can named them as top_level_dept.Education and Accounts dept have nearby median values so we can combine them as named as low_level_dept and unknown have least median value among all so we can name this as very_low_dept.So here we combine them first and then do label encoding on them.

S.no.	Department	Value_Counts	Encoding
1	very_low_level_dept	2778	0
2	low_level_dept	3066	1
3	mid_level_dept	17524	2
4	top_level_dept	1632	3

Tab: 38 Department Table After Combining & Encoding of the Labels

"Role"

S.no.	Role	Value_Counts
1	Others	2248
2	Bio statistician	1913
3	Analyst	1892
4	Project Manager	1850
5	Team Lead	1833
6	Consultant	1780
7	Business Analyst	1711
8	Sales Execituve	1574
9	Sales Manager	1427
10	Senior Researcher	1236
11	Financial Analyst	1182
12	CEO	1149
13	Scientist	1139
14	Head	1108
15	unknown	963
16	Associate	767
17	Data scientist	363
18	Principal Analyst	275
19	Area Sales Manager	134
20	Senior Analyst	128
21	Researcher	123
22	Sr. Business Analyst	114
23	Research Scientist	33
24	Professor	33
25	Lab Executuve	25

Tab: 39 Value Counts for Categorical Feature (Role)

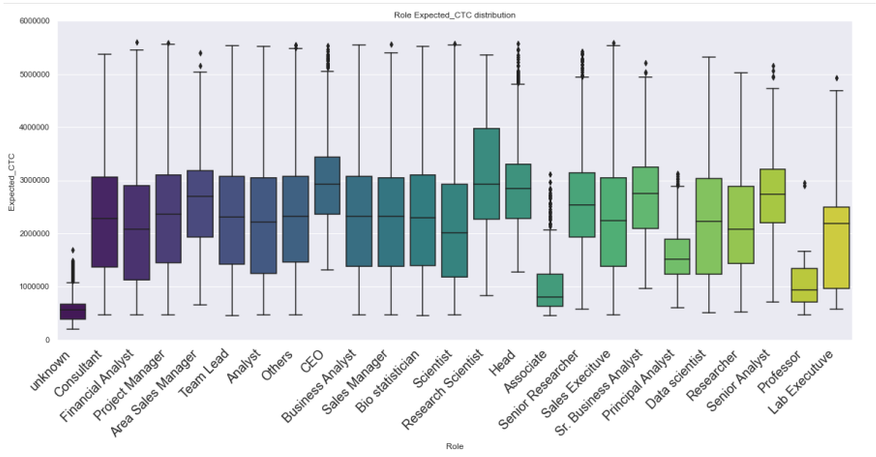


Fig: 57 Box- Plot of Role Vs Expected_CTC

Note -

As we saw in the above figure median values for expected_ctc of HR, Banking , Sales , Engineering, Others,Anlaytics/BI, Healthcare and IT-Software and Marketing is almost same / nearby so we can combine them into one and named them as mid_level_dept. As we saw in the plot Top Managemnet department applicants have higher median values than other so we can named them as top_level_dept.Education and Accounts dept have nearby median values so we can combine them as named as low_level_dept and unknown have least median value among all so we can name this as very_low_dept.So here we combine them first and then do label encoding on them.

S.no.	Department	Value_Counts	Encoding
1	very_low_level_dept	2778	0
2	low_level_dept	3066	1
3	mid_level_dept	17524	2
4	top_level_dept	1632	3

Tab: 40 Department Table After Combining & Encoding of the Labels

S.no.	Role	Value_Counts
1	Others	2248
2	Bio statistician	1913
3	Analyst	1892
4	Project Manager	1850
5	Team Lead	1833
6	Consultant	1780
7	Business Analyst	1711
8	Sales Execituve	1574
9	Sales Manager	1427
10	Senior Researcher	1236
11	Financial Analyst	1182
12	CEO	1149
13	Scientist	1139
14	Head	1108
15	unknown	963
16	Associate	767
17	Data scientist	363
18	Principal Analyst	275
19	Area Sales Manager	134
20	Senior Analyst	128
21	Researcher	123
22	Sr. Business Analyst	114
23	Research Scientist	33
24	Professor	33
25	Lab Executuve	25

Tab: 41 Value Counts for Categorical Feature (Role)

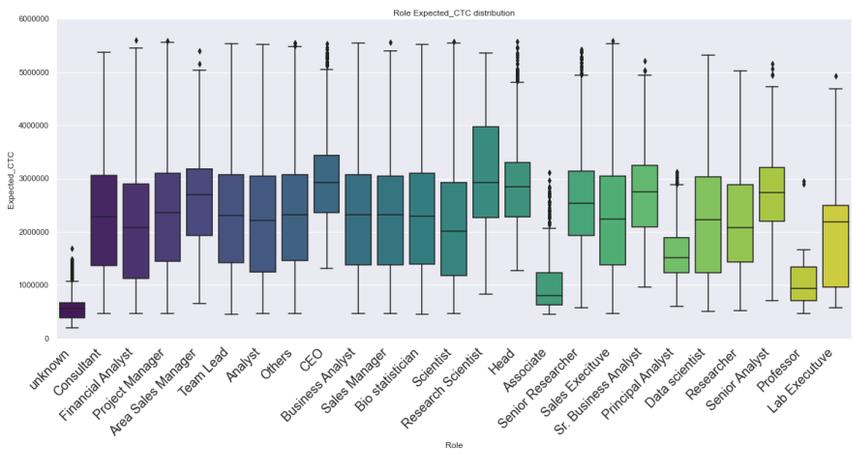


Fig: 58 Box- Plot of Role Vs Expected_CTC

Note -

As we saw in the above figure median values for expected_ctc of CEO , Research Scientist ,Head , Area Sales Manager , Senior Buisnes Analyst , Senior Researcher and Senior Analyst is quite nearby so we can combine them and name them as top_level_roles. Secondly, we saw that Consultant ,Financial Analyst ,Project Manager ,Team Lead ,Analyst ,Others ,Business Analyst ,Sales Manager ,Bio statistician , Scientist ,Sales Executive ,Data Scientist , Researcher and Lab Executives have near by median values so we can combine them and name them as mid_level_roles.Principal Analyst have low median value than top_level_roles and mid_level_roles so we can name them as mid_low_level_roles.Associate and Professor have lower median values than all other so we combine them and name them as low_level_roles. Atlast we have unknown which have least median among all so we can name them as extremely_low_level_roles.

S.no.	Role	Value_Counts	Encoding
1	extremely_low_level_roles	963	0
2	low_level_roles	800	1
3	mid_low_level_roles	275	2
4	mid_level_roles	19060	3
5	top_level_roles	3902	4

Tab: 42 Role Table After Combining & Encoding of the Labels

"Designation"

S.no.	Designation	Value_Counts
1	Unknown	3129
2	HR	1648
3	Others	1647
4	Manager	1628
5	Product Manager	1626
6	Sr.Manager	1617
7	Consultant	1606
8	Assistant Manager	1590
9	Marketing Manager	1590
10	Data Analyst	1575
11	Research Analyst	1563
12	Medical Office	1047
13	Software Developer	914
14	Web Designer	882
15	Network Engineer	862
16	Director	772
17	CA	715
18	Research Scientist	537
19	Scientist	52

Tab: 43 Value Counts for Categorical Feature (Designation)

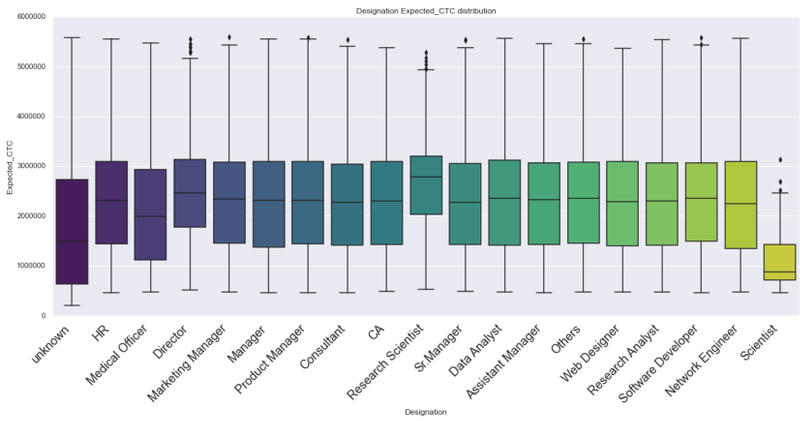


Fig: 59 Box- Plot of Designation Vs Expected_CTC

Note-

As we saw in the above figure median values for expected_ctc of Research Scientist is high as compared to all so we name it as top_designation.Secondly, we found that HR, Markerting Manager ,Director , Manager , Product Manager ,Consultant, CA ,Sr.Manger , Data Analyst ,Assistant Manager ,Others ,Web Desginers, Research Analyst ,Software Developer ,have almost nearby median values so we can combine them and name them mid_designation.Medical Office and Network Engineer has almost nearby median values so we can combine them and name them as mid_low_designation.Then we unknown so we can name them as low_designation as their median values is lower than above two.Atlast we have scientist which has least median value so we can name them as extremely_low_designation.

