**Capstone Project**

**ON**

**CTC Prediction**



**Debsmita Chakraborty**

**Batch-July ‘c’**

TABLE OF CONTENTS

1)Introduction of the business problem\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-3

2)EDA and Business Implication\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-4-12

3)Data Cleaning and Pre-processing\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-13-15

4)Model Building\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-15-21

5)Model Validation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-21-22

6)Final Interpretation/Recommendation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-23

FIGURES and PLOTS

1)Data Dictionary\_ Fig1.1-1.2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Pg-3

2)EDA and Application\_ Fig2.1-2.10\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Pg-4-12

3)Data Cleaning and Pre-processing\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-13-15

4)Model Building\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-15-21

5)Model Validation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Pg-21-22

1) Introduction of the business problem

**Business Problem:**

a) 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.

b) Need of the study/project-The idea is to minimise human intervention to calculate salaries of an employee who is joining the company but automate it with data of the past with reference to his interview. To keep all employees with similar profile under one umbrella pricing.

c) Understanding business/social opportunity- Helping business reduce manual labour and minimize human judgement which at times may be less practical.

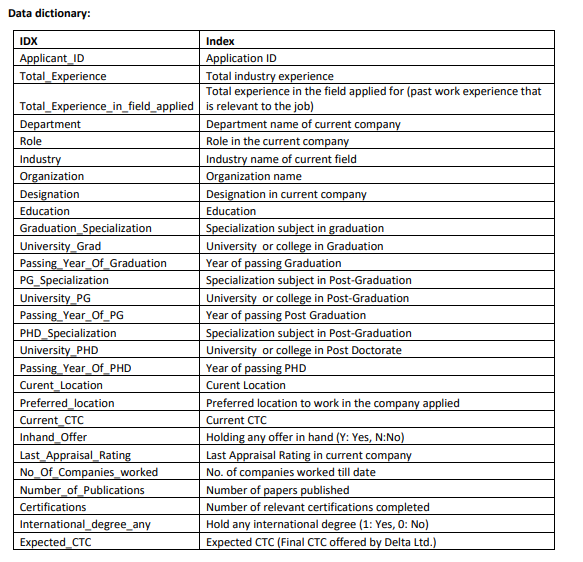


Fig 1.1- Data Dictionary

* The data looks like:

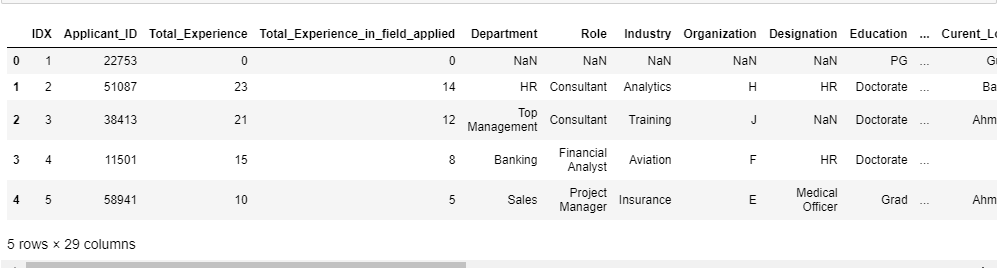


Fig 1.2- Data Heading

**2. EDA and Business Implication**

* The column headings being:

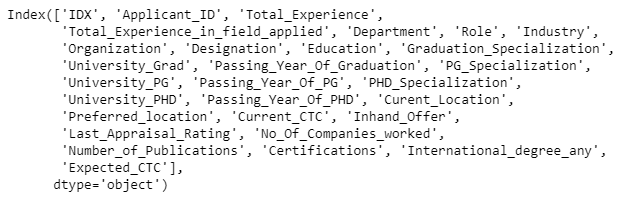


Fig 2.1- Data Column Headings

* The data has 29 columns and 2500 rows. We see below a lot of missing values least being University PHD, PHD specialization and passing year of PHD which is almost 10000 entries missing:

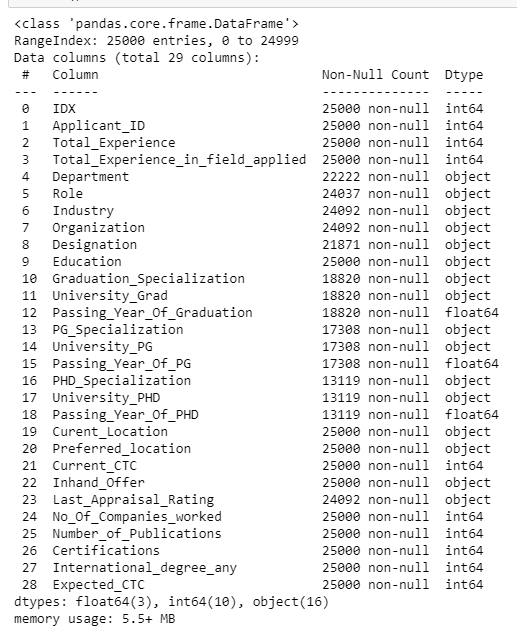


Fig 2.2 – Data Information

* The data shows descriptive analysis wherein we do see a lot of outliers as the difference in mean and maximum is high. A lot of fluctuations in the values can be seen:

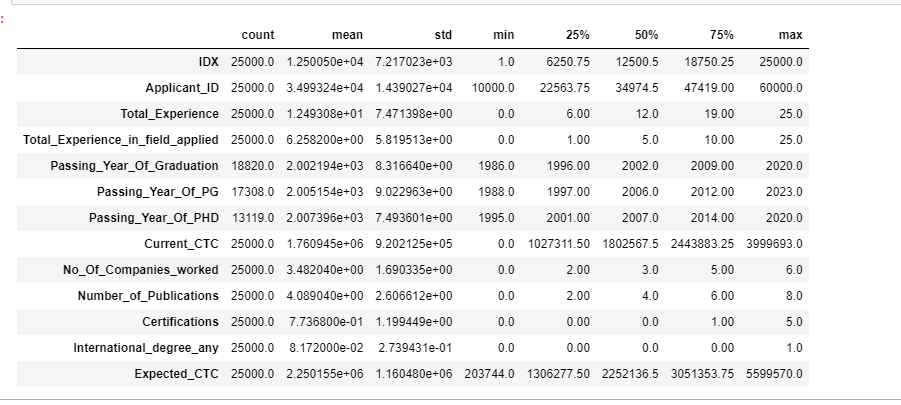


Fig 2.3 Data Described

* Data definitely from its describe function seemed to have outliers as missing values are high but number of duplicate rows is zero.

