***2022***

**Prepared by:**

**Arya Sonthalia**

**10/7/2022**

Capstone project: HR\_ data

**TABLE OF CONTENTS**

**CONTENTS**

1. **Introduction**
2. **EDA and Business Implication**
3. **Data Cleaning and Pre-processing**
4. **Model building**
5. **Model validation**
6. **Final interpretation / recommendation**

**LIST OF TABLES :**

**Table 1:Sample data………………………………………………………………………………………………………..**

**Table 2:Data Info…………………………………………………………………………………………………………...**

**Table 3 : describe table ………………………………………………………………………………………………….**

**Table 4: Skewness ………………………………………………………………………………………………………….**

**Table 5:Correlation…………………………………………………………………………………………………………**

**LIST OF FIGURES :**

**FIGURE 1: count plot of preferred location……………………………………………………………………..**

**Figure 2 : HIST PLOT of variables …………………………………………………………………………………….**

**FIGURE 3 : Box plot of the variable …………………………………………………………………………………**

**FIGURE 4 : count plot of inhand offer……………………………………………………………………………..**

**Figure 5 : scatter plot total experience and expected CTC ………………………………………………**

**Figure 6: scatter plot total experience and expected CTC ………………………………………………**

**Figure 7: bar plot of expected CTC and education…………………………………………………………..**

**Figure 8: bar plot of expected CTC and preferred location………………………………………………**

**Figure 9: bar plot of expected CTC and in hand offer……………………………………………………..**

**Figure 10: bar plot of expected CTC and Designation……………………………………………………..**

**Figure 11:pair plot of variables………………………………………………………………………………………**

**Figure 12:Heat map……………………………………………………………………………………………………….**

**FIGURE 13:Histogram and box plot before outlier treatment………………………………………..**

**Figure 14: Histogram and box plot before outlier treatment………………………………………….**

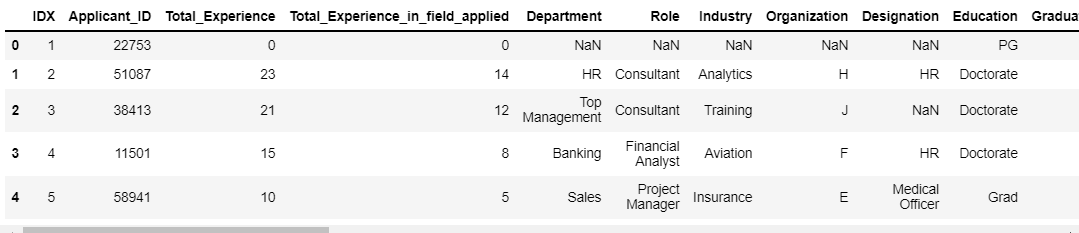
**Figure 16:HISTPLOT OF ALL THE VARIABLES AFTER OUTLIER TREATMENT……………………..**

**Introduction**

**The given data is of the individuals who have applied in delta Ltd.. The organisation is particular in giving salaries to individuals for their role. It understands the importance of determining and maintaining the right salary range for all the employees serving the organisation with similar profiles. The data gives us deep insight of the various factors and skills of employees that are considerable to determine the abilities and factors for which they were evaluated. Delta ltd wants to focus on avoiding the human judgement and is keen on building a non-biased model to determine the salary of employee based on their best abilities.**

**Indeed it’s very important for organisation to have a policy were in employees are rewarded with salary for their best ability rather than based on human judgement. It creates healthy work environment and shows high the ethical standards of the organisation. Data driven salary package will have accuracy and unbiased offers for the employees.**

**The objective here is to achieve a robust approach in offering salary so that best work force is created for the organisation. Right approach towards creating salary range will have high ethical standards for organisation and accurancy in determining the appropriate salary package without discrimination among the similar employee profile.**

****

**EDA and Business Implication**

**The data collection seems to done during interview process of Delta ltd were in the existing salary and other considerable factors regarding employee’s experience and other abilities to which they were already been evaluated**

**The data consists of 29 columns and 25000 rows. The columns present gives us insights of IDX, Applicant ID, Total experience, total experience in field applied, department, role, Industry, Organisation, designation, education, Graduation specialization, university graduation, passing year of graduation, specialization PG ,university PG, passing year of PG, Specialization PHD, university of PHD ,passing year of PHD, Current location, preferred location, current CTC, In-hand offer, Last appraisal rating, no. of companies worked, number of publications, certification, International degree and expected CTC. Though there are couple of columns which gives us details irrelevant to our objective and need data cleaning before we create a model.**

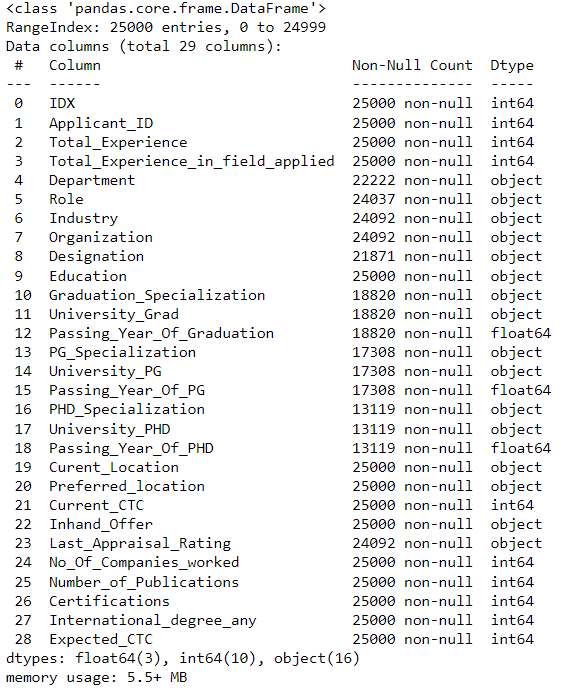
**‘IDX,’ApplicantID’,’Department','Organization','Passing\_Year\_Of\_Graduation','Passing\_Year\_Of\_PG','University\_Grad','PG\_Specialization','University\_PG','PHD\_Specialization','University\_PHD','Passing\_Year\_Of\_PHD','Curent\_Location','Graduation\_Specialization','Role','No\_Of\_Companies\_worked are the columns that do not help us in achieving the objective thus need to be dropped. The target variable here is the Expected CTC of employee. we do observe null values presence in data. Other variables do have an important significance as they give us insight of the qualification, certifications , preferred location were employee can be placed, the current salary and appraisal received in the previous company, their experience, department in which they are currently working.**

