CAPSTONE PROJECT

Salary Prediction

By: Asritha Malisetty

Mentor: Abhay Poddar

Contents:

1.INTRODUCTION

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.

Need of the Study/Project:

The objective of this exercise is to build a model, using historical data that will determine an employee's salary to be offered, such that manual judgments on selection are minimized. It is intended to have a robust approach and eliminate any discrimination in salary among similar employee profiles

Understanding business/social opportunity:

Predicting the expected CTC will help the Delta Ltd to hire its employees and offer the salary without any discrimination among similar employee profiles.

2.DATA REPORT

Visual inspection of data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 29 columns):
 #   Column                             Non-Null Count  Dtype  
---  ------                             --------------  -----  
 0   IDX                                25000 non-null  int64  
 1   Applicant_ID                       25000 non-null  int64  
 2   Total_Experience                   25000 non-null  int64  
 3   Total_Experience_in_field_applied  25000 non-null  int64  
 4   Department                         22222 non-null  object 
 5   Role                               24037 non-null  object 
 6   Industry                           24092 non-null  object 
 7   Organization                       24092 non-null  object 
 8   Designation                        21871 non-null  object 
 9   Education                          25000 non-null  object 
 10  Graduation_Specialization          18820 non-null  object 
 11  University_Grad                    18820 non-null  object 
 12  Passing_Year_Of_Graduation         18820 non-null  float64
 13  PG_Specialization                  17308 non-null  object 
 14  University_PG                      17308 non-null  object 
 15  Passing_Year_Of_PG                 17308 non-null  float64
 16  PHD_Specialization                 13119 non-null  object 
 17  University_PHD                     13119 non-null  object 
 18  Passing_Year_Of_PHD                13119 non-null  float64
 19  Curent_Location                    25000 non-null  object 
 20  Preferred_location                 25000 non-null  object 
 21  Current_CTC                        25000 non-null  int64  
 22  Inhand_Offer                       25000 non-null  object 
 23  Last_Appraisal_Rating              24092 non-null  object 
 24  No_Of_Companies_worked             25000 non-null  int64  
 25  Number_of_Publications             25000 non-null  int64  
 26  Certifications                     25000 non-null  int64  
 27  International_degree_any           25000 non-null  int64  
 28  Expected_CTC                       25000 non-null  int64  
dtypes: float64(3), int64(10), object(16)
```

Observations:

- From the table above we can see that the data contains 25000 rows and 29 attributes.
- There are 3 float variables, 10 int variables and 16 object variables.
- **IDX , Applicant_ID** are not required can be dropped.
- **Organization** has Miscllaneous data so have to drop this feature.
- There are missing values in **Department,Role, Industry, Designation, Graduation_Specialization, University_Grad, Passing_Year_Of_Graduation, PG_Specialization, University_PG, Passing_Year_Of_PG, PHD_Specialization, University_PHD, Passing_Year_Of_PHD, Last_Appraisal_Rating**

Understanding of attributes:

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.)

3. **EDA,Data Cleaning and Business Implication**

Removal of unwanted variables:

Following are the variables removed from the dataset

VARIABLES	REASON
IDX	Unique id which is not useful for predicting
Applicant_ID	Unique id which is not useful for predicting
Organization	Have miscellaneous data
Designation	There are 2 variable w.r.to job title Role and Designation from which designation tells about the generalized level of the employee whereas Role tells about the specific function or responsibilities assigned to an individual which helps in predicting ExpectedCTC.Hence,Droping Designation may reduce noise.
Graduation_Specialization	There is another variable Education which explains about the highest Qualification of Applicant and Graduation_Specialization does impact expectedCTC instead Experience and Experience in Current Industry will help in Predicting ExpectedCTC

University_Grad	This variable is not Required as Education explains about the Qualification of Applicant
Passing_Year_Of_Graduation	This variable is not Required as Education explains about the Qualification of Applicant
PG Specialization	This variable is not Required as Education explains about the Qualification of Applicant
University_PG	This variable is not Required as Education explains about the Qualification of Applicant
Passing_Year_of_PG	This variable is not Required as Education explains about the Qualification of Applicant
PHD_Specialization	This variable is not Required as Education explains about the Qualification of Applicant
University_PHD	This variable is not Required as Education explains about the Qualification of Applicant
Passing_Year_of_PHD	This variable is not Required as Education explains about the Qualification of Applicant
Current_Location	Current Location doesn't matter as the Preffered location is the one considered for the job.

Missing Value treatment

There are missing values in **Organization , Department,Role, Industry, Designation, Graduation_Specialization, University_Grad, Passing_Year_Of_Graduation, PG_Specialization, University_PG,**

Passing_Year_Of_PG, PHD_Specialization, University_PHD, Passing_Year_Of_PHD, Last_Appraisal_Rating.