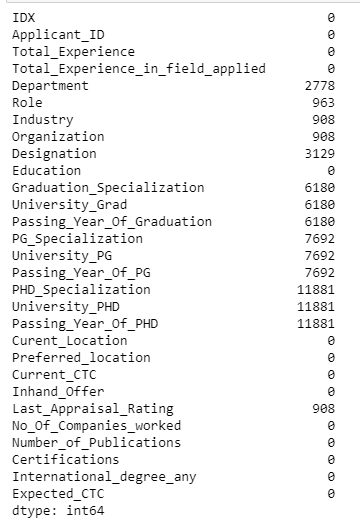
 

Fig 2.4 Null Values and Duplicate rows

* Getting the summary statistics of object data type:

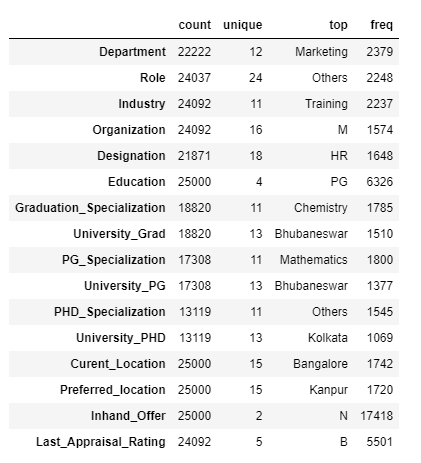
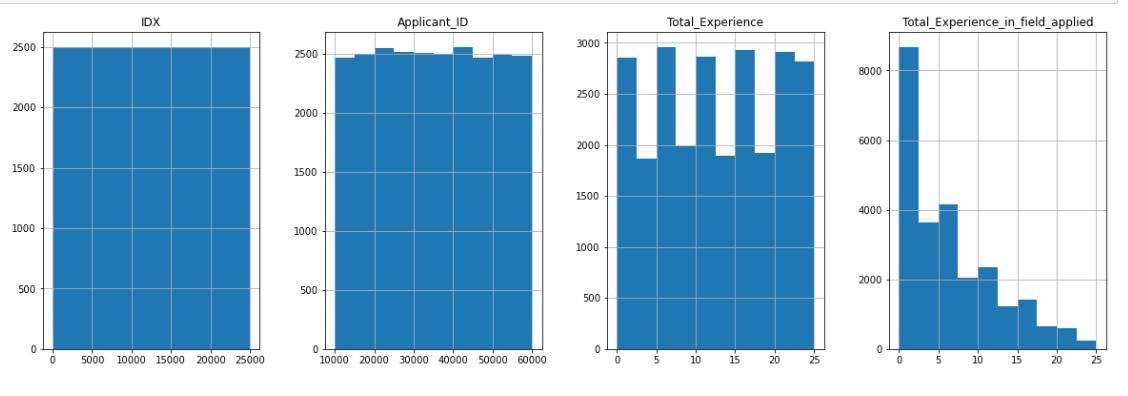


Fig 2.4 Data Summary

**Our target variable is Expected CTC since the entire modelling is about predicting the same.**

**UNIVARIATE ANALYSIS:**

****

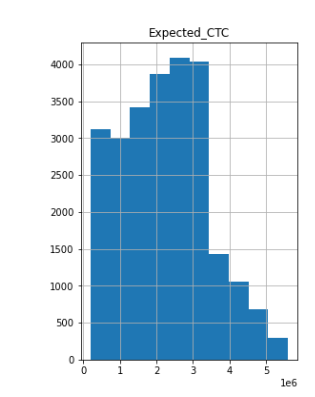
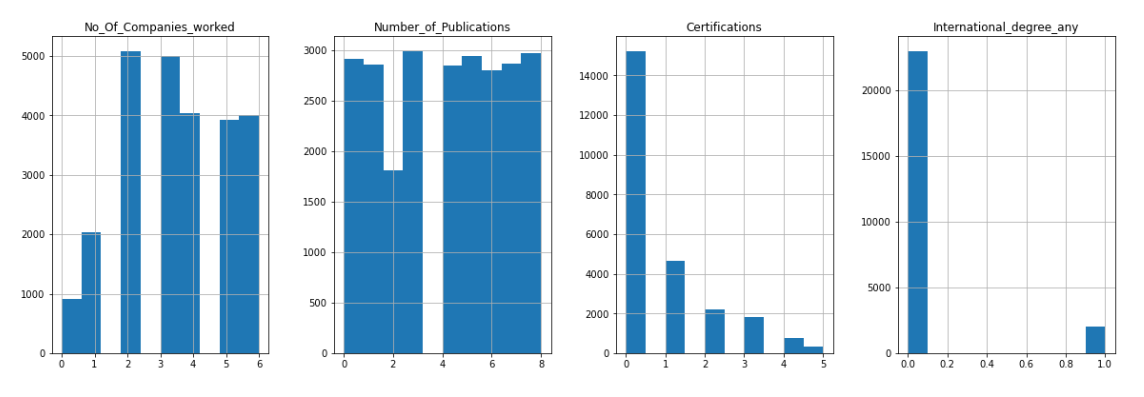
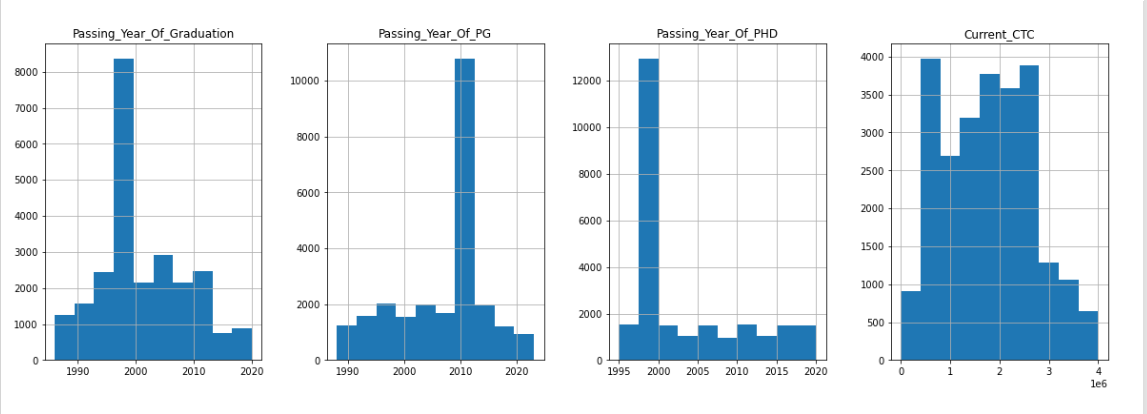
****

Fig 2.5 Histogram

INFERENCE:

* We see a lot of pass outs in the year 2000 and the least in 2014-2015.
* Passing year of PG is the maximum in 2010 and least post 2020.
* Total years of experience is maximum 10, 5 and 25. A uniform distribution of data.
* The total experience in the field applied is 0.
* More people have worked in 2 to 3 companies amongst applicates with 0 certifications Internationally.
* We see Expected CTC at a uniform distribution.