**Understanding of attributes :**

**The attributed of the data consists of 25000 enteries under 29 rows. Most of the attributes are continuous. Variables like department, role, industry, organisation, education, preferred location intergers, objects and float data type distributed in 29 columns. Expected CTC seems is the target variable present in the dataset. There are no duplicate values in the data. Once we dropped not so required columns we are left with 13 columns**

**Industry, designation and last appraisal rating do show null values which has to be treated.**

**, last appraisal and in-hand offer are categorical variable and thus needs to be converted into numerical. The data is having**

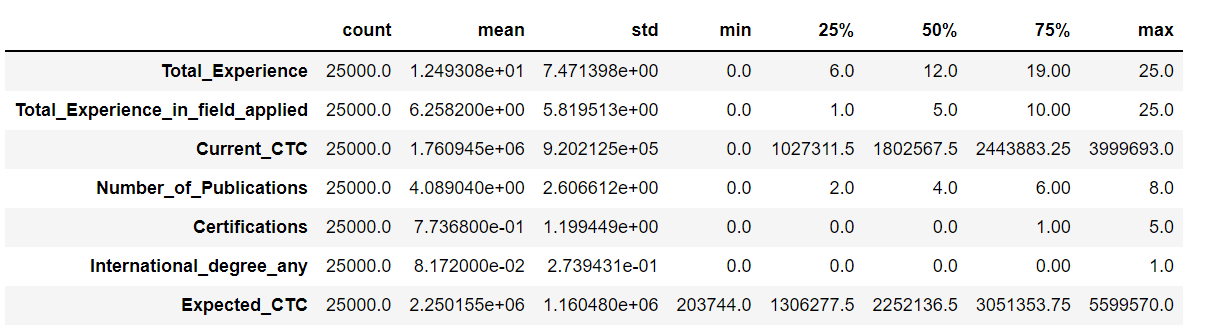
****

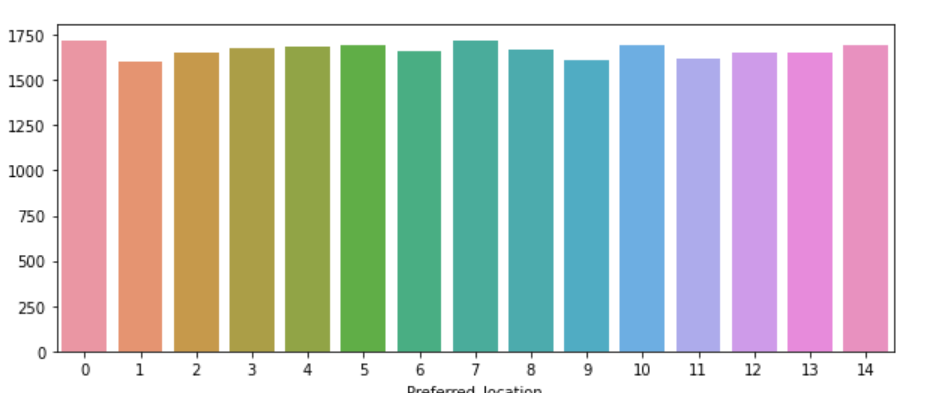
**Univariate analysis:**

**To understand the univariate analysis let us check the continuous variable and understand the summarised statistical dataset. In the table below we can find the count, minimum and maximum value,standard deviation of each continuous variable.**

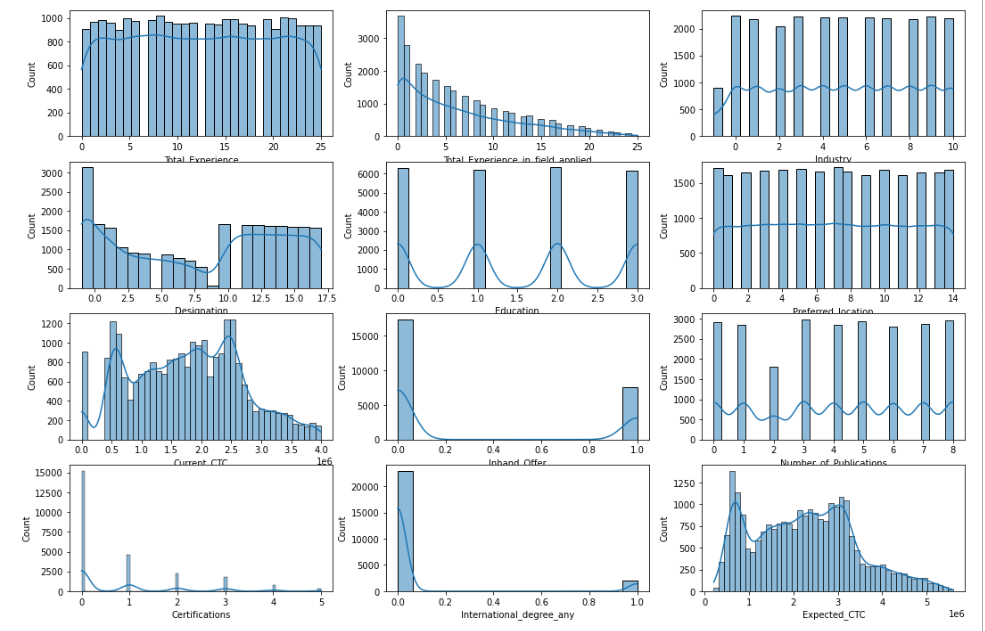
* **The minimum value of target variable Expected CTC 203744.0 were as the maximum 5599570.0**
* **The current CTC is minimum of 1027311.5 were as the maximum current CTC 3999693.0**
* **Employees applying for the Delta Ltd has maximum of 5 certification and a 75% of tem atleast have 1 certification**
* **The maximum experience in the field applied by employee is 25 years were as 75 % of the applicant do have experience have 10 year, we even see fresher applying and 25% of the total applicant have a minimum of 1 year experience. Thus the average applicant have atleast 6 years of experience**