S.no.	Designation	Value_Counts	Encoding
1	extremely_low_designation	52	0
2	low_designation	3129	1
3	mid_low_designation	1909	2
4	mid_designation	19373	3
5	top_designation	537	4

Tab: 44 Designation Table After Combining & Encoding of the Labels

"Highest_Education"

S.no.	Highest_Education	Value_Counts
1	Under_Grad	6180
2	Grad	6209
3	PG	6326
4	Doctorate	6285

Tab: 45 Value Counts for Categorical Feature (Highest_Education)

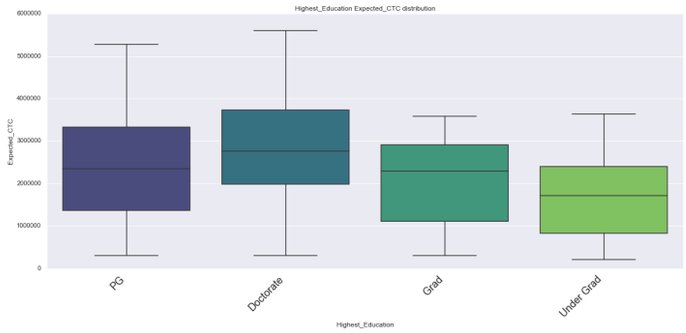


Fig: 60 Box- Plot of Highest Education Vs Expected_CTC

Note-

Here , we do the ordinal label encoding as simple we encode Under_Grad as 0 , Grad as 1 , PG as 2 and Doctorate as 3 as per their order.

S.no.	Highest_Education	Value_Counts	Encoding
1	Under_Grad	6180	0
2	Grad	6209	1
3	PG	6326	2
4	Doctorate	6285	3

Tab: 46 Highest_Education Table After Encoding of the Labels

"Inhand_Offer"

S.no.	Inhand_Offer	Value_Counts
1	N	17418
2	Y	7582

Tab: 47 Value Counts for Categorical Feature (Inhand_Offer)

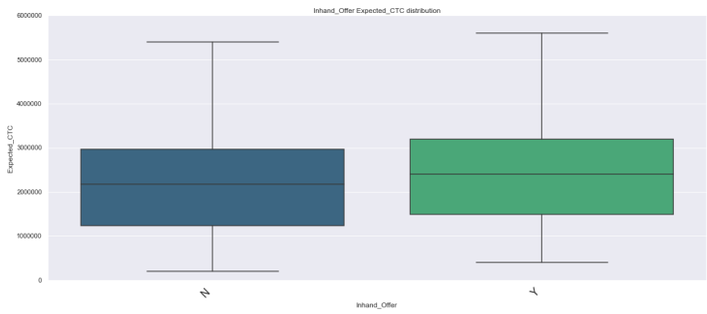


Fig: 61 Box- Plot of Inhand_Offer Vs Expected_CTC

Note-

Here , we do the label encoding as simple we encode N as 0 , Y as 1.

S.no.	Inhand_Offer	Value_Counts	Encoding
1	N	17418	0
2	Y	7582	1

Tab: 48 Inhand_Offer Table After Encoding of the Labels

"Last_Appraisal_Rating"

S.no.	Last_Appraisal_Rating	Value_Counts
1	Unknown	908
2	D	4917
3	C	4812
4	B	5501
5	A	4671
6	Key_Performer	4191

Tab: 49 Value Counts for Categorical Feature (Last_Appraisal_Rating)

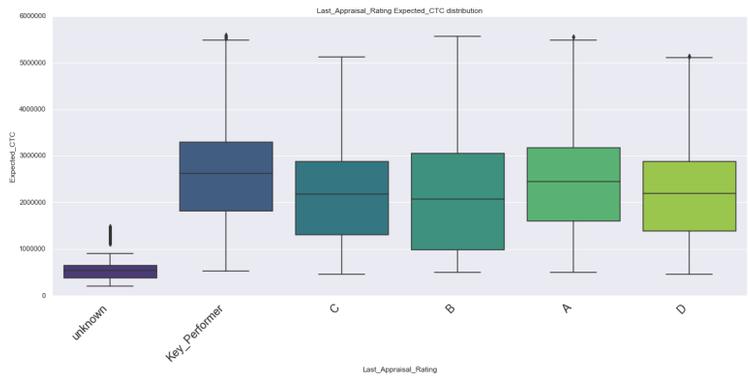


Fig: 62 Box- Plot of Last_Appraisal_Rating Vs Expected_CTC

Note -

As we saw that C and D have same median values for expected_ctc so we combine C and D and name them as C.

S.no.	Last_Appraisal_Rating	Value_Counts	Encoding
1	Unknown	908	0
2	C	9729	1
3	B	5501	2
4	A	4671	3
5	Key_Performer	4191	4

Tab: 50 Last_Appraisal_Rating Table After Combining & Encoding of the Labels

Checking the Dataset after Encoding

	Total_Experience	Total_Experience_in_field_applied	Department	Role	Designation	Highest_Education	Passing_Year_Of_Graduation	Passing_Year_Of_PG	Passing_Year_Of_PHD	Current_CTC
0	0	0	0	0	1	2	2020.0	0.0	0.0	0
1	23	14	2	3	3	3	1988.0	1990.0	1997.0	2702664
2	21	12	3	3	1	3	1990.0	1992.0	1999.0	2236661
3	15	8	2	3	3	3	1997.0	1999.0	2005.0	2100510
4	10	5	2	3	2	1	2004.0	2006.0	2010.0	1931644

Tab: 51 Checking the Dataset after Encoding

Checking the Appropriateness of Datatypes & Information of the Dataframe after Encoding :

The info() function is used to print a concise summary of a DataFrame. This method prints information about a DataFrame including the index d-type and column d-types, non-null values and memory usage.

S.no.	Features / Columns	Non-Null Count	D-Type
0	Total_Experience	25000 non-null	int64
1	Total_Experience_in_field_applied	25000 non-null	int64
2	Department	25000 non-null	Int64
3	Role	25000 non-null	int64
4	Designation	25000 non-null	int64
5	Highest Education	25000 non-null	int64
6	Passing_Year_Of_Graduation	25000 non-null	float64
7	Passing_Year_Of_PG	25000 non-null	float64
8	Passing_Year_Of_PHD	25000 non-null	Float64
9	Current_CTC	25000 non-null	Int64
10	Inhand_Offer	25000 non-null	float64
11	Last_Appraisal_Rating	25000 non-null	int64
12	No_Of_Companies_worked	25000 non-null	int64
13	Number_of_Publications	25000 non-null	int64
14	Certifications	25000 non-null	int64
15	International_degree_any	25000 non-null	int64
16	Expected_CTC	25000 non-null	int64
17	Percentage_Relevant_Exp_in_Field	25000 non-null	float64

Tab: 52 Appropriateness of Datatypes & Information of the Dataframe after Encoding

Insights -

From the above results we can see that there are no null values present in dataset.Their are total 25000 rows & 18 columns are in this dataset,indexed from 0 to 24999. Out of 18 variables 4 are float64 , and 14 variable are int64 d-type. Memory used by the dataset: 3.4 MB.

Note-

Now all the features in the numerical form and we can use them to build various machine learning regression model to predict the expected_ctc of the applicant.

Model Building

In this model building exercise we will build different regression models like Linear Regression , XG-Boost Regressor , Decision Tree , Random Forest and ANN Regressor models to predict the Expected_CTC for applicant who are applying/ joining for different roles in the Delta Ltd.The main objective of this problem is to provides salary estimates at time of joining for applicants based on their job title, location, years of experience, skill and profile to minimise human judgment with regard to salary to be offered. It is imperative to provide an unbiased salary for an employee which he/she truly deserves, and also has to be appropriate to the market demands.

Note:

Before proceeding to the model building we need to split the data-set into train and test set.Then we apply the supervised regression algorithm to the training set and check the prediction on test set.We are doing the train-test split down the line.

Train-Test Split for Regression Models -

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets.

In the given problem, we are advised to split the training and the testing data in the ratio of (70: 30).Here we are split the data into train and test part , like x_train , x_test , train_labels & test_labels ,by using train_test_split func() from sk-learn library here ,we are taking 70 % data for training and 30 % data for testing.

Model 1 - Linear Regression Model

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

[1] This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

[2]In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

[3] Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.

[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response

- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

- Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Note :

A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. ... Continuous predictor variables are also called covariates, and categorical predictor variables are also called factors.Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable.

Building A Linear Regression Model -

Let's start with building a linear model. Instead of simple linear regression, where you have one predictor and one outcome, we will go with multiple linear regression, where you have more than one predictors and one outcome.

Multiple linear regression follows the formula :

$$yi = \beta0 + \beta1xi1 + \beta2xi2 + ... + \beta pxip + \epsilon$$

Where:

- yi is the dependent or predicted variable.
- β0 is the y-intercept, i.e., the value of y when both xi and x2 are 0.
- β1 and β2 are the regression coefficients representing the change in y relative to a one-unit change in xi1 and xi2, respectively.
- βp is the slope coefficient for each independent variable.
- ϵ is the model's random error (residual) term.