**CATEGORICAL DATA ANALYSIS:**

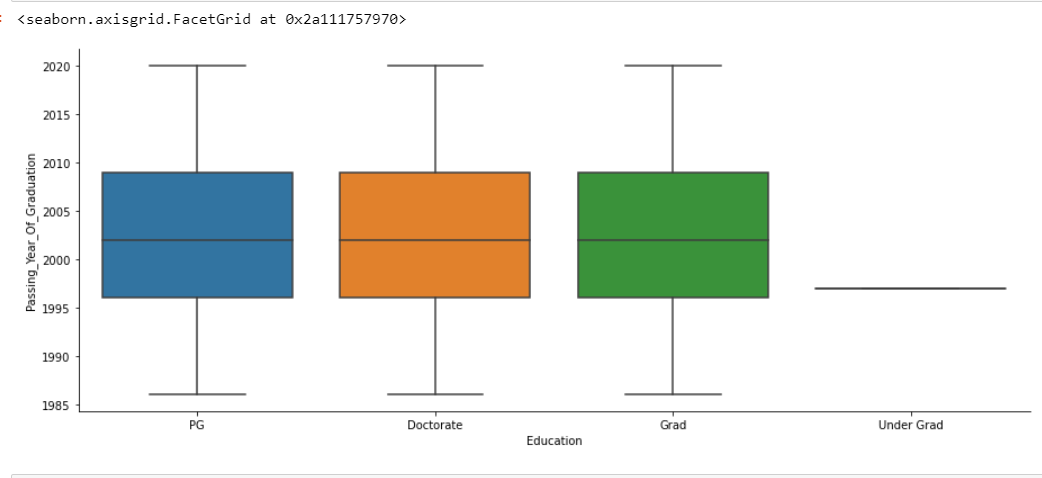
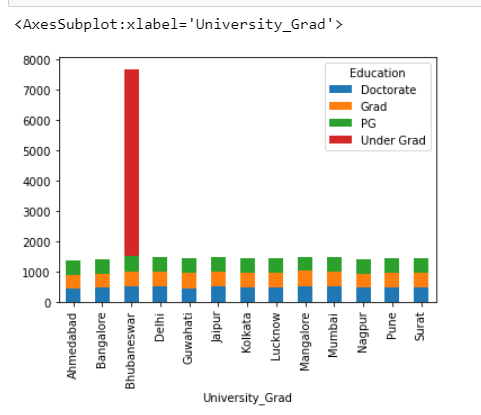
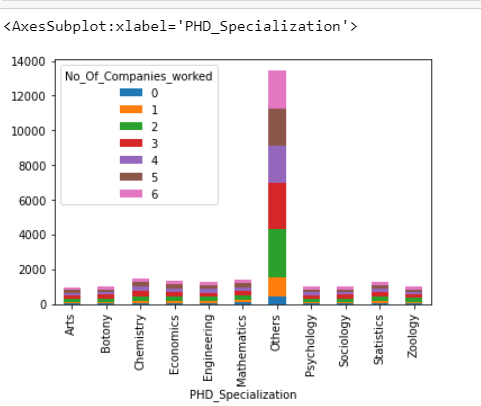
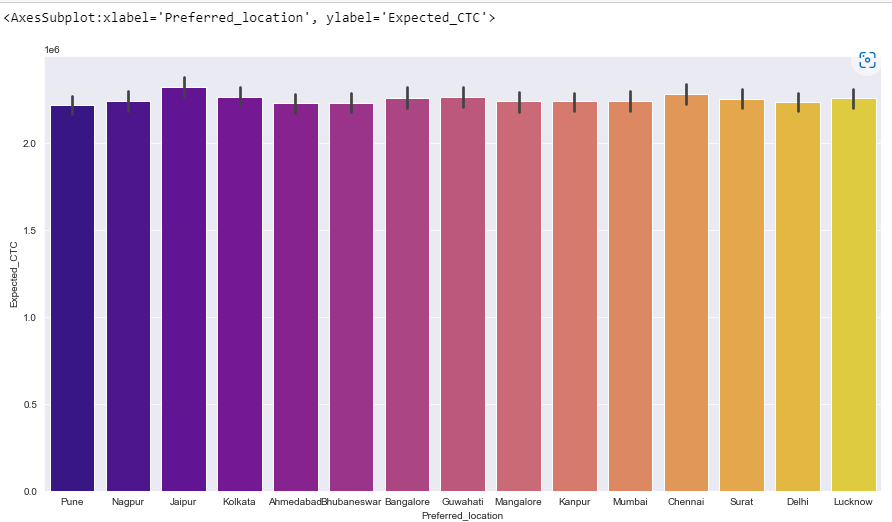
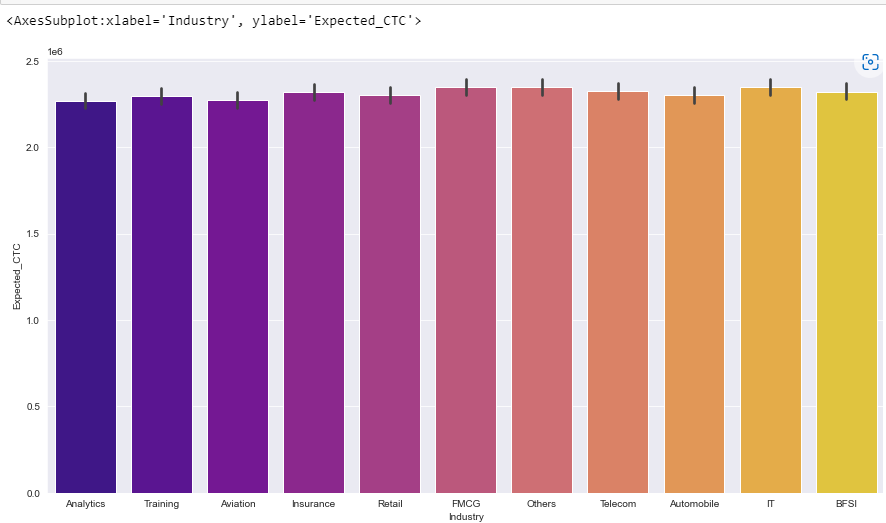
****

Fig 2.6 Bivariate Analysis

**Inference:**

* **From the categorical representation it is applicants who have 3 years of experience with other qualifications.**
* **Maximum applicants are undergraduates.**
* **The Education column with the year of passing has no outliers as such.**