**Table 3 : describe table**





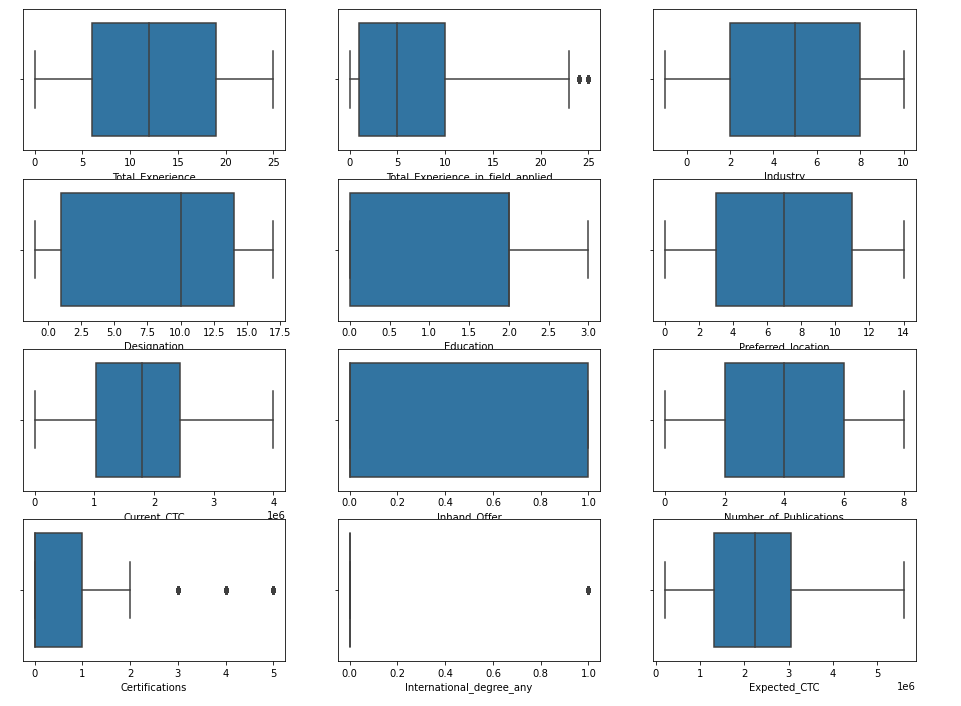
**FIGURE 1: count plot of preferred location**

****

**Figure 2 : HIST PLOT of variables**

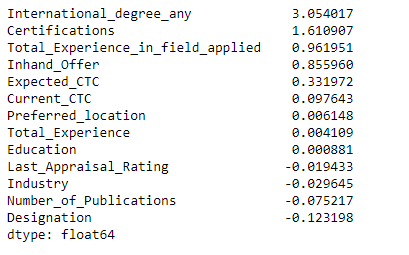
**INSIGHTS:**

**Hist plot above gives us deep insight of the data distribution. We do observe right side skewness.**

****

**INSIGHTS :**

**We do observe outlier through the box plot presentation and it is a right skewed data we shall use log transformation method to treat outlier**