```
Total_Experience                           0
Total_Experience_in_field_applied          0
Department                              2778
Role                                     963
Industry                                 908
Designation                             3129
Education                                  0
Graduation_Specialization               6180
University_Grad                         6180
Passing_Year_Of_Graduation              6180
PG_Specialization                       7692
University_PG                           7692
Passing_Year_Of_PG                      7692
PHD_Specialization                     11881
University_PHD                         11881
Passing_Year_Of_PHD                    11881
Curent_Location                            0
Preferred_location                         0
Current_CTC                                0
Inhand_Offer                               0
Last_Appraisal_Rating                    908
No_Of_Companies_worked                     0
Number_of_Publications                     0
Certifications                             0
International_degree_any                   0
Expected_CTC                               0
```

Variables with missing Values	Approach of Treating
Department	• Applicants with 0 experience will not have the department.so replaced with **None** • Applicants with experience >=1 are replaced with **Others.**
Role	• Applicants with 0 experience will not have the department.so

	replaced with **None** • Applicants with experience >=1 are replaced with **Others.**
Industry	• Applicants with 0 experience will not have the department.so replaced with **None** • Applicants with experience >=1 are replaced with **Others.**
Designation	Not treated as the variable is dropped
Graduation_Specialization	Not treated as the variable is dropped
University_Grad	Not treated as the variable is dropped
Passing_Year_Of_Graduation	Not treated as the variable is dropped
PG_Specialization	Not treated as the variable is dropped
University_PG	Not treated as the variable is dropped
Passing_Year_Of_PG	Not treated as the variable is dropped
PHD_Specialization	Not treated as the variable is dropped
University_PHD	Not treated as the variable is dropped
Passing_Year_Of_PHD	Not treated as the variable is dropped
Last_Appraisal_Rating	All the null values in Last_Apprasal_Rating belongs to Applicants who are freshers so replaced with **NA**(not applicable).