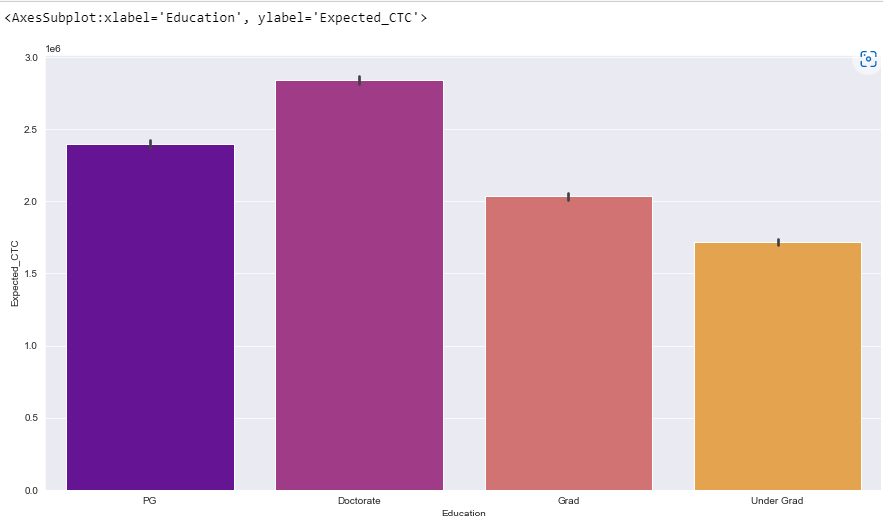
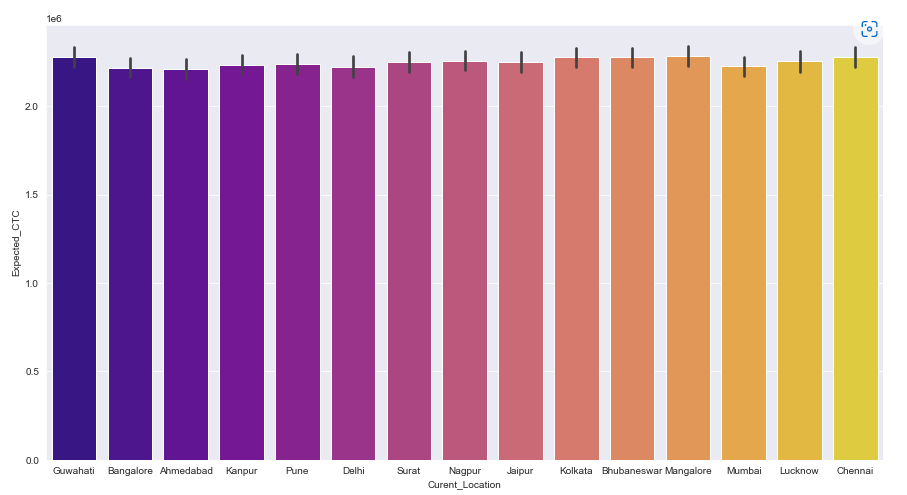
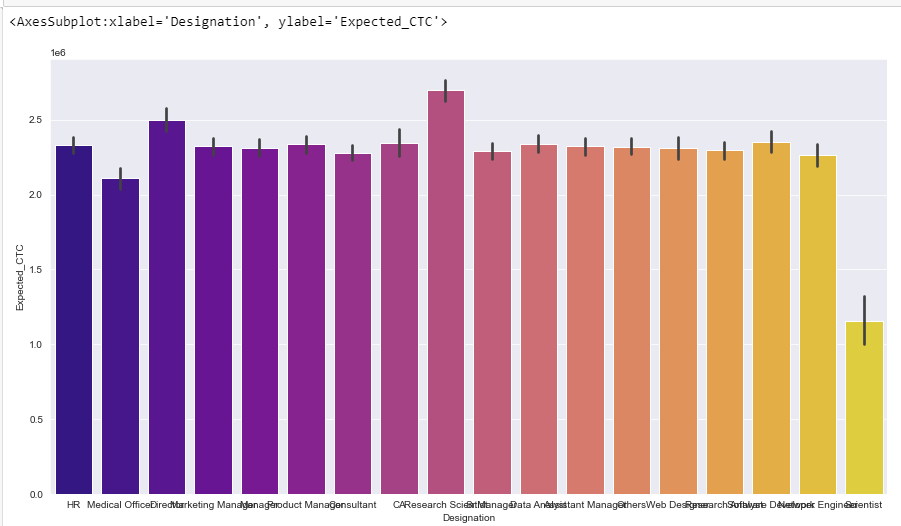
Expected\_CTC to Preferred\_location

** **

Expected\_CTC to Industry

**Inference:**

* We see that as per industry the expected salary is maximum for IT, Automobiles, FMCG and least is for Analytics.
* The location which offers more CTC is Jaipur hence it is the more preferred location. Least being Kanpur.

Expected\_CTC to Education

Expected\_CTC to Designation

Expected\_CTC to Current Location

Inference:

* The current location is mostly for Pune and Guwahati and the least is for Mumbai.
* The education qualification of doctorates earn the maximum salary.
* The designation Marketing earns the maximum followed by Software development. The least being of HR.

Fig 2.7 Bivariate Analysis for Categorical variables

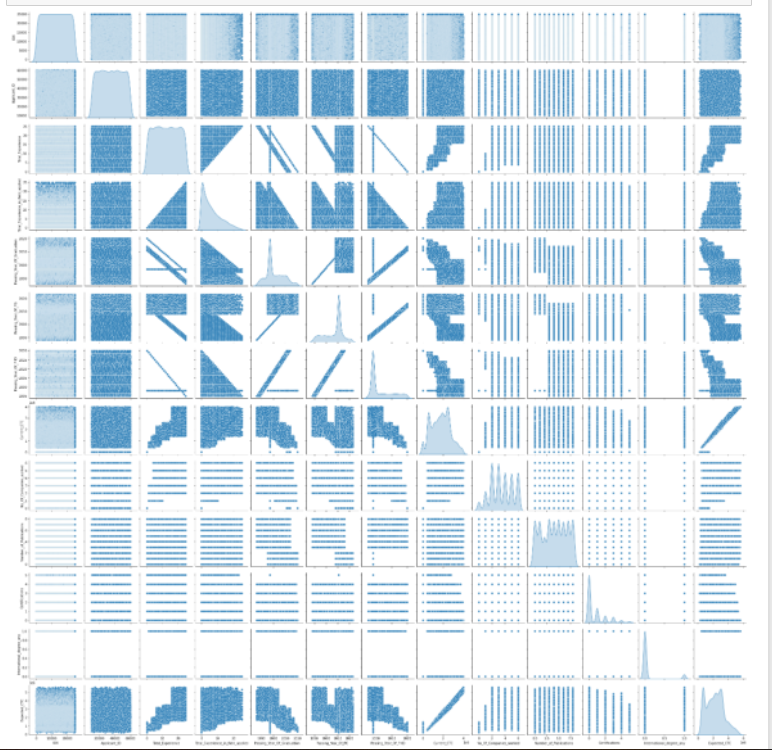
**MULTIVARIATE ANALYSIS** 

Fig 2.8 Pair Plot

Inference:

* There exists multicollinearity in data for which we will remove outliers and scale the data.