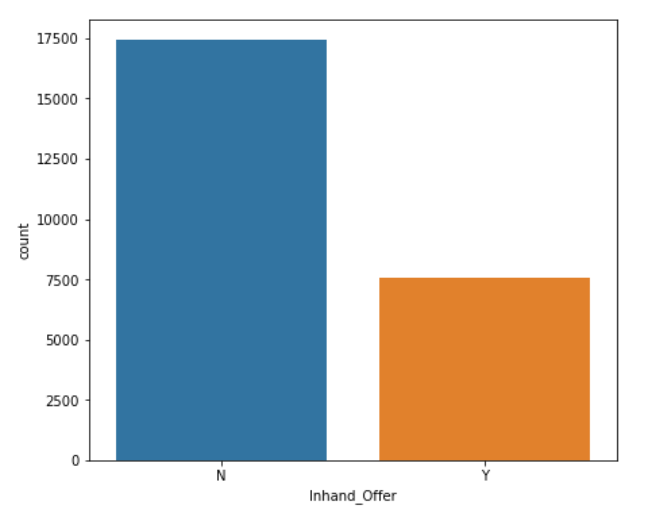
****

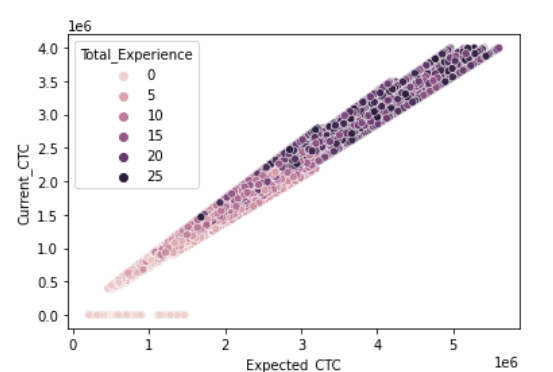
**Insight:**

**Positive skewness can be observed**

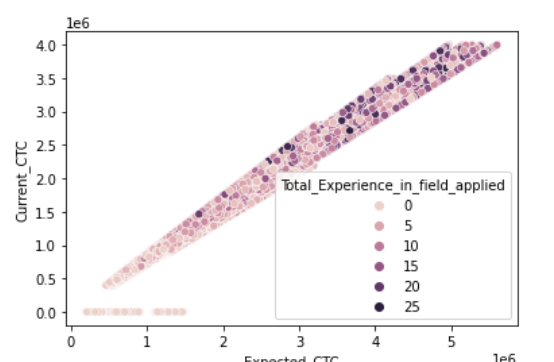
**Bivariate analysis**

**FIGURE 4 : count plot of inhand offer**

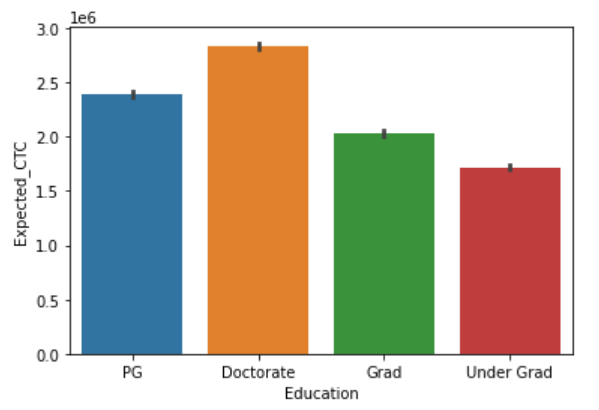
****

****

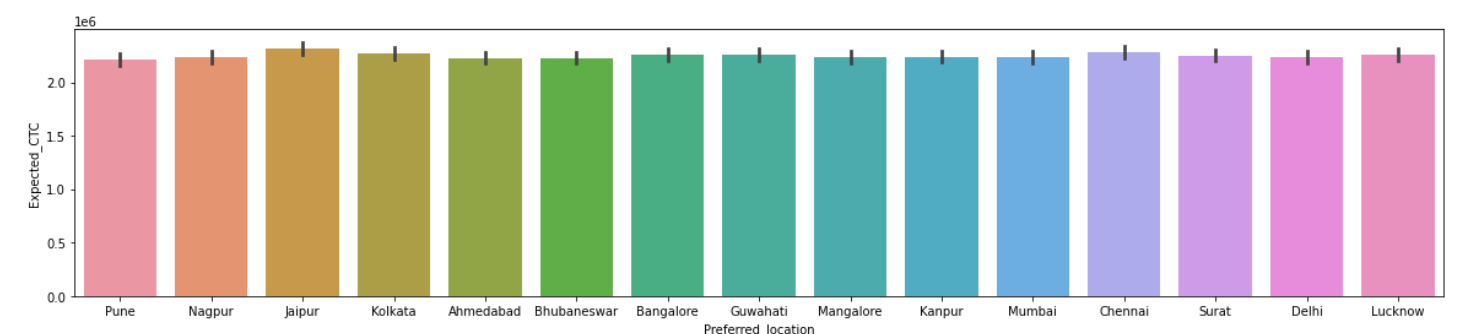
**Figure 6: scatter plot total experience and expected CTC**

****

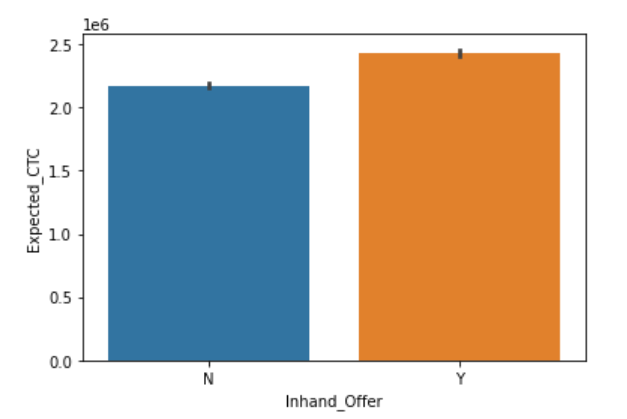
**Insight: In general scenario when the experience is high the expected ctc and currect ctc are higher yet It can be seen that employees with less experience in the field as well are having higher current CTC and expecting high CTC. Thus having experience in the field does play a key role.**

****

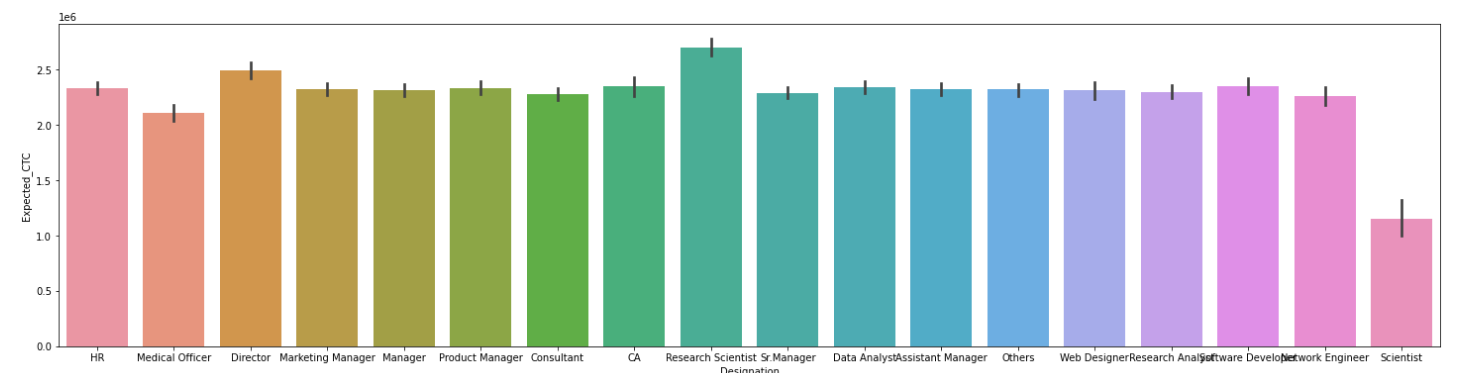
**Figure 8: bar plot of expected CTC and preferred location**