The coefficients in this linear equation denote the magnitude of additive relation between the predictor and the response. In simpler words, keeping everything else fixed, a unit change in x1 will lead to change of β1 in the outcome, and so on.

Invoke the Linear Regression function (from sklearn.linear_model import LinearRegression) fit the function on the train & test data and build the linear regression model. In this problem we are advised to build various linear regression model and check the performance of Predictions on Train and Test sets using R-square, RMSE & Adj R-square & atlast we need to compare these models and select the best one.

Model 1- Linear Regression (By Sklearn library)

Explore the coefficients for each of the independent attributes.

- The coefficient for Total_Experience is -5708.371196039307
- The coefficient for Total_Experience_in_field_applied is 7627.991833760015
- The coefficient for Department is -26964.332458536745
- The coefficient for Role is -92250.08636479377
- The coefficient for Designation is -34531.52969682892
- The coefficient for Highest_Education is 91983.43367532847
- The coefficient for Passing_Year_Of_Graduation is -3810.3385464441985
- The coefficient for Passing_Year_Of_PG is -28.22170310228962
- The coefficient for Passing_Year_Of_PHD is -15.403892960030497
- The coefficient for Current_CTC is 1.250672063596419
- The coefficient for Inhand_Offer is 40181.77232177193
- The coefficient for Last_Appraisal_Rating is 69501.37686798647
- The coefficient for No_Of_Companies_worked is -10873.972069499954
- The coefficient for Number_of_Publications is 4492.3749344316475
- The coefficient for Certifications is 612.4293931556061
- The coefficient for International_degree_any is 36401.78854350737
- The coefficient for Percentage_Relevant_Exp_in_Field is -1275.8616627535325

The coefficients in this linear equation denote the magnitude of additive relation between the predictor and the response. In simpler words, keeping everything else fixed, a unit change in x1 will lead to change of β1 in the outcome, and so on.

Intercept for the model -

The regression constant is also known as the intercept thus, regression models without predictors are also known as intercept only models.As such, we will begin with intercept only models for OLS regression and then move on to logistic regression models without predictors. Intercept for the model is **7951913.5554777.**

R square - is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. 100% indicates that the model explains all the variability of the response data around its mean.

R square on Training Data - 0.9893059671397622
R square on Testing Data - 0.9897924521445122

R-Squared value of 0.989 would indicate that 98.9% of the variance of the dependent variable being studied is explained by the variance of the independent variable.

RMSE - The root mean square error (RMSE) for a regression model is similar to the standard deviation (SD) for the ideal measurement model. The SD estimates the deviation from the sample mean x. The RMSE estimates the deviation of the actual y-values from the regression line.

Root mean square error, which is a metric that tells us the average distance between the predicted values from the model and the actual values in the dataset. The lower the RMSE, the better a given model is able to “fit” a dataset.

RMSE on Training data - 119545.42733621509
RMSE on Testing data - 118283.86008401512

Model 1 - Linear Regression using Stats-Model ols

This is the expression of Model 1 for Training Data

expr= 'Expected_CTC ~ Total_Experience + Total_Experience_in_field_applied + Department + Role + Designation + Highest_Education + Passing_Year_Of_Graduation + Passing_Year_Of_PG + Passing_Year_Of_PHD + Current_CTC + Inhand_Offer + Last_Appraisal_Rating + No_Of_Companies_worked + Number_of_Publications + Certifications +International_degree_any + Percentage_Relevant_Exp_in_Field'

Explore the coefficients for each of the independent attributes of train data.

- The coefficient for Total_Experience is -5708.371196039307
- The coefficient for Total_Experience_in_field_applied is 7627.991833760015
- The coefficient for Department is -26964.332458536745
- The coefficient for Role is -92250.08636479377
- The coefficient for Designation is -34531.52969682892
- The coefficient for Highest_Education is 91983.43367532847
- The coefficient for Passing_Year_Of_Graduation is -3810.3385464441985
- The coefficient for Passing_Year_Of_PG is -28.22170310228962
- The coefficient for Passing_Year_Of_PHD is -15.403892960030497
- The coefficient for Current_CTC is 1.250672063596419
- The coefficient for Inhand_Offer is 40181.77232177193
- The coefficient for Last_Appraisal_Rating is 69501.37686798647
- The coefficient for No_Of_Companies_worked is -10873.972069499954
- The coefficient for Number_of_Publications is 4492.3749344316475
- The coefficient for Certifications is 612.4293931556061
- The coefficient for International_degree_any is 36401.78854350737
- The coefficient for Percentage_Relevant_Exp_in_Field is -1275.8616627535325

The coefficients in this linear equation denote the magnitude of additive relation between the predictor and the response. In simpler words, keeping everything else fixed, a unit change in x1 will lead to change of β1 in the outcome, and so on.

Intercept for the model on train data -

The regression constant is also known as the intercept thus, regression models without predictors are also known as intercept only models. ... As such, we will begin with intercept only models for OLS regression and then move on to logistic regression models without predictors. Intercept for the model is **7.951914e+06.**

OLS Regression Results of train data -

OLS is a common technique used in analyzing linear regression. In brief, it compares the difference between individual points in your data set and the predicted best fit line to measure the amount of error produced. ... ols() function requires two inputs the formula for producing the best fit line, and the dataset.

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.989
Model:	OLS	Adj. R-squared:	0.989
Method:	Least Squares	F-statistic:	9.513e+04
Date:	Sat, 23 Apr 2022	Prob (F-statistic):	0.00
Time:	10:35:03	Log-Likelihood:	-2.2943e+05
No. Observations:	17500	AIC:	4.589e+05
Df Residuals:	17482	BIC:	4.590e+05
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.952e+06	5.33e+05	14.906	0.000	6.91e+06	9e+06
Total_Experience	-5708.3712	361.518	-15.790	0.000	-6416.983	-4999.760
Total_Experience_in_field_applied	7627.9918	358.648	21.269	0.000	6925.006	8330.977
Department	-2.696e+04	1337.203	-20.165	0.000	-2.96e+04	-2.43e+04
Role	-9.225e+04	1487.514	-62.016	0.000	-9.52e+04	-8.93e+04
Designation	-3.453e+04	1362.882	-25.337	0.000	-3.72e+04	-3.19e+04
Highest_Education	9.198e+04	1289.197	71.349	0.000	8.95e+04	9.45e+04
Passing_Year_Of_Graduation	-3810.3385	265.021	-14.377	0.000	-4329.807	-3290.870
Passing_Year_Of_PG	-28.2217	1.717	-16.433	0.000	-31.588	-24.856
Passing_Year_Of_PHD	-15.4039	1.852	-8.318	0.000	-19.034	-11.774
Current_CTC	1.2507	0.002	551.869	0.000	1.246	1.255
Inhand_Offer	4.018e+04	2229.571	18.022	0.000	3.58e+04	4.46e+04
Last_Appraisal_Rating	6.95e+04	867.810	80.088	0.000	6.78e+04	7.12e+04
No_Of_Companies_worked	-1.087e+04	604.064	-18.001	0.000	-1.21e+04	-9689.946
Number_of_Publications	4492.3749	469.075	9.577	0.000	3572.942	5411.808
Certifications	612.4294	888.161	0.690	0.490	-1128.456	2353.314
International_degree_any	3.64e+04	3369.645	10.803	0.000	2.98e+04	4.3e+04
Percentage_Relevant_Exp_in_Field	-1275.8617	48.378	-26.373	0.000	-1370.687	-1181.036

Omnibus:	6129.924	Durbin-Watson:	1.989
Prob(Omnibus):	0.000	Jarque-Bera (JB):	58782.864
Skew:	1.407	Prob(JB):	0.00
Kurtosis:	11.526	Cond. No.	1.17e+09

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.17e+09. This might indicate that there are
strong multicollinearity or other numerical problems.

Tab : 53 Summary of Linear Regression Model - 1 (Train Data)

Insights :

- We get the intercept & coefficient values from the summary .
- Looking at p value of Certifications we conclude that their is no relationship between Certifications &Expected_CTC (dependent variable) , so we can drop it from sample , will do further analysis.

- R-sqd value for this model is 0.989 which is very good.
- Adj.- R-sqd value for this model is 0.989 which very good.
- The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample , here Durbin Watson value is 1.989 this shows there is no autocorrelation detected in the sample.
- Here, Kurtosis value found be 11.526 which indicates he dataset has heavier tails than a normal distribution (more in the tails).
- Here skew is 1.407 indicated data slightly right skewed
- Prob(Omnibus) – a test of the skewness and kurtosis of the residual (characteristic #2). We hope to see a value close to zero which would indicate normalcy. The Prob (Omnibus) performs a statistical test indicating the probability that the residuals are normally distributed. Here omnibus value is 0.000 indicates normalcy.
- Standard Errors assume that the covariance matrix of the errors is correctly specified.
- The condition number is large, 1.17e+09. This might indicate that there are strong multicollinearity or other numerical problems , we can do treatment of multicollinearity for better results.As we found in note -1 by the help of heat-map and feature selection method year of graduation , PG and PHD have correlation with each other that's why this problem arises we can drop any of the feature by taking advice of domain expert.