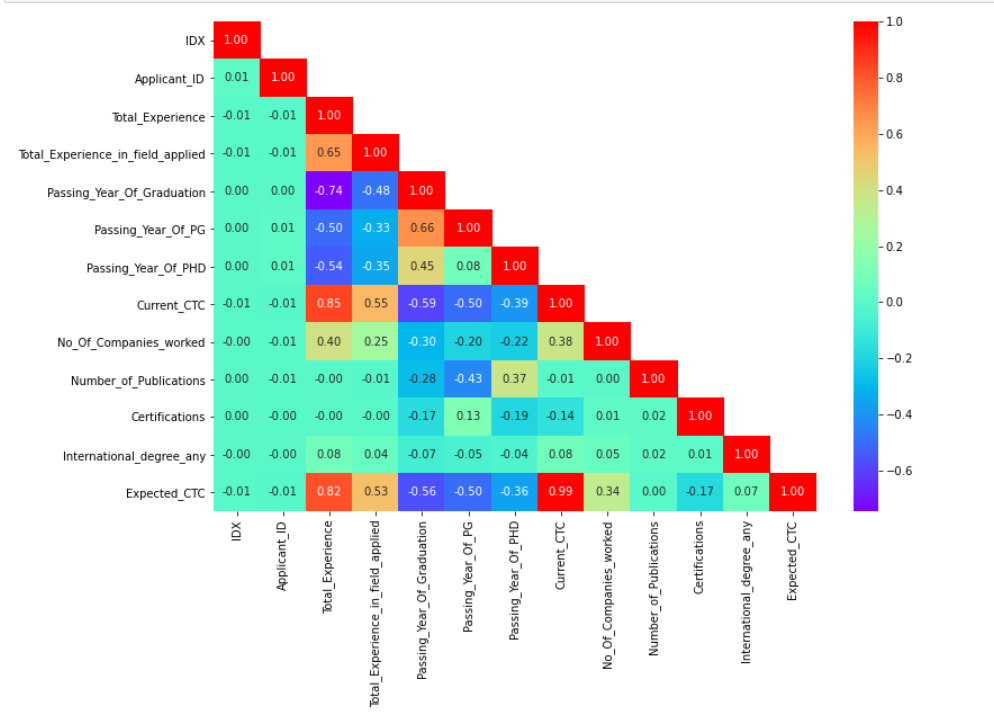
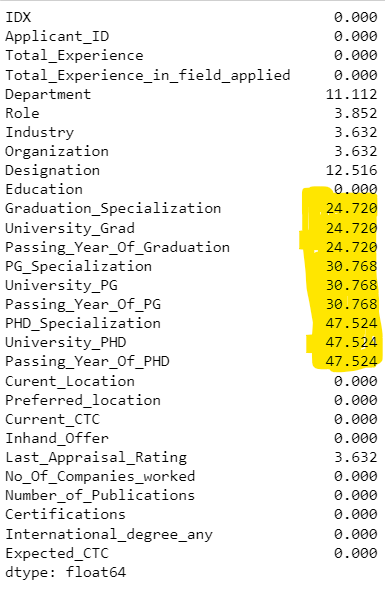


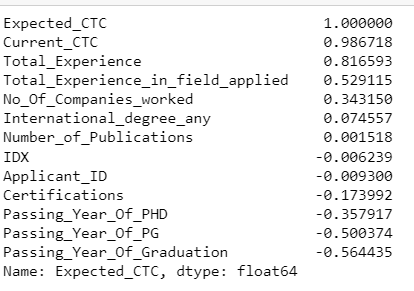
Fig 2.9 Correlation Matrix

**INFERENCE:**

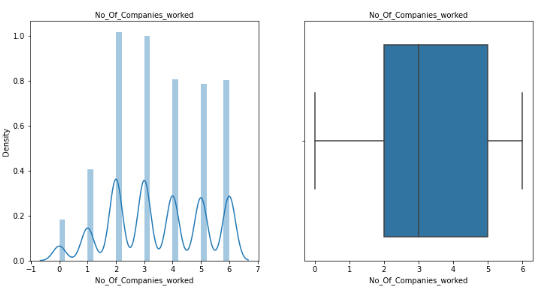
* **The data is not following normal distribution indicating missing values in the pair plot. The number of missing data contribution of each field is as:**

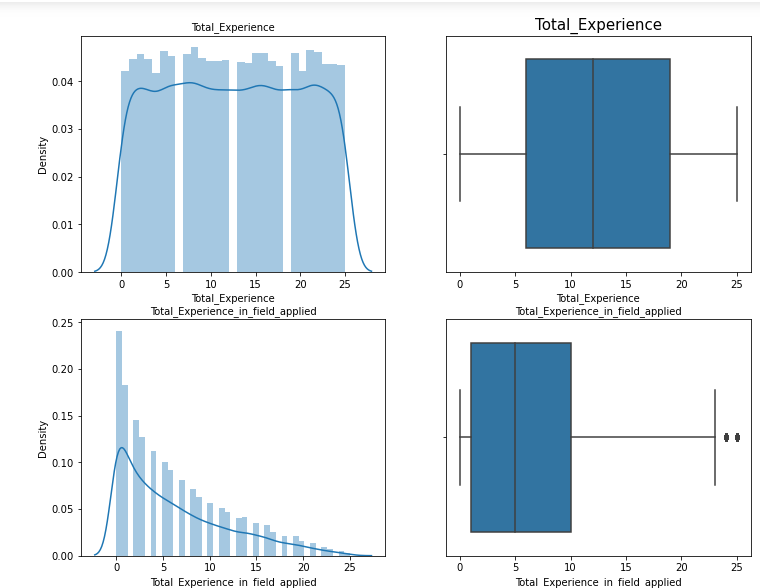
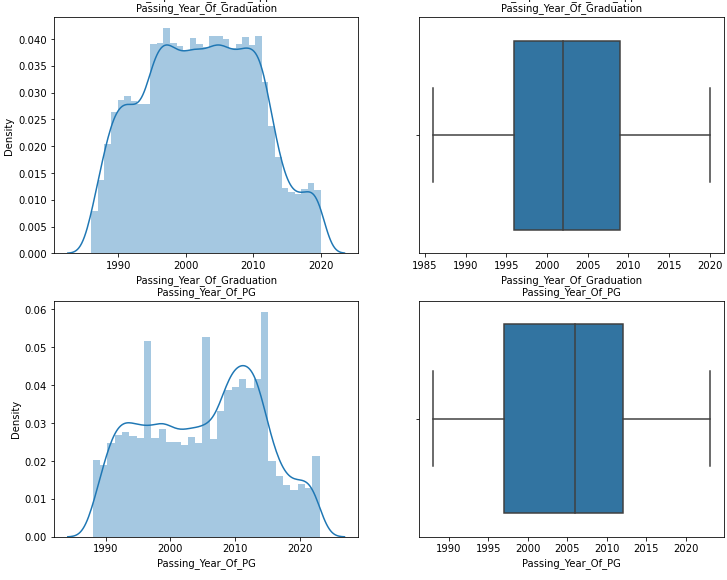


* **The data having more than 10%-15% missing values we will impute them later.**
* **Correlation between -1 to +1 are the best which is valid for fields – Expected CTC to Current CTC, Total experience to current CTC and expected CTC to Total experience.**
* **The correlations of all fields to Expected CTC is as in descending to ascending order with respect to Expected CTC our target variable:**

****

**UNIVARIATE ANALYSIS**



****

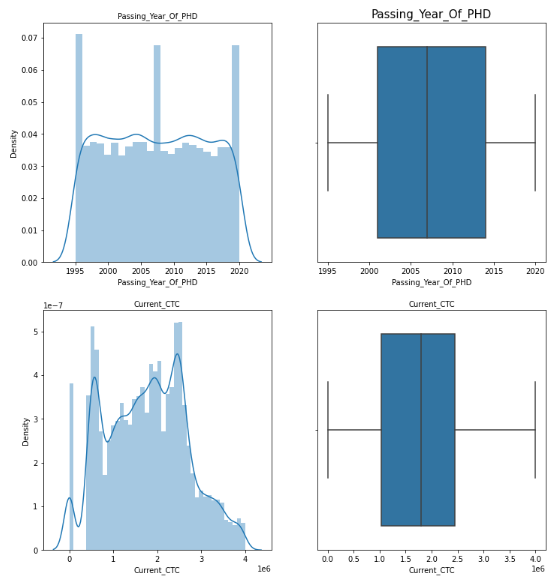
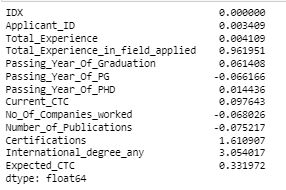