****

**Figure 9: bar plot of expected CTC and in hand offer**

****

**Figure 10: bar plot of expected CTC and Designation**

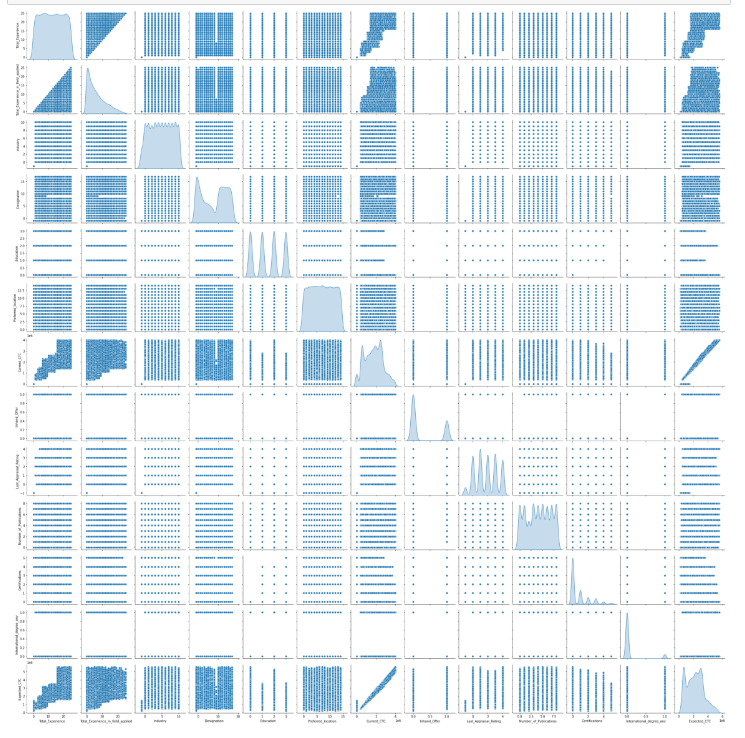
****

**INSIGHTS:**

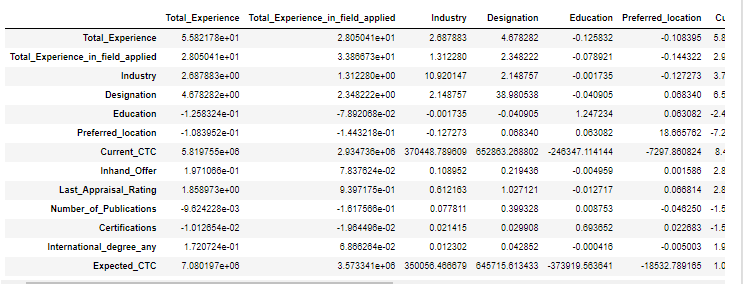
* **There are many candidates who do have inhand offer**
* **Doctarates have higher Expected CTC**
* **Total experience does matter in determining the CTC, also experience in the field applied also is related to the Expected CTC**
* **Those candidates who have In hand offer are having higher expected CTC**

**Multi-variant Analysis**

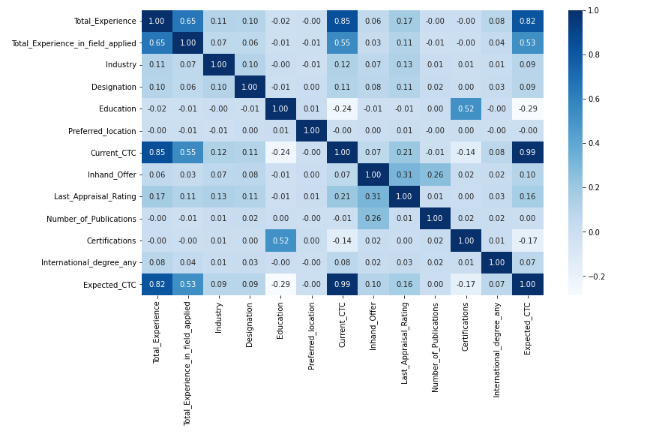
**Figure 11:pair plot of variables**



**Table 5:Correlation**



**Figure 12:Heat map**



**Insights:**

* **Current CTC and Expected CTC hare highly correlated**
* **Education and expected CTC are having fair correlation**
* **Correlation can be seen between total years of experience in the same field and expected CTC**

**Data Cleaning and Pre-processing**

1. **Removal of unwanted variables:**

**There are data present which do not have any relation or significance with the target variable so we have dropped those variables**

**'IDX','Applicant\_ID','Department','Organization','Passing\_Year\_Of\_Graduation','Passing\_Year\_Of\_PG','University\_Grad','PG\_Specialization','University\_PG','PHD\_Specialization','University\_PHD','Passing\_Year\_Of\_PHD','Curent\_Location','Graduation\_Specialization','Role','No\_Of\_Companies\_worked'**

**b)Missing Value treatment**

**Data does have missing values in it and after transformation we have imputed the data using KNN imputer**

**e)Variable transformation**

**There are variable which are categorical and hence we have converted them into numerical by transforming them into binary digits**

* **Industry**
* **Designation**
* **Education**
* **Preferred location**
* **In hand offer**
* **Last appraisal rating**

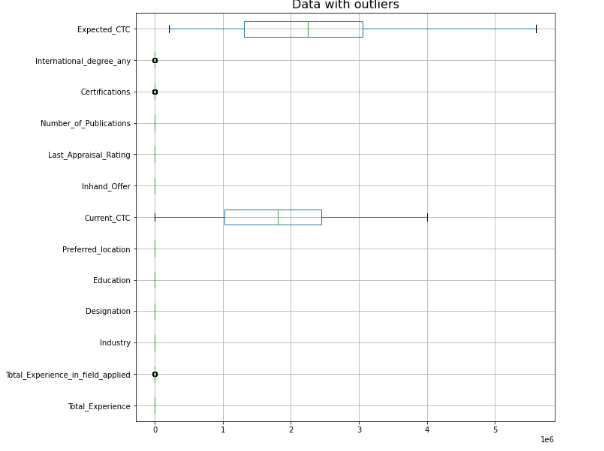
**d)OUTLIERS:**

**We do observe data to be positively skewed and skewness is towards the right**

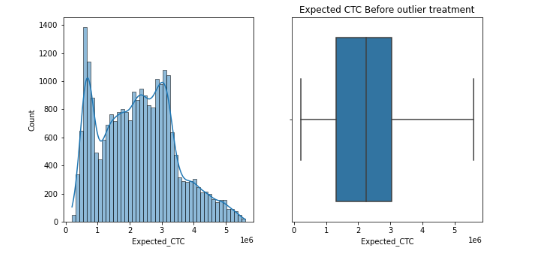
**Thus we have choosen to treat outliers by log transformation method and we could clearly identify the change in the data by treating the outlier.**