RMSE on Training data - 119545.42733621519

Prediction on Train Data -

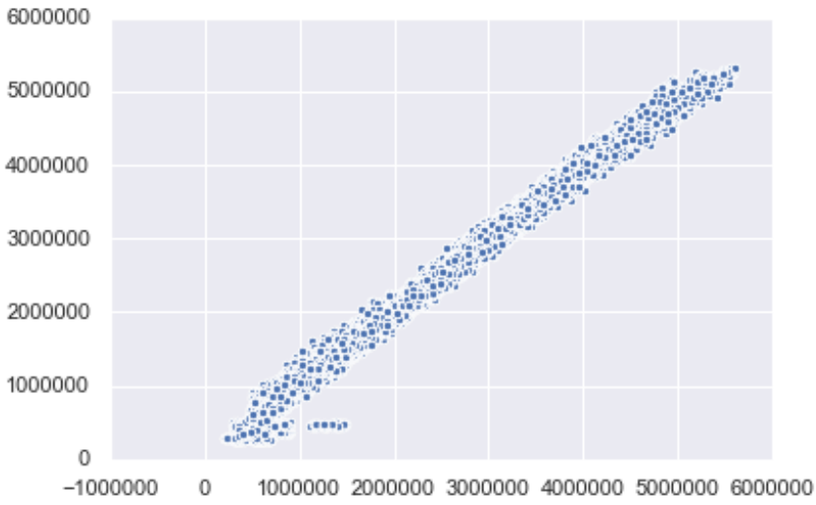


Fig: 63 Prediction on Train Data Model 1 (Scatter Plot Showing Distribution of Actual y & Predicted y)

Linear Regression Expression for Train Data.

(7951913.56) * Intercept + (-5708.37) * Total_Experience + (7627.99) * Total_Experience_in_field_applied + (-26964.33) * Department + (-92250.09) * Role + (-34531.53) * Designation + (91983.43) * Highest_Education + (-3810.34) * Passing_Year_Of_Graduation + (-28.22) * Passing_Year_Of_PG + (-15.4) * Passing_Year_Of_PHD + (1.25) * Current_CTC + (40181.77) * Inhand_Offer + (69501.38) * Last_Appraisal_Rating + (-10873.97) * No_Of_Companies_worked + (4492.37) * Number_of_Publications + (612.43) * Certifications + (36401.79) * International_degree_any + (-1275.86) * Percentage_Relevant_Exp_in_Field +

This is the expression of Model 1 for Test Data

expr= 'Expected_CTC ~ Total_Experience + Total_Experience_in_field_applied + Department + Role + Designation + Highest_Education + Passing_Year_Of_Graduation + Passing_Year_Of_PG + Passing_Year_Of_PHD + Current_CTC + Inhand_Offer + Last_Appraisal_Rating + No_Of_Companies_worked + Number_of_Publications + Certifications +International_degree_any + Percentage_Relevant_Exp_in_Field'

Explore the coefficients for each of the independent attributes of Test data.

Intercept	8.289336e+06
Total_Experience	-5.660149e+03
Total_Experience_in_field_applied	7.487810e+03
Department	-2.756411e+04
Role	-8.955167e+04
Designation	-3.214243e+04
Highest_Education	9.110346e+04
Passing_Year_Of_Graduation	-3.988796e+03
Passing_Year_Of_PG	-2.441003e+01
Passing_Year_Of_PHD	-1.792036e+01
Current_CTC	1.251031e+00
Inhand_Offer	5.073873e+04
Last_Appraisal_Rating	6.770867e+04
No_Of_Companies_worked	-9.086677e+03
Number_of_Publications	3.913095e+03
Certifications	-3.405709e+02
International_degree_any	3.087677e+04
Percentage_Relevant_Exp_in_Field	-1.232950e+03

The coefficients in this linear equation denote the magnitude of additive relation between the predictor and the response. In simpler words, keeping everything else fixed, a unit change in x1 will lead to change of β1 in the outcome, and so on.

Intercept for the model on test data -

The regression constant is also known as the intercept thus, regression models without predictors are also known as intercept only models. ... As such, we will begin with intercept only models for OLS regression and then move on to logistic regression models without predictors. Intercept for the model is **8.289336e+06**

OLS Regression Results of test data -

OLS is a common technique used in analyzing linear regression. In brief, it compares the difference between individual points in your data set and the predicted best fit line to measure the amount of error produced. ... ols() function requires two inputs the formula for producing the best fit line, and the dataset.

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.990
Model:	OLS	Adj. R-squared:	0.990
Method:	Least Squares	F-statistic:	4.287e+04
Date:	Sat, 23 Apr 2022	Prob (F-statistic):	0.00
Time:	10:35:03	Log-Likelihood:	-98232.
No. Observations:	7500	AIC:	1.965e+05
Df Residuals:	7482	BIC:	1.966e+05
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.289e+06	8.01e+05	10.351	0.000	6.72e+06	9.86e+06
Total_Experience	-5660.1488	528.843	-10.703	0.000	-6696.830	-4623.468
Total_Experience_in_field_applied	7487.8097	543.526	13.776	0.000	6422.346	8553.274
Department	-2.756e+04	2009.275	-13.718	0.000	-3.15e+04	-2.36e+04
Role	-8.955e+04	2242.384	-39.936	0.000	-9.39e+04	-8.52e+04
Designation	-3.214e+04	2067.160	-15.549	0.000	-3.62e+04	-2.81e+04
Highest_Education	9.11e+04	1940.747	46.942	0.000	8.73e+04	9.49e+04
Passing_Year_Of_Graduation	-3988.7957	397.919	-10.024	0.000	-4768.829	-3208.763
Passing_Year_Of_PG	-24.4100	2.572	-9.492	0.000	-29.451	-19.369
Passing_Year_Of_PHD	-17.9204	2.825	-6.344	0.000	-23.458	-12.383
Current_CTC	1.2510	0.003	374.354	0.000	1.244	1.258
Inhand_Offer	5.074e+04	3366.961	15.070	0.000	4.41e+04	5.73e+04
Last_Appraisal_Rating	6.771e+04	1299.604	52.099	0.000	6.52e+04	7.03e+04
No_Of_Companies_worked	-9086.6765	918.784	-9.890	0.000	-1.09e+04	-7285.601
Number_of_Publications	3913.0952	714.087	5.480	0.000	2513.283	5312.907
Certifications	-340.5709	1319.868	-0.258	0.796	-2927.884	2246.742
International_degree_any	3.088e+04	5111.393	6.041	0.000	2.09e+04	4.09e+04
Percentage_Relevant_Exp_in_Field	-1232.9497	73.716	-16.726	0.000	-1377.455	-1088.445

Omnibus:	2593.503	Durbin-Watson:	2.021
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24564.453
Skew:	1.382	Prob(JB):	0.00
Kurtosis:	11.424	Cond. No.	1.17e+09

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.17e+09. This might indicate that there are strong multicollinearity or other numerical problems.

Tab : 54 Summary of Linear Regression Model - 1 (Test Data)

Insights :

- We get the intercept & coefficient values from the summary .
- Looking at p value of Certifications we conclude that their is no relationship between Certifications &Expected_CTC (dependent variable) , so we can drop it from sample , will do further analysis.
- R-sqd value for this model is 0.990 which is very good.
- Adj.- R-sqd value for this model is 0.990 which very good.
- The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample , here Durbin Watson value is 2.0 this shows there is no autocorrelation detected in the sample.
- Here, Kurtosis value found be 11.424 which indicates he dataset has heavier tails than a normal distribution (more in the tails).
- Here skew is 1.382 indicated data slightly right skewed.
- Prob(Omnibus) – a test of the skewness and kurtosis of the residual (characteristic #2). We hope to see a value close to zero which would indicate normalcy. The Prob (Omnibus) performs a statistical test indicating the probability that the residuals are normally distributed. Here omnibus value is 0.000 indicates normalcy.
- Standard Errors assume that the covariance matrix of the errors is correctly specified.
- The condition number is large, 1.17e+09. This might indicate that there are strong multicollinearity or other numerical problems , we can do treatment of multicollinearity for better results.As we found in note -1 by the help of heat-map and feature selection method year of graduation , PG and PHD have correlation with each other that's why this problem arises we can drop any of the feature by taking advice of domain expert.