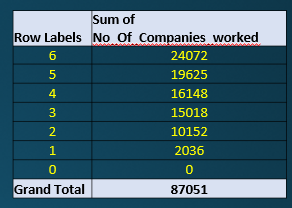
Fig 2.10 Univariate Analysis

INFERENCE:

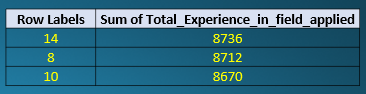
* Data seems to be normally distributed with the skewness in the fields as below:



* We have removed the outliers.
* The data can be seen in the histogram with normal distribution.
* No of companies worked with shows a lot of variations so the data in there looks like:



* Total years of experience in the field is also showing a declining trend like below because 14 is highest and 25 years which is maximum has lowest relative experience:



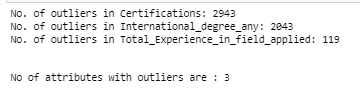
Insights from EDA

* We see Expected CTC is mostly making dependency on Predicted CTC.
* Location wise we can see Ahmedabad the top most paid and Surat the least. We see the education in top 3 earning cities is maximum in Chemistry and Economics. Existing employees in this case can be educated more in the respected EDUCATION and given a different role rather than hiring from outside at high pay scale.
* Relevant experience in the applied field is ‘ZERO’ in maximum cases so employees who aim or have upskilled themselves in the company should be given internal transfers as it will be profitable to retention plans as well as employee welfare.

**3) Data Cleaning and Pre-processing**

- Approach used for identifying and treating missing values and outlier treatment (and why):

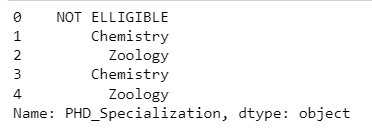
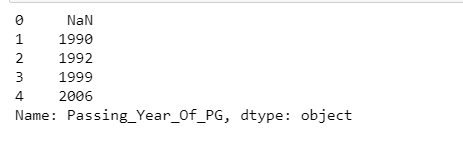
* Data has outliers in the fields:



* As in the box plots above we have identified outliers but did not treat it as missing values were filled in with median and mode. The number of outliers to total data is not making much of a difference.

- Variables removed or added and why (if any)

* People who are not a Graduate or are in Post Graduate won’t do PHD specialization hence making a special feature in the column as “NOT ELLIGIBLE”.

* Applicants who are undergrads will not have Graduation Year, Passing year of Graduation, PG Specialization, University\_ PG, Passing Year of PG or PHD Specialization, University PHD or passing year of PHD.
* We have few sequential categorical columns which needs to be replaced:

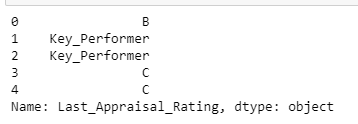
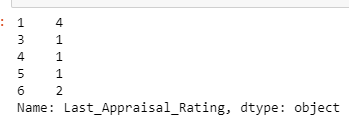
EDUCATION: Doctorate – 4

PG - 3,

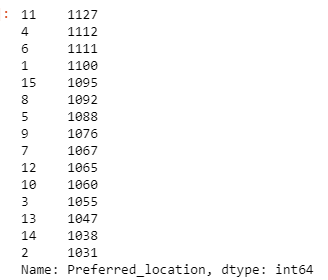
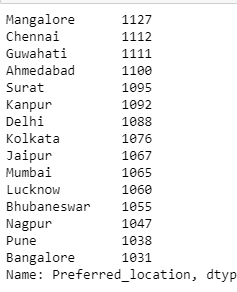
Grad – 2

Under Grad – 1

* Key Performaer is 4 and then C=1. Highest to lowest ranking as 4 to 0

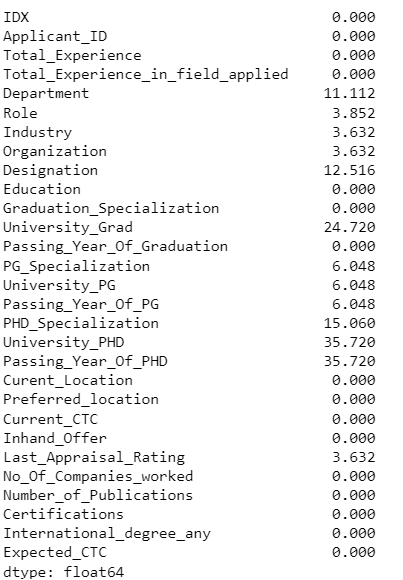
 

* **Preferred and Current location is as:**

1127 is Mangalore data where 11 is the coding so we know Mangalore is 11, Chennai is 4, Delhi is 5 so it is in alphabetical order.Where locations with A is 1 and Z would be 26.

* We can see the contribution of missing values prior to applying ‘not eligible and dropping further NA values’ is :



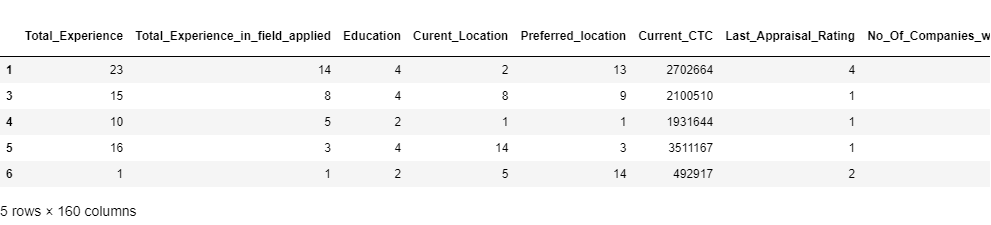
* IDX column adds no benefit hence dropping that:
* We have dropped columns like 'IDX’, 'Passing\_Year\_Of\_PG', 'Passing\_Year\_Of\_Graduation', 'Passing\_Year\_Of\_PHD’ as it won’t add any benefit to our findings as we have been given a column called ‘years of experience’ . The current rows and columns is:



- Need for variable transformation (if any)

We have then converted everything into dummy variables as, machine learning algorithms won’t understand categorical data.

Post encoding our data has 16164 rows and 161 columns.



4) Model building

Models that we have used is Regression Models since our prediction was how much salary an employee should get. Salary is a continuous variable which is the variable which we are predicting, so model should be a regression model.