**e)Variable transformation**

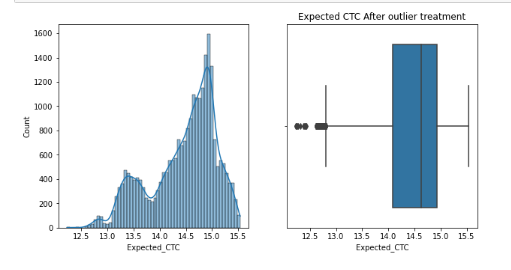
**There are variable which are categorical and hence we have converted them into numerical by transforming them into binary digits**



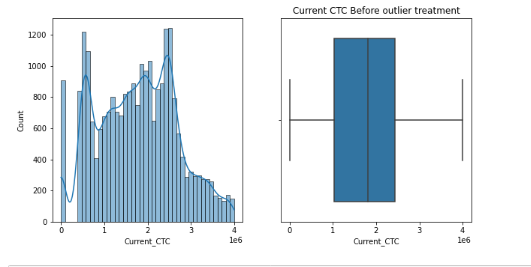
**Figure 12:Histogram and box plot before outlier treatment**

****

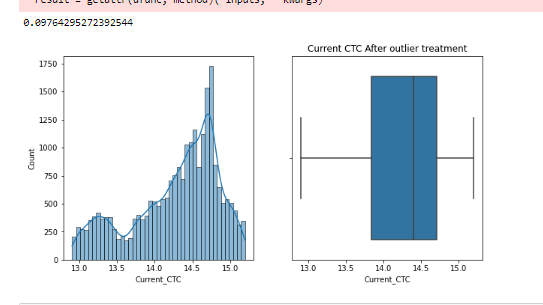
**FIGURE 13:Histogram and box plot before outlier treatment**



**Figure 14: Histogram and box plot before outlier treatment**

****

**Figure 15: Histogram and box plot before outlier treatment**

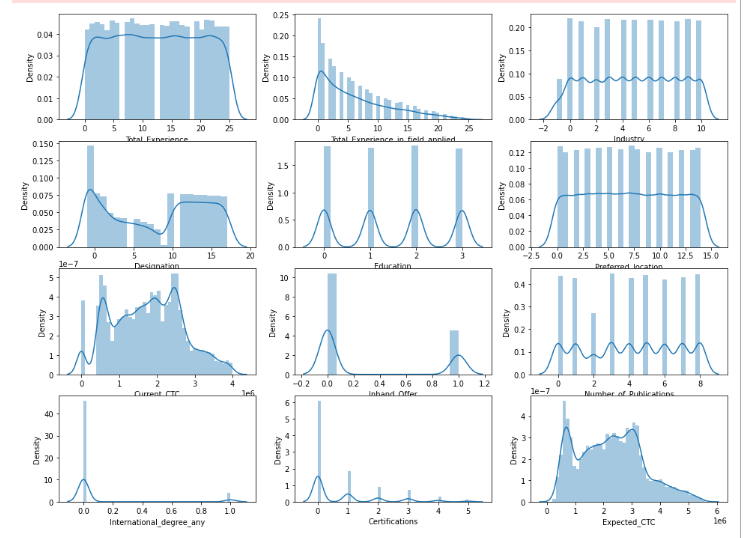


**INSIGHTS:**

**We do observe outliers in the data and treated it using log method.**

**The data showed positive skewness which was taken care by treating it**

**Figure 16:HISTPLOT OF ALL THE VARIABLES AFTER OUTLIER TREATMENT**



**4. Business insights from EDA**

**a) Is the data unbalanced?**

**Solution:**

**The given data is a unbalanced data, there are missing values present, we do see outliers and needs imputation**

**If so, what can be done?**

**The categorical column is converted into numerical so that we are able to get that the entire data and it’s insight**

**As we observe the given data is not very huge data and removing any part of data can impact the entire analysis process hence we only treat the outliers using log transformation method and try to achieve bell curve**

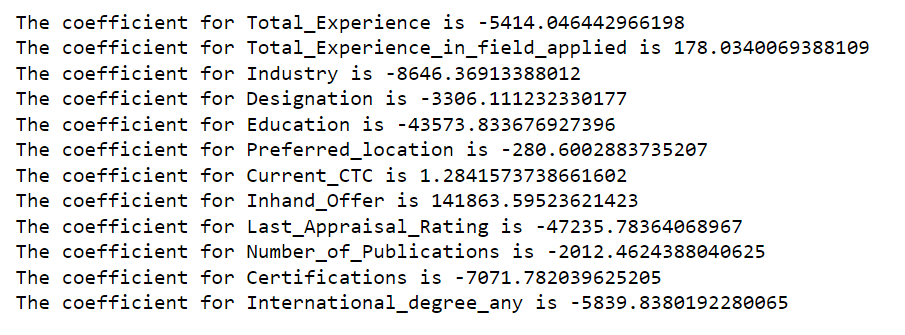
**and treat null values using KNN imputer**

**b) Any business insights using clustering (if applicable): we have used imputation technique of k nearest neighbour which fills the null values as per the values present near the data.**

**c) Any other business insights: The data could have been clearer and more specific details would have been helpful. Unbiased environment is definitely a very important step. we have seen the data highly co-related . we can build a model which can categorise the different variable, it relation with the target variable.**

**Model building - Clear on why was a particular model(s) chosen. - Effort to improve model performance.**

Table 2: coefficient



INSIGHTS:

* We have applied linear regression model as the variables are continuous. we can see the coefficient of each variable.
* Variable Total expected CTC, Total current CTC and current variable have positive coefficient and rest of the variable’s coefficient is negative.