RMSE on Test data - 118024.20092797445

Prediction on Test Data -

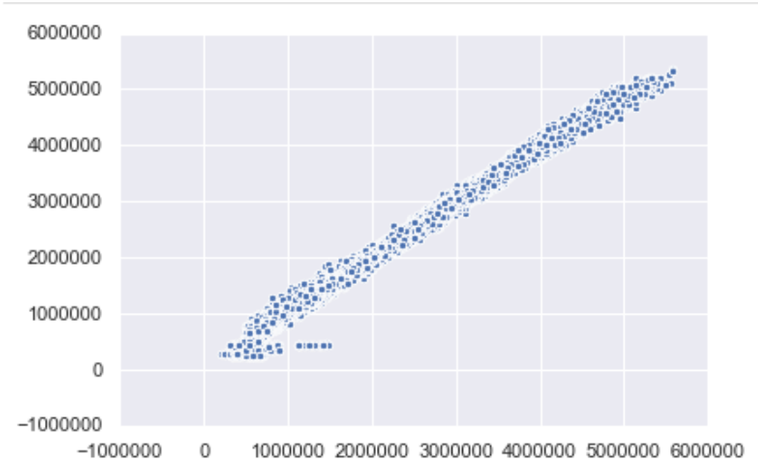


Fig: 64 Prediction on Test Data Model 1 (Scatter Plot Showing Distribution of Actual y & Predicted y)

Linear Regression Expression for Test Data.

(8289335.82) * Intercept + (-5660.15) * Total_Experience + (7487.81) * Total_Experience_in_field_applied + (-27564.11) * Department + (-89551.67) * Role + (-32142.43) * Designation + (91103.46) * Highest_Education + (-3988.8) * Passing_Year_Of_Graduation + (-24.41) * Passing_Year_Of_PG + (-17.92) * Passing_Year_Of_PHD + (1.25) * Current_CTC + (50738.73) * Inhand_Offer + (67708.67) * Last_Appraisal_Rating + (-9086.68) * No_Of_Companies_worked + (3913.1) * Number_of_Publications + (-340.57) * Certifications + (30876.77) * International_degree_any + (-1232.95) * Percentage_Relevant_Exp_in_Field +

Model 2- Linear Regression with Z-Score Scaling (By Sklearn library)

Explore the coefficients for each of the independent attributes.

- The coefficient for Total_Experience is -0.036910256337198334
- The coefficient for Total_Experience_in_field_applied is 0.03867264545859081
- The coefficient for Department is -0.017369049439972986
- The coefficient for Role is -0.06411092488498438
- The coefficient for Designation is -0.02157293210723535
- The coefficient for Highest_Education is 0.08895524323659658
- The coefficient for Passing_Year_Of_Graduation is -0.023838465335146983
- The coefficient for Passing_Year_Of_PG is -0.02257131933597273
- The coefficient for Passing_Year_Of_PHD is -0.013356747190299341
- The coefficient for Current_CTC is 0.992608926805067
- The coefficient for Inhand_Offer is 0.01597323953337604
- The coefficient for Last_Appraisal_Rating is 0.07079957993233385
- The coefficient for No_Of_Companies_worked is -0.015903120805298227
- The coefficient for Number_of_Publications is 0.01015971990337714
- The coefficient for Certifications is 0.0006337864931054203
- The coefficient for International_degree_any is 0.008647721923217262
- The coefficient for Percentage_Relevant_Exp_in_Field is -0.03740496119882783

The coefficients in this linear equation denote the magnitude of additive relation between the predictor and the response. In simpler words, keeping everything else fixed, a unit change in x1 will lead to change of β1 in the outcome, and so on.

Intercept for the model -

The regression constant is also known as the intercept thus, regression models without predictors are also known as intercept only models.As such, we will begin with intercept only models for OLS regression and then move on to logistic regression models without predictors. Intercept for the model is **-2.6537853445585968e-16.**

R square - is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. 100% indicates that the model explains all the variability of the response data around its mean.

R square on Training Data - 0.9893059671397622
R square on Testing Data - 0.9897874551097149

R-Squared value of 0.989 would indicate that 98.9% of the variance of the dependent variable being studied is explained by the variance of the independent variable.

RMSE - The root mean square error (RMSE) for a regression model is similar to the standard deviation (SD) for the ideal measurement model. The SD estimates the deviation from the sample mean x. The RMSE estimates the deviation of the actual y-values from the regression line.

Root mean square error, which is a metric that tells us the average distance between the predicted values from the model and the actual values in the dataset. The lower the RMSE, the better a given model is able to “fit” a dataset.

RMSE on Training data - 0.10341195704674494
RMSE on Testing data - 0.10081062150738966

Model 2 - Linear Regression using Stats-Model ols with Z-Score Scaling

This is the expression of Model 2 for Training Data

expr_2= 'Expected_CTC ~ Total_Experience + Total_Experience_in_field_applied + Department + Role + Designation + Highest_Education + Passing_Year_Of_Graduation + Passing_Year_Of_PG + Passing_Year_Of_PHD + Current_CTC + Inhand_Offer + Last_Appraisal_Rating + No_Of_Companies_worked + Number_of_Publications + Certifications + International_degree_any + Percentage_Relevant_Exp_in_Field'

Explore the coefficients for each of the independent attributes of train data.

- The coefficient for Total_Experience is -0.036910256337198334
- The coefficient for Total_Experience_in_field_applied is 0.03867264545859081
- The coefficient for Department is -0.017369049439972986
- The coefficient for Role is -0.06411092488498438
- The coefficient for Designation is -0.02157293210723535
- The coefficient for Highest_Education is 0.08895524323659658
- The coefficient for Passing_Year_Of_Graduation is -0.023838465335146983
- The coefficient for Passing_Year_Of_PG is -0.02257131933597273
- The coefficient for Passing_Year_Of_PHD is -0.013356747190299341
- The coefficient for Current_CTC is 0.992608926805067
- The coefficient for Inhand_Offer is 0.01597323953337604
- The coefficient for Last_Appraisal_Rating is 0.07079957993233385
- The coefficient for No_Of_Companies_worked is -0.015903120805298227
- The coefficient for Number_of_Publications is 0.01015971990337714
- The coefficient for Certifications is 0.0006337864931054203
- The coefficient for International_degree_any is 0.008647721923217262
- The coefficient for Percentage_Relevant_Exp_in_Field is -0.03740496119882783

The coefficients in this linear equation denote the magnitude of additive relation between the predictor and the response. In simpler words, keeping everything else fixed, a unit change in x1 will lead to change of β1 in the outcome, and so on.

Intercept for the model on train data -

The regression constant is also known as the intercept thus, regression models without predictors are also known as intercept only models. ... As such, we will begin with intercept only models for OLS regression and then move on to logistic regression models without predictors. Intercept for the model is **-1.734723e-18.**

OLS Regression Results of train data -

OLS is a common technique used in analyzing linear regression. In brief, it compares the difference between individual points in your data set and the predicted best fit line to measure the amount of error produced. ... ols() function requires two inputs the formula for producing the best fit line, and the dataset.

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.989
Model:	OLS	Adj. R-squared:	0.989
Method:	Least Squares	F-statistic:	9.513e+04
Date:	Sat, 23 Apr 2022	Prob (F-statistic):	0.00
Time:	10:35:04	Log-Likelihood:	14877.
No. Observations:	17500	AIC:	-2.972e+04
Df Residuals:	17482	BIC:	-2.958e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.735e-18	0.001	-2.22e-15	1.000	-0.002	0.002
Total_Experience	-0.0369	0.002	-15.790	0.000	-0.041	-0.032
Total_Experience_in_field_applied	0.0387	0.002	21.269	0.000	0.035	0.042
Department	-0.0174	0.001	-20.165	0.000	-0.019	-0.016
Role	-0.0641	0.001	-62.016	0.000	-0.066	-0.062
Designation	-0.0216	0.001	-25.337	0.000	-0.023	-0.020
Highest_Education	0.0890	0.001	71.349	0.000	0.087	0.091
Passing_Year_Of_Graduation	-0.0238	0.002	-14.377	0.000	-0.027	-0.021
Passing_Year_Of_PG	-0.0226	0.001	-16.433	0.000	-0.025	-0.020
Passing_Year_Of_PHD	-0.0134	0.002	-8.318	0.000	-0.017	-0.010
Current_CTC	0.9926	0.002	551.869	0.000	0.989	0.996
Inhand_Offer	0.0160	0.001	18.022	0.000	0.014	0.018
Last_Appraisal_Rating	0.0708	0.001	80.088	0.000	0.069	0.073
No_Of_Companies_worked	-0.0159	0.001	-18.001	0.000	-0.018	-0.014
Number_of_Publications	0.0102	0.001	9.577	0.000	0.008	0.012
Certifications	0.0006	0.001	0.690	0.490	-0.001	0.002
International_degree_any	0.0086	0.001	10.803	0.000	0.007	0.010
Percentage_Relevant_Exp_in_Field	-0.0374	0.001	-26.373	0.000	-0.040	-0.035

Omnibus:	6129.924	Durbin-Watson:	1.989
Prob(Omnibus):	0.000	Jarque-Bera (JB):	58782.864
Skew:	1.407	Prob(JB):	0.00
Kurtosis:	11.526	Cond. No.	7.65

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tab : 55 Summary of Linear Regression Model - 2 (Train Data)

Insights :

- We get the intercept & coefficient values from the summary .
- Looking at p value of Certifications we conclude that their is no relationship between Certifications &Expected_CTC (dependent variable) , so we can drop it from sample , will do further analysis.
- R-sqd value for this model is 0.989 which is very good.
- Adj.- R-sqd value for this model is 0.989 which very good.
- The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample , here Durbin Watson value is 1.989 this shows there is no autocorrelation detected in the sample.
- Here, Kurtosis value found be 11.526 which indicates he dataset has heavier tails than a normal distribution (more in the tails).
- Here skew is 1.407 indicated data slightly right skewed
- Prob(Omnibus) – a test of the skewness and kurtosis of the residual (characteristic #2). We hope to see a value close to zero which would indicate normalcy. The Prob (Omnibus) performs a statistical test indicating the probability that the residuals are normally distributed. Here omnibus value is 0.000 indicates normalcy.
- Standard Errors assume that the covariance matrix of the errors is correctly specified.