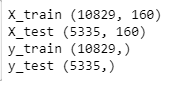
* **LINEAR REGRESSION**
* **DECISION TREE**
* **RANDOM FOREST**
* **LASSO and RIDGE**
* **GRADIENT BOOSTING**

The train and test shape of the data is as:



MODEL 1: LINEAR REGRESSION MODEL

 Llinear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.  Our train and test data looks like:

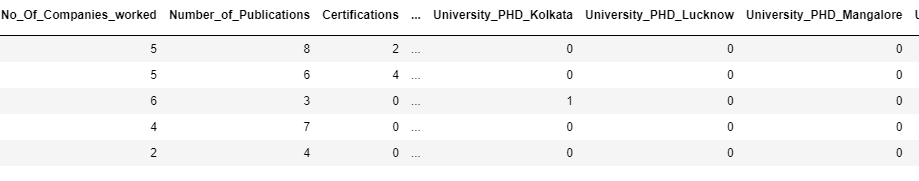
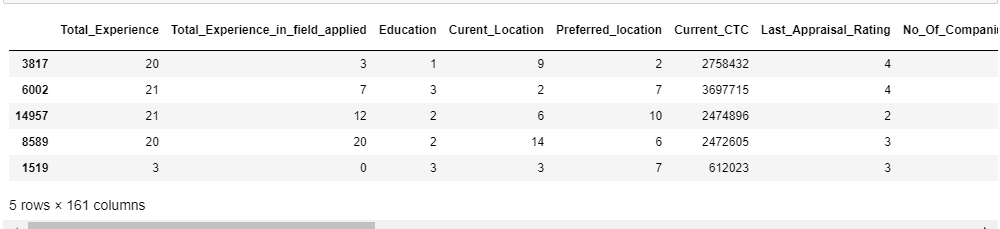


Although zero is a good intercept ours have come to :



The accuracy score of our linear regressor is:

TEST:  TRAIN : 

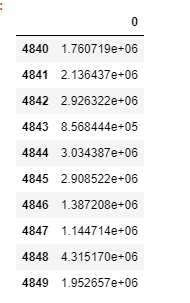
Since the data has too many categorical variables, we also tried polynomial regression and got fairly good results near to zero:



* Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors) and the mean squared error (MSE)[[1]](https://en.wikipedia.org/wiki/Mean_squared_error#cite_note-:1-1) or mean squared deviation (MSD) of an [estimator](https://en.wikipedia.org/wiki/Estimator) (of a procedure for estimating an unobserved quantity) measures the [average](https://en.wikipedia.org/wiki/Expected_value) of the squares of the [errors](https://en.wikipedia.org/wiki/Error_(statistics))—that is, the average squared difference between the estimated values and the actual value.



* Let’s see the actual and predicted results difference:

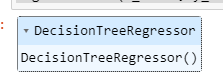
 

ACTUAL

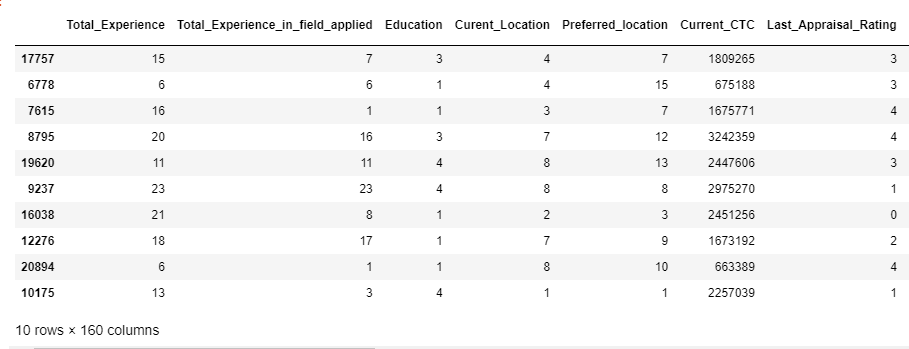
PREDICTED

There is a lot of difference between the two results which needs to be reconsidered for this model.

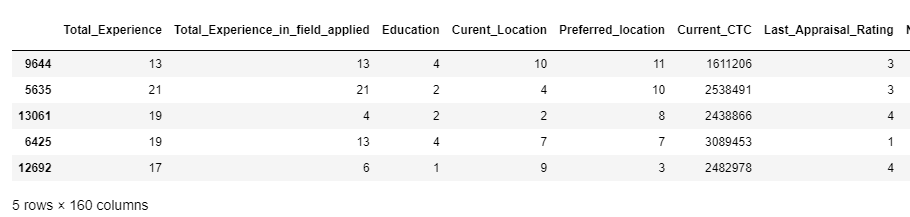
MODEL 2:DECISION TREE REGRESSOR

*  Let’s see the train and test data:

TRAIN



TEST



The accuracy score of our linear regressor is:

* TEST:  TRAIN : 

Both train and test models have a score of 99.9%

* Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors) and the mean squared error (MSE)[[1]](https://en.wikipedia.org/wiki/Mean_squared_error#cite_note-:1-1) or mean squared deviation (MSD) of an [estimator](https://en.wikipedia.org/wiki/Estimator) (of a procedure for estimating an unobserved quantity) measures the [average](https://en.wikipedia.org/wiki/Expected_value) of the squares of the [errors](https://en.wikipedia.org/wiki/Error_(statistics))—that is, the average squared difference between the estimated values and the actual value.



* Let’s see the actual and predicted results difference:

PREDICTED

ACTUAL

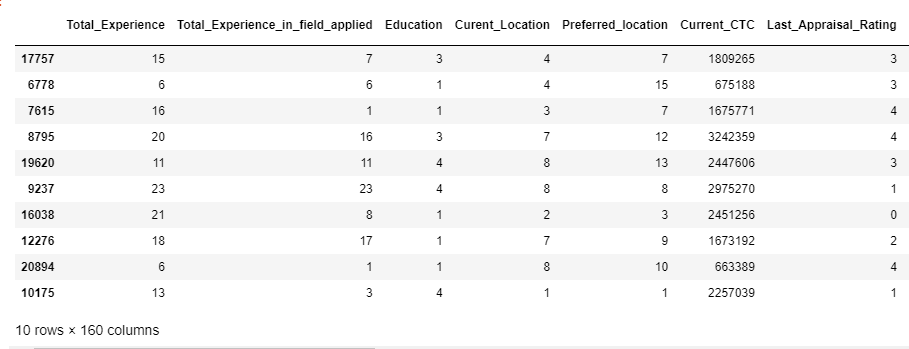
Since the salaries are in lacs the difference is valid and not at all much. Hence Decision tree here is a better model than liner regressor.

RANDOM FOREST REGRESSOR

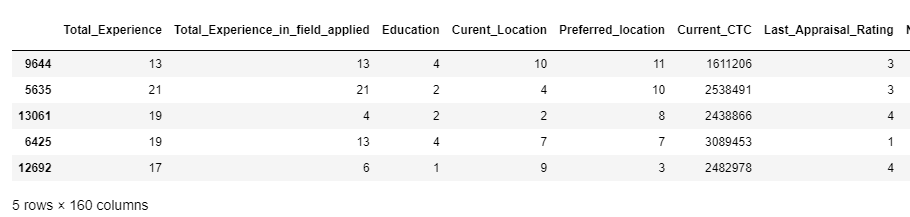
MODEL 3:



* Lets see the train and test data:



TRAIN



TEST

The accuracy score of our linear regressor is:

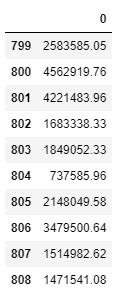
* TEST:  TRAIN : 

Both train and test models have a score of 99.9%

* Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors) and the mean squared error (MSE)[[1]](https://en.wikipedia.org/wiki/Mean_squared_error#cite_note-:1-1) or mean squared deviation (MSD) of an [estimator](https://en.wikipedia.org/wiki/Estimator) (of a procedure for estimating an unobserved quantity) measures the [average](https://en.wikipedia.org/wiki/Expected_value) of the squares of the [errors](https://en.wikipedia.org/wiki/Error_(statistics))—that is, the average squared difference between the estimated values and the actual value.