R Square of linear regression model for train:



R Square of linear regression model for testing:



RMSE TRAIN :



RMSE TEST :



Insights:

We observed a very good r square value of the model for both train and test

**CONCATING X AND Y AS SINGLE DATASET**

**Table:2 dataset(concate x and y)**

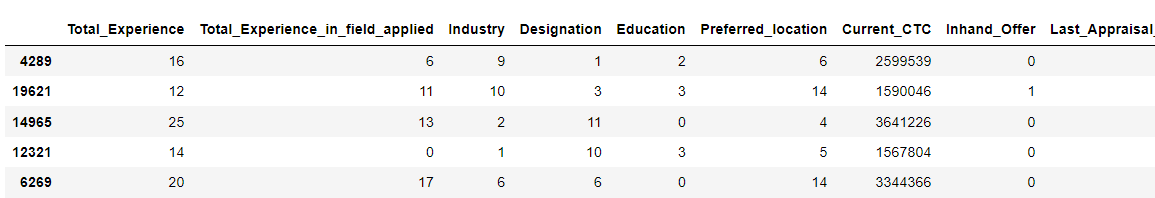


Table 3:LM1 params

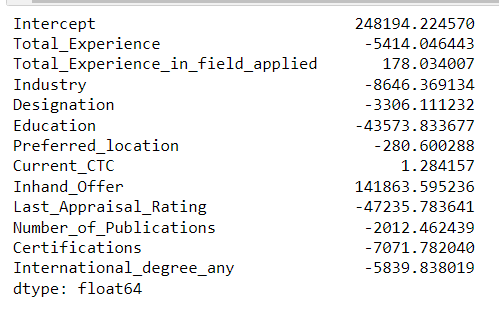
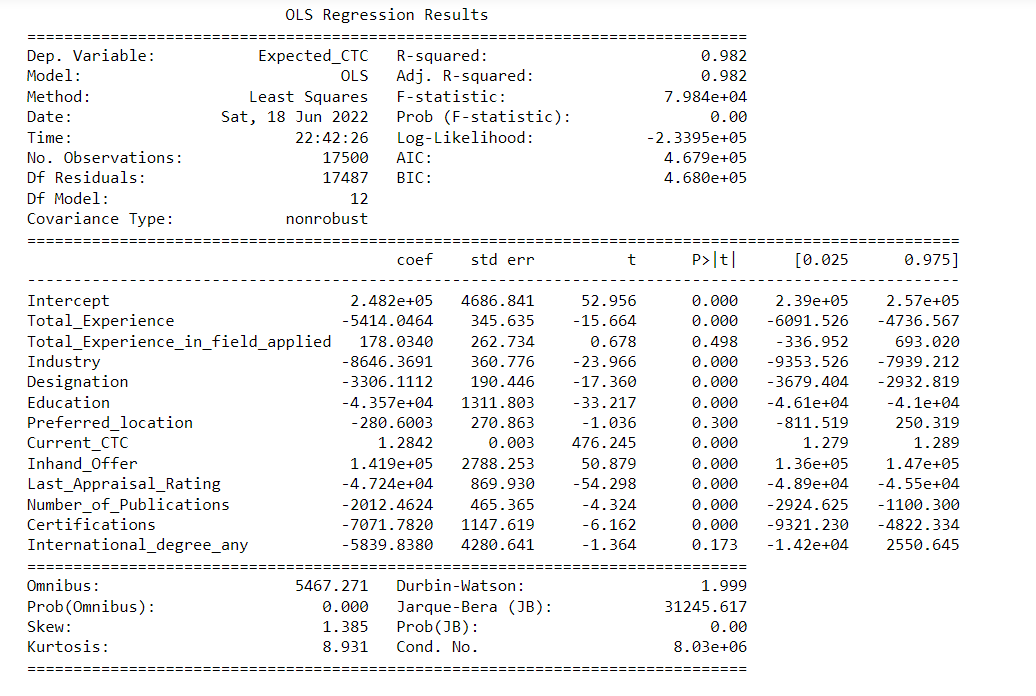


Table 4: summary lm1



IINSIGHTS:

* Includes 12 variables
* Number of observation is 17500
* Intercept is 2.482e+05
* We do have colinerarity
* R-square is 0.98 which is very good score
* P-value is very very low for most of the variable ,though for total-experience in the field applied has 0.4,preffered location has 0.3 and international degree has 0.173 p value

Figure2: scatter plot

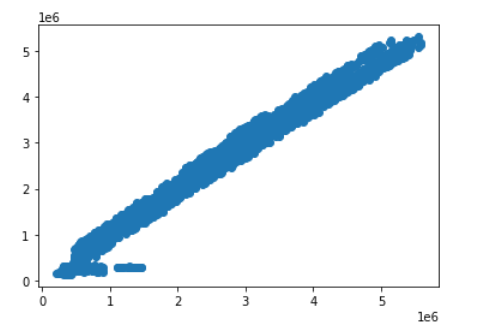
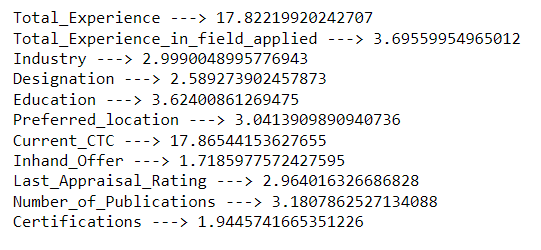


Table 5: vif



VIF determines the collinearity among the variables. We observe the total experience is having higher collinearity so, we must drop the variable and get the results again.