RMSE on Training data - 0.10341195704674508

Prediction on Train Data -

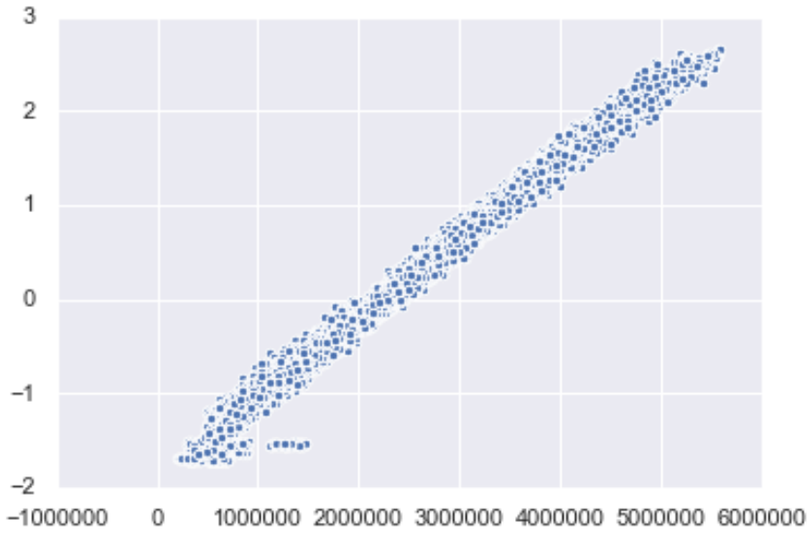


Fig: 65 Prediction on Train Data Model 2 (Scatter Plot Showing Distribution of Actual y & Predicted y)

Linear Regression Expression for Train Data.

(-0.0) * Intercept + (-0.04) * Total_Experience + (0.04) * Total_Experience_in_field_applied + (-0.02) * Department + (-0.06) * Role + (-0.02) * Designation + (0.09) * Highest_Education + (-0.02) * Passing_Year_Of_Graduation + (-0.02) * Passing_Year_Of_PG + (-0.01) * Passing_Year_Of_PHD + (0.99) * Current_CTC + (0.02) * Inhand_Offer + (0.07) * Last_Appraisal_Rating + (-0.02) * No_Of_Companies_worked + (0.01) * Number_of_Publications + (0.0) * Certifications + (0.01) * International_degree_any + (-0.04) * Percentage_Relevant_Exp_in_Field +

This is the expression of Model 2 for Test Data

expr_2= 'Expected_CTC ~ Total_Experience + Total_Experience_in_field_applied + Department + Role + Designation + Highest_Education + Passing_Year_Of_Graduation + Passing_Year_Of_PG + Passing_Year_Of_PHD + Current_CTC + Inhand_Offer + Last_Appraisal_Rating + No_Of_Companies_worked + Number_of_Publications + Certifications + International_degree_any + Percentage_Relevant_Exp_in_Field'

Explore the coefficients for each of the independent attributes of test data.

Intercept	1.942890e-16
Total_Experience	-3.608045e-02
Total_Experience_in_field_applied	3.659121e-02
Department	-1.751725e-02
Role	-6.067375e-02
Designation	-1.974041e-02
Highest_Education	8.666216e-02
Passing_Year_Of_Graduation	-2.445569e-02
Passing_Year_Of_PG	-1.934000e-02
Passing_Year_Of_PHD	-1.534801e-02
Current_CTC	9.900043e-01
Inhand_Offer	1.993500e-02
Last_Appraisal_Rating	6.822196e-02
No_Of_Companies_worked	-1.311259e-02
Number_of_Publications	8.650356e-03
Certifications	-3.510010e-04
International_degree_any	7.182087e-03
Percentage_Relevant_Exp_in_Field	-3.519465e-02

The coefficients in this linear equation denote the magnitude of additive relation between the predictor and the response. In simpler words, keeping everything else fixed, a unit change in x1 will lead to change of β1 in the outcome, and so on.

Intercept for the model on test data -

The regression constant is also known as the intercept thus, regression models without predictors are also known as intercept only models. ... As such, we will begin with intercept only models for OLS regression and then move on to logistic regression models without predictors. Intercept for the model is **1.942890e-16.**

OLS Regression Results of test data -

OLS is a common technique used in analyzing linear regression. In brief, it compares the difference between individual points in your data set and the predicted best fit line to measure the amount of error produced. ... ols() function requires two inputs the formula for producing the best fit line, and the dataset.

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.990
Model:	OLS	Adj. R-squared:	0.990
Method:	Least Squares	F-statistic:	4.287e+04
Date:	Sat, 23 Apr 2022	Prob (F-statistic):	0.00
Time:	10:35:05	Log-Likelihood:	6566.8
No. Observations:	7500	AIC:	-1.310e+04
Df Residuals:	7482	BIC:	-1.297e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.943e-16	0.001	1.67e-13	1.000	-0.002	0.002
Total_Experience	-0.0361	0.003	-10.703	0.000	-0.043	-0.029
Total_Experience_in_field_applied	0.0366	0.003	13.776	0.000	0.031	0.042
Department	-0.0175	0.001	-13.718	0.000	-0.020	-0.015
Role	-0.0607	0.002	-39.936	0.000	-0.064	-0.058
Designation	-0.0197	0.001	-15.549	0.000	-0.022	-0.017
Highest_Education	0.0867	0.002	46.942	0.000	0.083	0.090
Passing_Year_Of_Graduation	-0.0245	0.002	-10.024	0.000	-0.029	-0.020
Passing_Year_Of_PG	-0.0193	0.002	-9.492	0.000	-0.023	-0.015
Passing_Year_Of_PHD	-0.0153	0.002	-6.344	0.000	-0.020	-0.011
Current_CTC	0.9900	0.003	374.354	0.000	0.985	0.995
Inhand_Offer	0.0199	0.001	15.070	0.000	0.017	0.023
Last_Appraisal_Rating	0.0682	0.001	52.099	0.000	0.066	0.071
No_Of_Companies_worked	-0.0131	0.001	-9.890	0.000	-0.016	-0.011
Number_of_Publications	0.0087	0.002	5.480	0.000	0.006	0.012
Certifications	-0.0004	0.001	-0.258	0.796	-0.003	0.002
International_degree_any	0.0072	0.001	6.041	0.000	0.005	0.010
Percentage_Relevant_Exp_in_Field	-0.0352	0.002	-16.726	0.000	-0.039	-0.031

Omnibus:	2593.503	Durbin-Watson:	2.021
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24564.453
Skew:	1.382	Prob(JB):	0.00
Kurtosis:	11.424	Cond. No.	7.43

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Tab : 56 Summary of Linear Regression Model - 2 (Test Data)

Insights :

- We get the intercept & coefficient values from the summary .
- Looking at p value of Certifications we conclude that their is no relationship between Certifications &Expected_CTC (dependent variable) , so we can drop it from sample , will do further analysis.
- R-sqd value for this model is 0.990 which is very good.
- Adj.- R-sqd value for this model is 0.990 which very good.
- The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample , here Durbin Watson value is 2.0 this shows there is no autocorrelation detected in the sample.
- Here, Kurtosis value found be 11.424 which indicates he dataset has heavier tails than a normal distribution (more in the tails).
- Here skew is 1.382 indicated data slightly right skewed.
- Prob(Omnibus) – a test of the skewness and kurtosis of the residual (characteristic #2). We hope to see a value close to zero which would indicate normalcy. The Prob (Omnibus) performs a statistical test indicating the probability that the residuals are normally distributed. Here omnibus value is 0.000 indicates normalcy.
- Standard Errors assume that the covariance matrix of the errors is correctly specified.