Univariate analysis

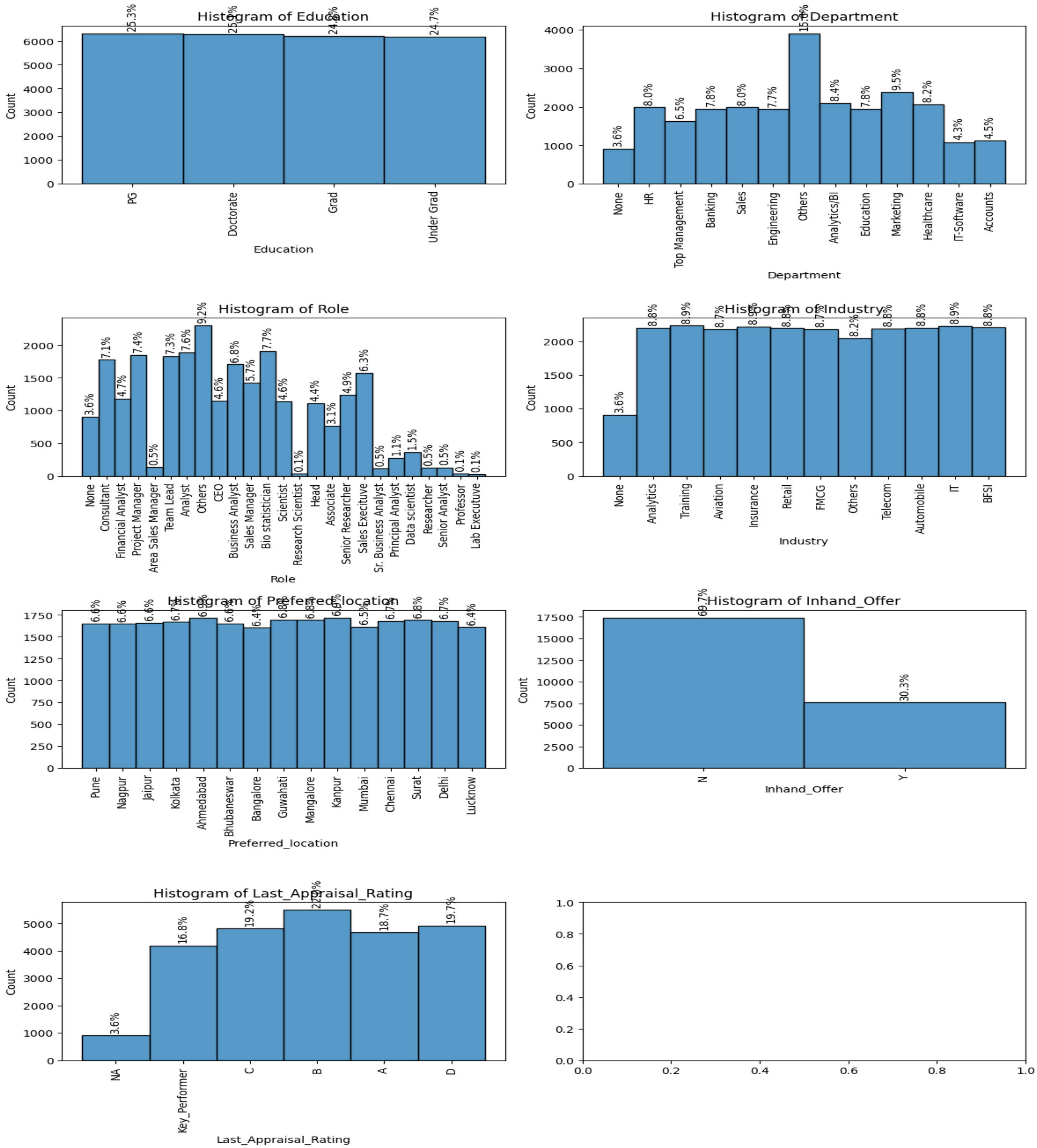

Figure 1: Univariate Analysis of Categorical Variable

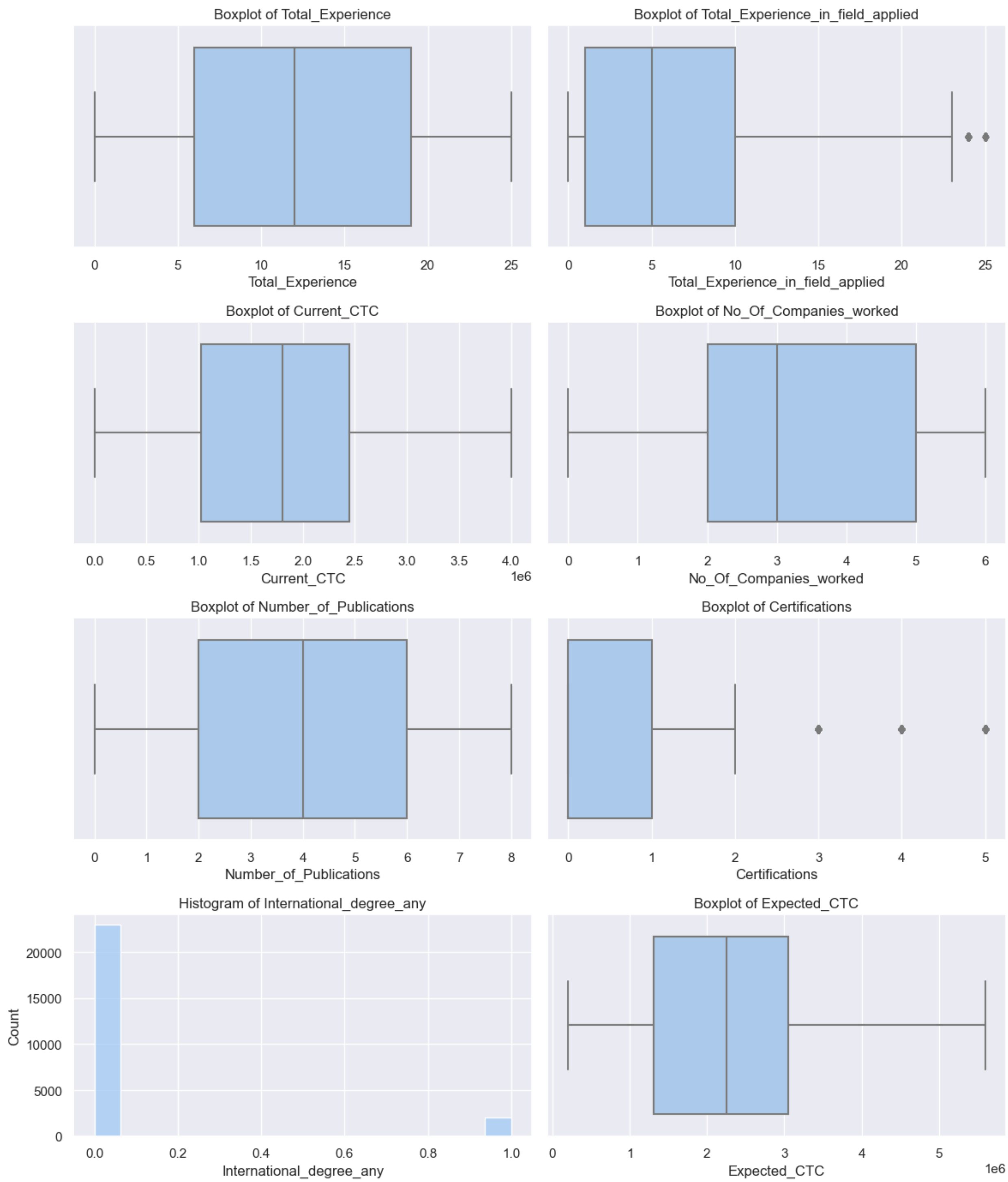

Figure 2: Univariate Analysis of Numerical Variable

Bivariate Analysis :

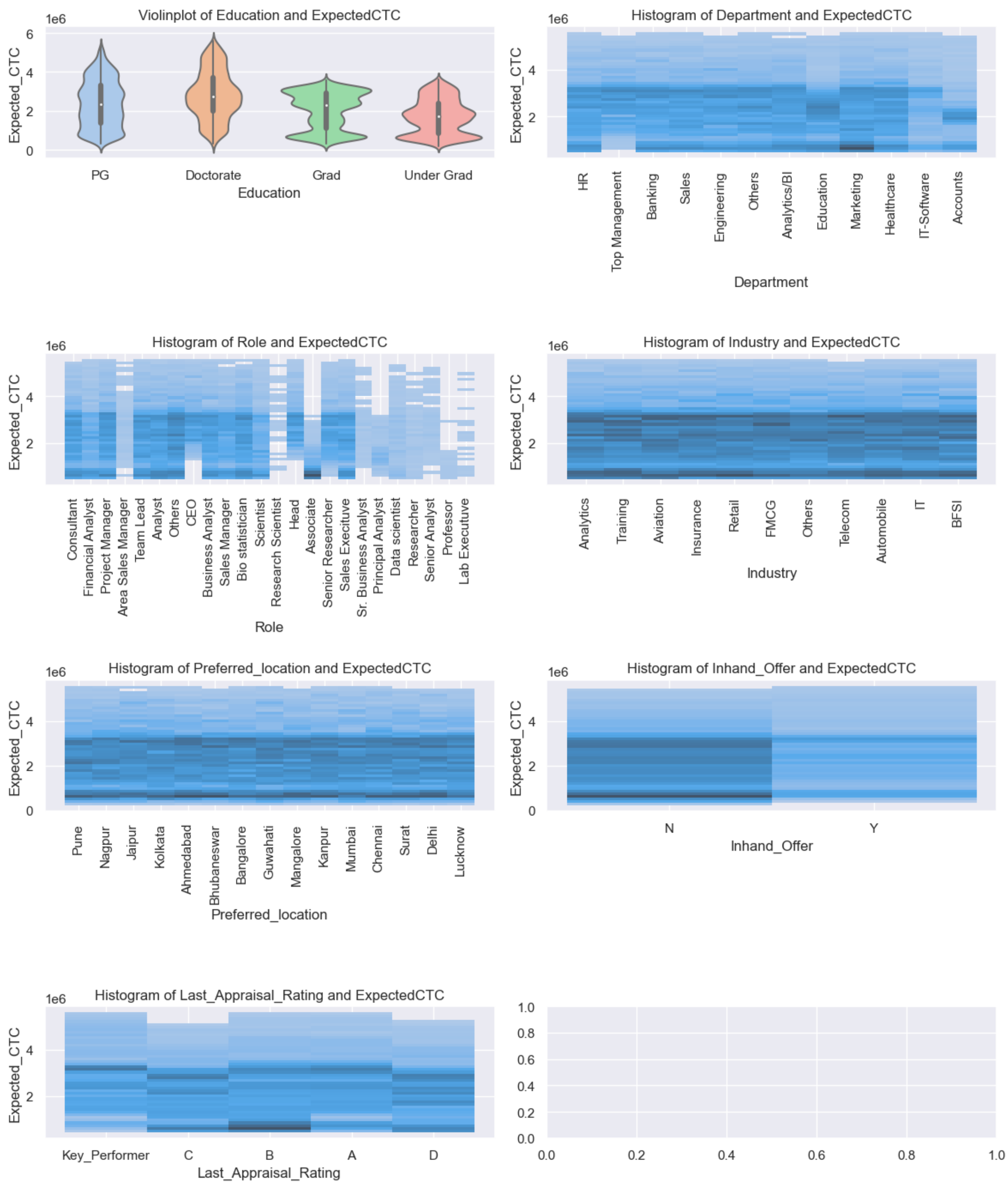

Figure 3: Bivariate Analysis With Target Variable Expected CTC(categorical)

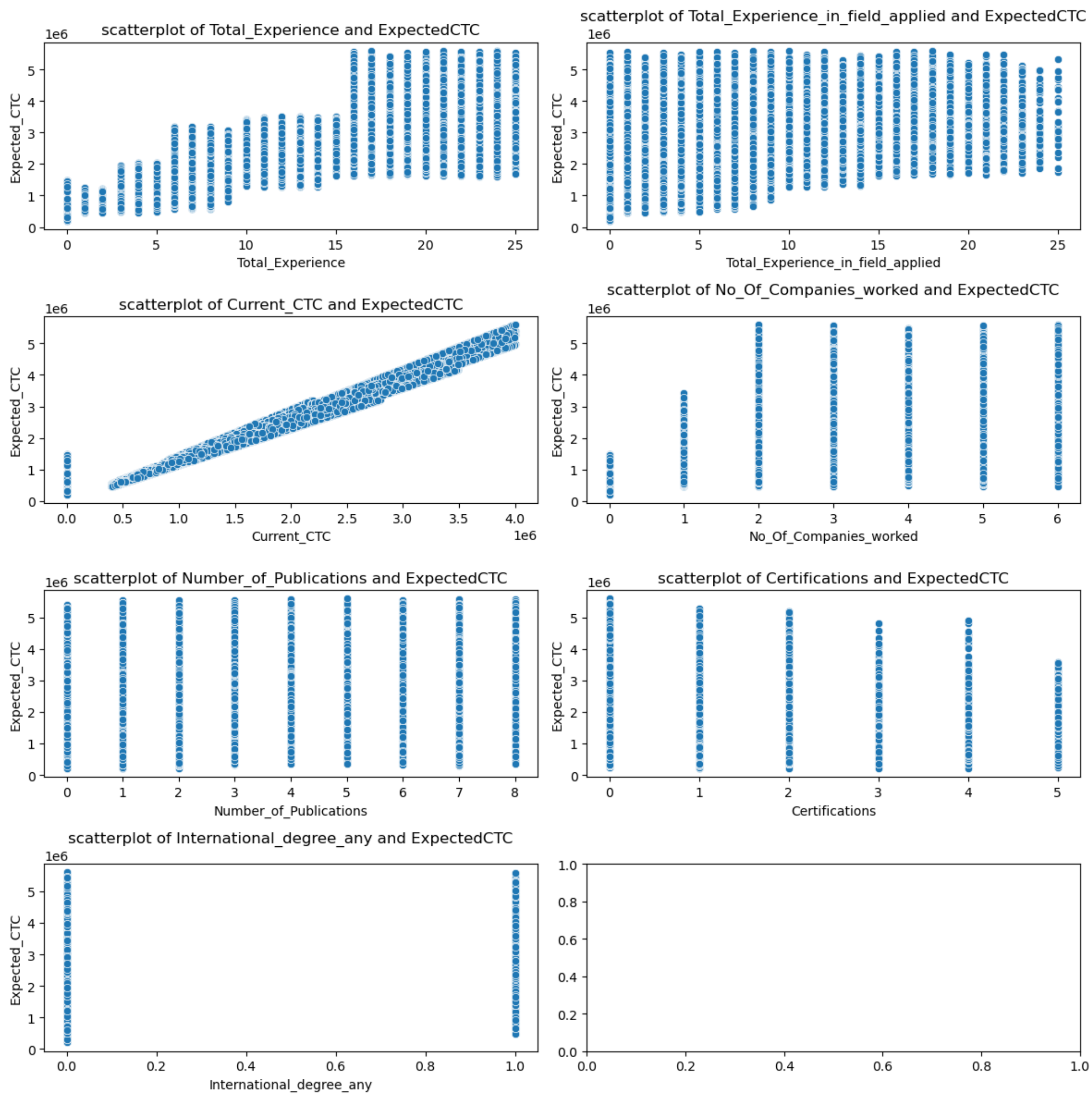

Figure 4: Bivariate Analysis With Target Variable Expected CTC(numerical)

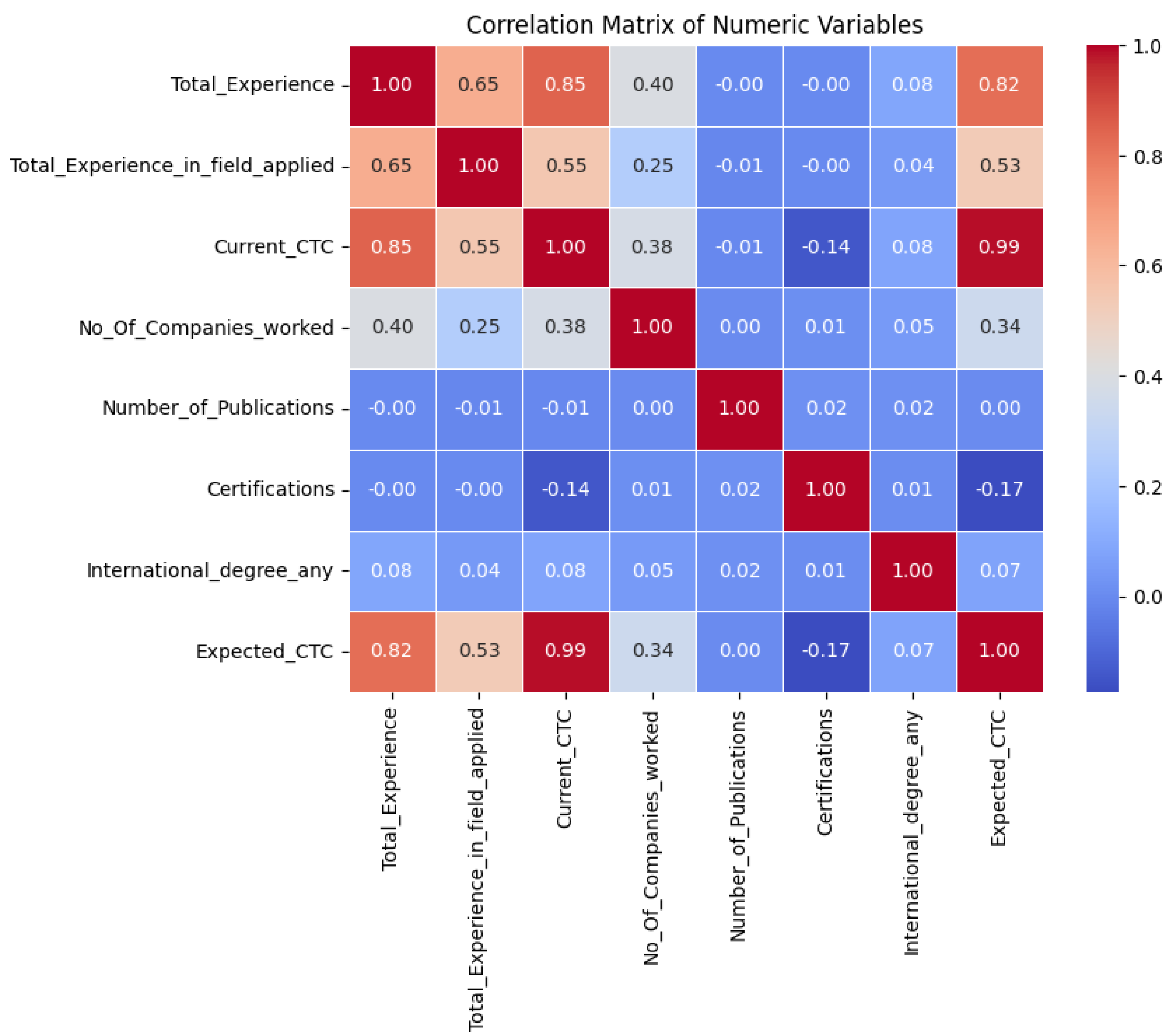

Figure 5: Correlation Matrix of Numerical Variables

Observations:

- Marketing department has more applicants followed by Analytics/BI.
- Applicants with no inhand offer are more compared to Applicants with inhand offer.
- There are outliers in Total_Experience_in_field_applied, Certifications.
- The Expeted CTC is higher for applicants with higher qualifications PG ,Doctorate compared to Grad and UnderGrad.

- Expected CTC does not change much for different locations in Preferred_location.so this might not contribute much on predicting ExpectedCTC and can be dropped.
- There is a high correlation between Current CTC and Expected CTC.