Table 6:lm2 params

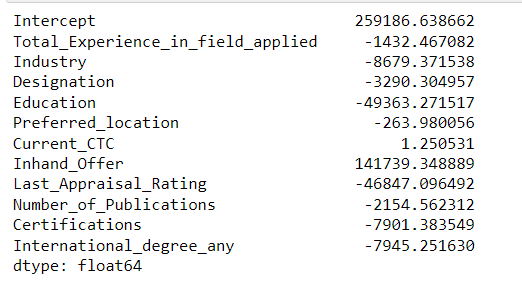
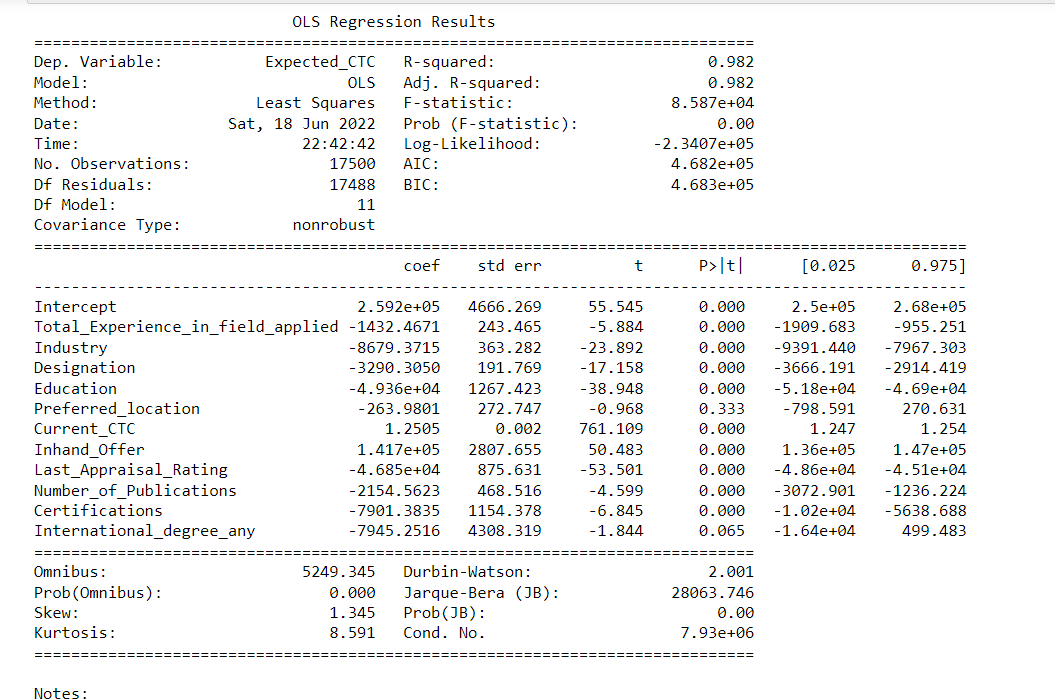


Table 3 : lm2 summary



INSIGHTS:

Insights:

* outliers treated
* R-square-98%
* Adj.R-square-98%
* F-statistic-good amount of relationship between variables and target variable
* P-value of variables are below .05
* High P-value of Variable graduation specialisation 0.763

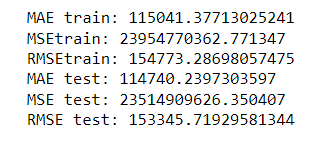
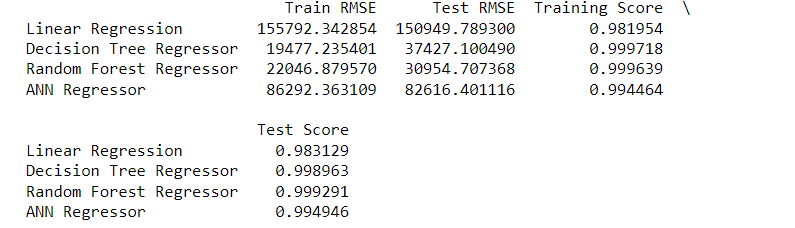
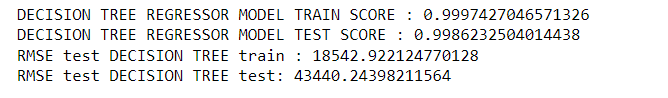


Table 8: regression model:RMSE,TRAIN AND TEST SCORE



* We have built and done a comparative study on the RMSE value of various model and when we observe linear regression to be the best suited as other model fail miserably while testing.

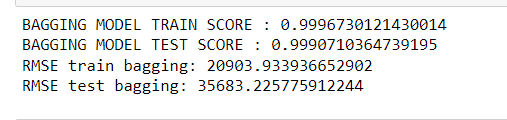
Decision Tree Regressor



* RMSE test is much higher than train

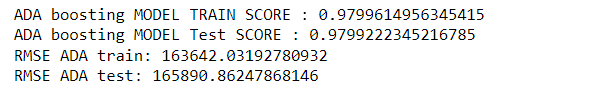
Bagging:

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.



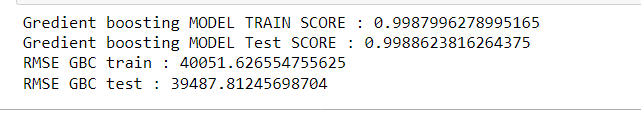
* RMSE test is higher than train

ADA Boosting:



* RMSE test is higher than train

Gredient Boosting:



* RMSE train is higher than test

|  |  |
| --- | --- |
| R-square train | 98.24% |
| R-square test | 98.38% |

**Model validation –**

**Once we build various models we observe and conclude that Linear Regression is the best suited model among all as the R- Square value of the model is excellent. We have got the score of 98%which shows the model is very well made and we can rely on this model.**

**While other failed miserably while tesing**

**Final interpretation / recommendation:**

* **Equalise the salary of employees,steps can be taken for correcting the salary structure.**
* **We can recruit more of Fresher’s and offer them equal salaries based on their educational qualification**
* **Recruitment can be focused on all the states equally, for the benefit of the company.**
* **We can build a fixed salary structure based on experience, education and certifications.**
* **Salary packages should be made data driven rather than human judgements.**
* **The current CTC is one variable which impacts the expected CTC of the employees so, must be taken into consideration while creating the salary structure**
* **We do observe the Experience matters a lot and most of the employees who have applied for the job and have experience are expecting higher CTC.**
* **Most of the locations are preferred choice of the employee hence do not have major effect on the CTC.**
* **Company can think of creating a range of salary or slabs rather than a fixed amount for employees based on their experience and attributes**