RMSE on Test data - 0.10081062150738977

Prediction on Test Data -

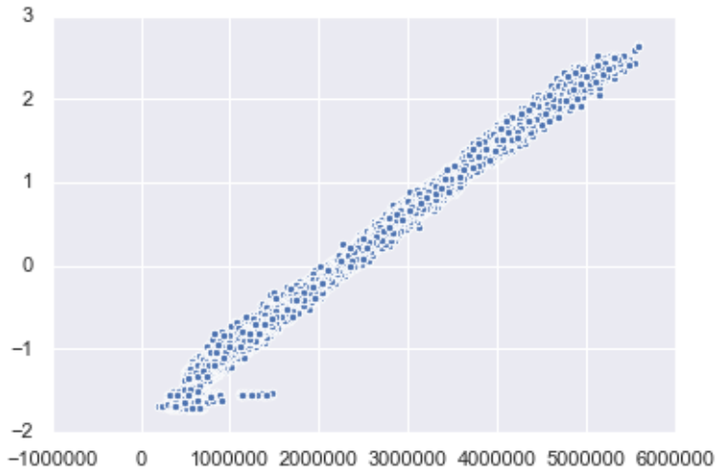


Fig: 66 Prediction on Test Data Model 2 (Scatter Plot Showing Distribution of Actual y & Predicted y)

Linear Regression Expression for Test Data.

(0.0) * Intercept + (-0.04) * Total_Experience + (0.04) * Total_Experience_in_field_applied + (-0.02) * Department + (-0.06) * Role + (-0.02) * Designation + (0.09) * Highest_Education + (-0.02) * Passing_Year_Of_Graduation + (-0.02) * Passing_Year_Of_PG + (-0.02) * Passing_Year_Of_PHD + (0.99) * Current_CTC + (0.02) * Inhand_Offer + (0.07) * Last_Appraisal_Rating + (-0.01) * No_Of_Companies_worked + (0.01) * Number_of_Publications + (-0.0) * Certifications + (0.01) * International_degree_any + (-0.04) * Percentage_Relevant_Exp_in_Field +

Conclusion :

Factors of Comparison	Linear Regression Model-1 (Train Data)	Linear Regression Model-1 (Test Data)	Linear Regression Model-2 (Z-Score Scaled)(Train Data)	Linear Regression Model-2 (Z-Score Scaled)(Test Data)
R-Square	0.989	0.989	0.989	0.990
Adj.R-Square	0.989	0.990	0.989	0.990
RMSE	119545.427336215	118024.200927974	0.103411957046745	0.10081062150739

Tab : 57 Comparison Table of Linear Regression Model 1 & 2 on Train and Test Data

As we saw from the above results that Linear Regression Model 1 and Model 2 don't have any issues of Overfitting and Underfitting plus R-square and Adj.R-Square values are also good for both models (nearly same).

Model 3 - XG Boost Regressor

XGBoost is a powerful approach for building supervised regression models. XGBoost is an efficient implementation of gradient boosting that can be used for regression predictive modeling.The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners.

The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e how far the model results are from the real values. The most common loss functions in XGBoost for regression problems is reg:linear, and that for binary classification is reg:logistics.

Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and XGBoost is one of the ensemble learning methods. XGBoost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancels out and better one sums up to form final good predictions.

Results :

Factors of Comparison	XG-Boost Regressor Model -3 (Train Data)	XG-Boost Regressor Model -3 (Test Data)
R-Square	0.99	0.99
Adj. R-Square	0.99	0.99
RMSE	36162.740820	35673.038967

Tab : 58 Result XG-BoostRegressor Model on Train and Test Data

Conclusion :

As we saw from the above results that XG-Boost Regressor Model-3 don't have any issues of Overfitting and Underfitting plus R-square and Adj.R-Square values are also good for both train and test. Even, here we get less RMSE on train and test as compared above Linear Regression Model.

Model - 4 Building 3 models using Decision Tree, Random Forest and ANN Regressor

Decision Tree Algorithm

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.

With a particular data point, it is run completely through the entirely tree by answering True/False questions till it reaches the leaf node. The final prediction is the average of the value of the dependent variable in that particular leaf node. Through multiple iterations, the Tree is able to predict a proper value for the data point.

Random Forest Regressor

A random forest regressor. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

The random forest algorithm follows a two-step process:

Builds n decision tree regressors (estimators). The number of estimators n defaults to 100 in Scikit Learn (the machine learning Python library), where it is called n_estimators. The trees are built following the specified hyperparameters (e.g. minimum number of samples at the leaf nodes, maximum depth that a tree can grow, etc.).

Average prediction across estimators. Each decision tree regression predicts a number as an output for a given input. Random forest regression takes the average of those predictions as its ‘final’ output.

ANN/MLP Regressor

Regression ANNs predict an output variable as a function of the inputs. The input features (independent variables) can be categorical or numeric types, however, for Regression ANNs, we require a numeric dependent variable.As we have to predict a continuous variable Expected_CTC ,that's why we are going to use this algorithm.

MLP Regressor trains iteratively since at each time step the partial derivatives of the loss function with respect to the model parameters are computed to update the parameters. It can also have a regularization term added to the loss function that shrinks model parameters to prevent overfitting.

Results Decision Tree, Random Forestand ANN Regressor

Models	Train RMSE	Test RMSE	Training Score	Test Score
Decision Tree Regressor	23964.035376	35992.407229	0.999570	0.999055
Random Forest Regressor	24745.396503	32929.832228	0.999542	0.999209
ANN Regressor	39581.455885	40714.261494	0.998820	0.998791

Tab : 59 Result of ANN, Decision Tree, Random Forest Regressor Model on Train and Test Data

Conclusion :

As we saw from the above results that Decision Tree , Random Forest and ANN Regressor don't have any issues of Overfitting and Underfitting plus model scores are also good for both train and test. Even, here we get good RMSE on train and test data.

Note - (Effort to improve model performance)

From the above results we clearly infer that no model have any issues of Overfitting and Underfitting , we have good accuracy scores and also have good R-square , Adj-R square and RMSE on train and test set. But we are doing hyper-parameter tuning on Decision Tree , Random Forest and ANN Regressor and check their accuracy ,and RMSE on train and test set whether there is any impact or not.

Model 5 - Hyper-parameter Tuning for Descision Tree / Random Forest / ANN Regressor

Note : For code reference please check the code file.

We are using Grid search to build a model for every combination of hyper-parameters specified and evaluates each model. A more efficient technique for hyper-parameter tuning is the Randomized search — where random combinations of the hyper-parameters are used to find the best solution.

As per the industries standards we are taking various hyper parameters to build our Decision Tree , Random Forest and ANN Regressor Model ,hyperparametrs are listed below.

Grid Search Results on Decision Tree

{'max_depth': 20, 'min_samples_leaf': 3, 'min_samples_split': 15}

Note : From Grid Search CV we get our best_params_ , we uses these parameters to build our Decision Tree Regressor.

Grid Search Results on Random Forest Regressor

{'max_depth': 10, 'max_features': 6, 'min_samples_leaf': 3, 'min_samples_split': 30, 'n_estimators': 500}

Note : From Grid Search CV we get our best_params_ , we uses these parameters to build our Random Forest Regressor.

Grid Search Results on ANN Regressor

{'activation': 'relu', 'hidden_layer_sizes': 500, 'solver': 'sgd'}

Note : From Grid Search CV we get our best_params_ , we uses these parameters to build our ANN Regressor Model.

Results Tuned Decision Tree, Tuned Random Forest and Tuned ANN Regressor

Models	Train RMSE	Test RMSE	Training Score	Test Score
Tuned-Decision Tree Regressor	23764.035376	34992.407229	0.98957	0.989055
Tuned-Random Forest Regressor	23745.396503	31929.832228	0.989542	0.989209
Tuned- ANN Regressor	38581.455885	39714.261494	0.98882	0.988791

Tab : 60 Result of Tuned ANN, Tuned Decision Tree, Tuned Random Forest Regressor Model on Train and Test Data

Conclusion :

As we saw from the above results that Tuned Decision Tree , Tuned Random Forest and Tuned ANN Regressor don't have any issues of Overfitting and Under-fitting , plus model scores are also good for both train and test and almost similar with above results. Even, here we get almost same RMSE on train and test data too as compared to the above results.

Model Validation and Comaparsion of all Model

For model validation we did the train -test split , trained the model on train data and validate it by using the test set. For model performance we have to check Accuracy metrics and R-Sqaure and Adj. R-Square and RMSE values all the models.As we build various Regression models like Linear Regression Models like Linear Regression Model Linear Regression (With Z-Score Scaling) , XG-Boost Regressor Model , Decision Tree Regressor , Random Forest Regressor , ANN Regressor Model and also did the hyper-parameter tuning of the Decision Tree ,Random Forest and ANN and calculate the R-Square and RMSE on train and test data for all models.Now we compare the all the models which we have been build durning the model building exercise and choose our generalised model for deployment i.e. Model which have highest R-Square / Model Score and Least RMSE on train and test data.