- There is a good correlation between Total_Experience and Expected_CTC
- The Expected_CTC of applicants with more certifications is less compared to applicants with no certifications.
- **Key Performer,A,B Appraisal rating applicants Expected_ CTC more compared to others.**
- **Fresher Applications are Very Less compared to Experienced and Expected_CTC is also low w.r.to Experienced**
-

Outlier treatment

Identified Outliers using BoxPlot .There are outliers in Total_Experience_in_field_applied, Certifications.But outliers seems genuine so not treating it.

Variable transformation

- Performed one-hot encoding for Education,Department,Role,Industry and Last_appraisal_rating
- Performed label encoding for Inhand_Offer.

Addition of new variables

There is no need of adding new variable .The variables present are enough for predicting Expected CTC.

4. **Model building and Model Validation**

Salary Prediction Data is a Regression Data ,So I used Regression models for Predicting the Salary. Following are the Models Used for Prediction.

- **Linear Regression:**Linear regression is straightforward and provides easily interpretable coefficients that show the relationship between the features and the target variable.
- **Lasso Regression:**Lasso Regression is the Extension of Linear Regression. Lasso regression performs L1 regularization, which can shrink some coefficients to zero, effectively selecting a simpler model that may generalize better
- **Decission Tree:**Splits data into subsets based on feature values, Can capture complex interactions between features, Does not require feature scaling
- **Gradiant Boost Ensemble Model:**Sequentially builds models to correct errors of previous models, Can handle both linear and non-linear relationships, Often requires careful tuning of hyperparameters
- **XG Boost Regressor:**An optimized implementation of gradient boosting, Handles missing data well and can be parallelized
- **Random Forest Regressor:**Combines multiple decision trees to improve predictive performance and reduce overfitting

All the Models are Evaluated using R2 and MAPE

Model	**Dataset**	**R-Square**	**MAE**	**MAPE**
Linear Regression	Train	0.9957	50733	0.0421
	Test	0.9958	50441	0.0412
Lasso	Train	0.9957	50731	0.0421

Regression				
	Test	0.9958	50438	0.0412
Decission Tree	Train	0.9994	3304	0.0053
	Test	0.9990	13834	0.0121
Gradiant Boost Ensemble Model	Train	0.9982	27880	0.0214
	Test	0.9982	27929	0.0205
XG Boost Regressor	Train	0.9993	11177	0.0100
	Test	0.9992	13362	0.0112
Random Forest Regressor	Train	0.9994	6866	0.0076
	Test	0.9992	12525	0.0112

- From the Above table we can see that Random Forest Regressor and XG Boost Regressor are best models for predicting.
- XG Boost Regressor got 99.93 and 99.92 accuracy for Train and Test Data
- Random Forest Regressor got 99.94 and 99.92 accuracy for Train and Test Data.

5.Final Interpretation/recommendations

- Random Forest Regressor Model is Able to Predict Salary More accurately Compared to Other Models
- Total Experience,Total Experience in field Applied,Current CTC,Inhand Offer,No of Companies worked,No of Publications,Certifications,International Degree any,Expected

CTC,Department,Role,Industry,Education,Last Appraisal Rating . These are the features that are important for fair salary predictions

- Accurate salary predictions help the company forecast its payroll expenses more precisely. This enables better budget planning and financial allocation, ensuring that resources are optimally distributed across departments and projects