Models	Training Score /R-Square (Train)	Test Score / R-Square (Test)	Adj.R-Square(Train)	Adj.R-Square (Test)	Training RMSE	Test RMSE
Linear Regression Model -1	0.989	0.989	0.989	0.990	119545.427336215	118024.200927974
Linear Regression Model -2 (Z-Score Scaling)	0.989	0.990	0.989	0.990	0.103411957046745	0.100811062150739
XG-Boost Regressor	0.99	0.99	0.99	0.99	36162.740820	35673.038967
Decision Tree Regressor	0.999570	0.999055	----------	----------	23964.035376	35992.407229
Random Forest Regressor	0.999542	0.999209	----------	----------	24745.396503	32929.832228
ANN Regressor	0.998820	0.998791	----------	----------	39581.455885	40714.261494
Tuned-Decision Tree Regressor	0.98957	0.989055	----------	----------	23764.035375	34992.407229
Tuned-Random Forest Regressor	0.989542	0.989209	----------	----------	23745.396503	31929.832228
Tuned-ANN Regressor	0.98882	0.988791	----------	----------	38581.261494	39714.261494

Tab : 61 Comparison of All Models on Train and Test Data

Conclusion :

As we saw from the above results that all regression models don't have any issues of Overfitting and Underfitting plus model scores are also good and almost similar for all models on train and test. On comparing the Model Scores from the Linear Regression models and other regression models ,Tuned Random Forest Regressor seems to be an optimum model as we get lowest RMSE on Tuned Random Forest Regressor for train and test and have very good model score of 0.99 on train and test.

We can also use XG_Boost Regressor and it also have very good model score and least difference in between RMSE of train and test set , as it is also powerful technique now-a-days.

But as per the data given associated with the problem we finalise that Tuned Random Forest Regressor Model will be the final generalised model for deployment.As It can perform both regression and classification tasks. A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

The Top Attributes reasonable & influencing the Expected_CTC are : **Total_Experience_in_field_applied, Highest_Education ,Inhand_offer , Last_Appraisal_Rating , Role, Designation , Current_CTC, International_degree_any and No_Of_Companies_worked.**

Final Insights from EDA and Visualization

- Most of applicants are of Marketing Department (2379) ,Analytics/BI (2096) , Health-care (2062) and others (2041).
- Less applicants belongs to IT-Software Department(1078).
- Only 25 applicants worked as Lab executive (Role).
- Majority of applicants worked in Training Industry.
- There is not too much variations in the Organization columns equivalent number of applicants worked in 16 different organization.

- Majority of applicants worked as HR i.e.(1648).
- Only 52 applicants worked as Scientist.
- 6180 applicants are Under Graduate.
- 6209 applicants are Graduate.
- 6326 applicants are Post Graduate.
- 6285 applicants are Doctorate.
- Highest number of applicants did their graduation specialization in chemistry.
- Most of applicants did their graduation from Bhubaneswar University (1510) and Delhi University(1492).
- Most of applicants did their post-graduation specialization in mathematics (1800) and chemistry (1796).
- Most of applicants did their post-graduation from Bhubaneswar University (1510) and Delhi University(1492).
- Most of applicants did their PhD specialization in others (1545) and chemistry (1458).
- Most of applicants did their PhD from Kolkata University (1069) and Delhi University(1064).
- Most of applicant's current location is Bangalore i.e.(1742) but preferred location is Kanpur i.e. (1720).In current location banglore is on top but for preferred location bangalore is at last number.
- 17418 applicants don't have In-hand job offer while 7582 applicants have In-hand job offer.
- 4191 applicants Last_Appraisal_Rating is Key_Performer while 4671 applicants Last_Appraisal_Rating is A ,5501 applicants Last_Appraisal_Rating is B. , 4812 applicants Last_Appraisal_Rating is C. , 4917 applicants Last_Appraisal_Rating is D.

- Total_Experience and the Expected_CTC is showing a strong relationship,with increase in Total_Experience(Independent Variable),Expected_CTC (Target Variable)is also increases.
- Total_Experience_in_field_applied and the Expected_CTC is showing a positive relationship,with increase in Total_Experience_in_field_applied(Independent Variable),Expected_CTC (Target Variable)is slightly increases.
- Passing_Year_Of_Graduation and the Expected_CTC is showing a negative relationship, as the Passing_Year_Of_Graduation increases the Expected_CTC goes on decreases.
- Passing_Year_Of_PHD and the Expected_CTC is showing a negative relationship, as the Passing_Year_Of_PHD increases the Expected_CTC goes on decreases.
- Current_CTC and the Expected_CTC is showing a positive relationship, as the Current_CTC increases the Expected_CTC goes on increases.
- No_Of_Companies_worked and Expected_CTC.As No_Of_Companies_worked increase there is also some increase in Expected_CTC.
- Applicants for Top Management have higher median value than others for Expected_CTC.
- Median values of CEO and Research Scientists for Expected_CTC are quite high as compared to others but distribution is wider for Research Scientists.
- Median values of Research Scientists for Expected_CTC are quite high as compared to others.
- Marketing Manger , Manager ,Product Manager and HR almost have equivalent median values for Expected_CTC.
- Similarly Data Analyst , Assistant Manger , Others , Web Designers and Research Analyst have equivalent median values for Expected_CTC.
- Box-plot of Doctorate have higher median values for Expected_CTC as compared to others.
- Under Grad Box-plot have lowest median values for Expected_CTC.
- We infer that Expected_CTC for recently graduated applicants is least as compared to others.
- Expected_CTC w.r.t Passing_Year_Of_PG , early 1990s applicants have high median for Expected_CTC , then in 20s there is some fall which keeps on increasing by each year passed , this variation may be caused as some of them unable to complete their PG in specific 2 year span or unable to complete their PG by any reasons.
- Expected_CTC for recently PhD passed applicants is less than applicants who completed PhD in early 1990s and 2000s.
- Median values for Key_Performers are higher than others.

Final Insights from Models

(7951913.56) * Intercept + (-5708.37) * Total_Experience + (7627.99) * Total_Experience_in_field_applied + (-26964.33) * Department + (-92250.09) * Role + (-34531.53) * Designation + (91983.43) * Highest_Education + (-3810.34) * Passing_Year_Of_Graduation + (-28.22) * Passing_Year_Of_PG + (-15.4) * Passing_Year_Of_PHD + (1.25) * Current_CTC + (40181.77) * Inhand_Offer + (69501.38) * Last_Appraisal_Rating + (-10873.97) * No_Of_Companies_worked + (4492.37) * Number_of_Publications + (612.43) * Certifications + (36401.79) * International_degree_any + (-1275.86) * Percentage_Relevant_Exp_in_Field +

- When Total_Experience_in_field_applied increases by 1 unit, Expected_CTC increases by 7627.99 units, keeping all other predictors constant.
- When Total_Experience increases by 1 unit, Expected_CTC decreases by -5708.37 units, keeping all other predictors constant.
- When Highest_Education increases by 1 unit, Expected_CTC increases by 91983.43) units, keeping all other predictors constant.
- The Top Attributes reasonable & influencing the Expected_CTC are : Total_Experience_in_field_applied, Highest_Education ,Inhand_offer , Last_Appraisal_Rating , Role, Designation , Current_CTC, International_degree_any and No_Of_Companies_worked.

Final Recommendations

- Based on our analysis we found that most variables like Organization , Graduation_Specialization , University_Grad , PG_Specialization , University_PG , PHD_Specialization , University_PHD , Current_location and Preferred_location of these variables is not showing any variation with the target variable (Expected_CTC) and there is no specific relation between them and target so instead of these variables we suggest company to take different variables from applicants like **Employment_Gap , Marital_Status , No_dependent_in_family and Interview_Test_Scores . Secondly** , we saw in our data that there are Passing_Year_of_UG , Passing_Year_of_PG , Passing_Year_of_PHD is asked by applicants instead of asking for all years we can only ask for passing_year w.r.t to highest education applicant can have.

- As we saw in our analysis that applicants who have in_hand_offer for job have higher expection of ctc , but applicants who didn't have any offer in hand have lesser expected_ctc. So company can focus such applicants if they are fit for company , then company should hire them immediately , it help company in cost costing as we get good applicants at lower expected_ctc.

- Recently graduates applicants asking for lower ctc if they best fit the positions then company should hire recently grads .Comapny will get good employees at lower salary.

- Company should check for the applicants whether they have completed their degrees in standard duration or not. Applicants which have backlog or incomplete degree will be given lower ctc.

- As we saw in our analysis fresher applicants have zero total_experience and zero current_ctc. As these being the important predictors of target . So we need to build a separate model for such applicants.

- Applicants with PhD qualifications are the most expensive applicants in terms of CTC , so hire PhD holders only when company actually need them.

- Applicants with higher number of companies worked have higher Expected_CTC.Business should look for those applicants who worked for less number of comapnies but have experienced and perform well in the company.

- Most of applicants preferred location is Kanpur but current location is Bangalore. As we know Bangalore is tier1 city which has more expensive living cost than Kanpur , but in our analysis we infer that current_location and preferred_location both features didn't play any vital role even we drop these variables from models too , so we suggest company shouldn't give high or low ctc to applicants on basis of location.

- Company should focus on trends of market salary for different industries , roles and designation so every applicant will get an unbiased salary which he/she truly deserves.

- Company should make new HR strategies that satisfy the demands of the applicants also it will help company to get their employee in their desired budget and also help company to reduce their attrition rate.

Note : For more deatils please check the code .

Thank You !