* Let’s see the actual and predicted results difference:

ACTUAL

PREDICTED

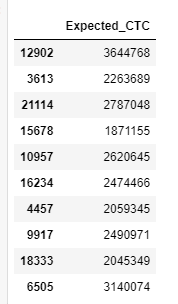
Since the salaries are in lacs the difference is valid and not at all much. Hence Random Forest here is a better model than liner regressor and decision as predictions are near buy.

MODEL 4: LASSO and RIDGE

We will take Lasso and Ridge for checking the Accuracy score and we get the below results:

Ridge Train and Test Scores: Lasso Train and Test Scores:  

Since the prediction to actual has a lot of difference the model is not an apt model:

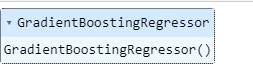


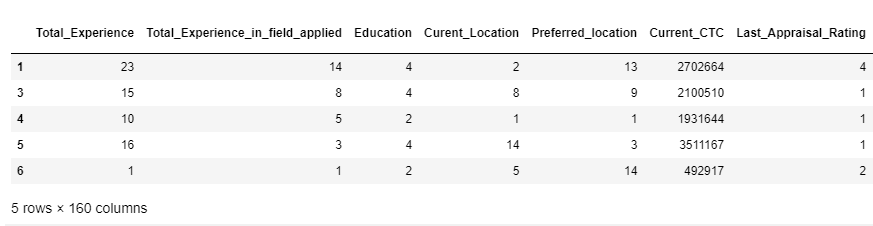
ACTUAL

PREDICTED

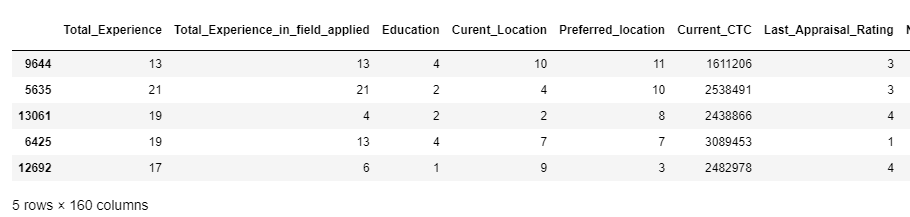
This model is not a good model.

MODEL 5: GRADIENT BOOSTING



TRAIN DATA:

TEST DATA:



The accuracy score of our linear regressor is:

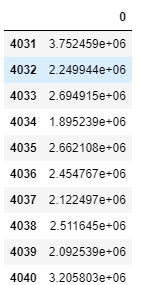
* TEST:  TRAIN: 

Both train and test models have a score of 99.9%

* Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors) and the mean squared error (MSE)[[1]](https://en.wikipedia.org/wiki/Mean_squared_error#cite_note-:1-1) or mean squared deviation (MSD) of an [estimator](https://en.wikipedia.org/wiki/Estimator) (of a procedure for estimating an unobserved quantity) measures the [average](https://en.wikipedia.org/wiki/Expected_value) of the squares of the [errors](https://en.wikipedia.org/wiki/Error_(statistics))—that is, the average squared difference between the estimated values and the actual value.



* Let’s see the actual and predicted results difference:

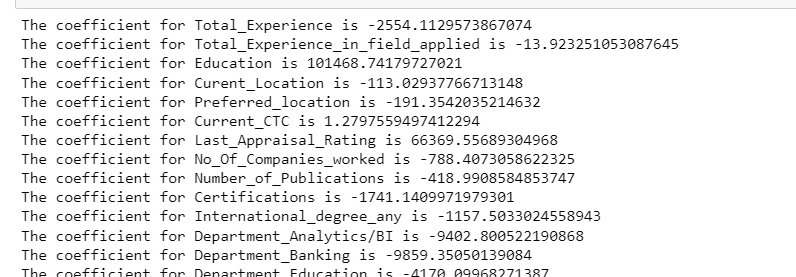
ACTUAL

PREDICTED

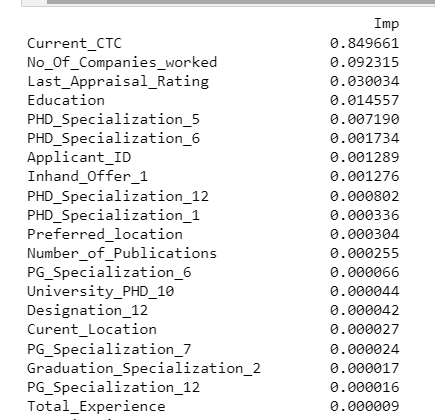
Not a great model than Random and Decision Tree as we see that the numbers do vary a lot.

- Effort to improve model performance.

**Since we see model is more dependant on Current CTC and last appraisal rating so we perform linear regression by dropping these two and check our accuracy.**



* The feature of importance gave us results as below. This shows Expected CTC will most depend on current\_ctc then on companies worked with followed by last appraisal rating and so on:

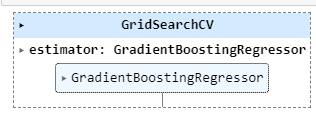


DECISION TREE we have performed and got below accuracy score. OUR MODEL IS A ROBUST MODEL WITH A SCORE OF 77.5%

TEST -  TRAIN - 

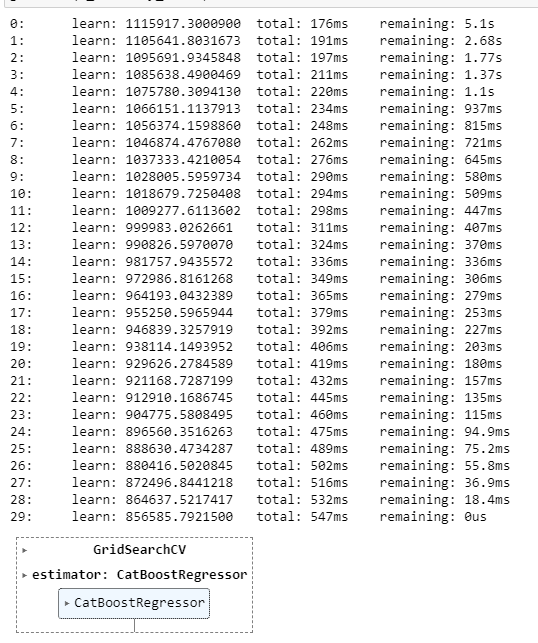
**5) Model validation**

Yes, we have used Model tuning as hyperparameter optimization. Hyperparameters are variables that control the training process. These are configuration variables that do not change during a model training job. Model tuning provides optimized values for hyperparameters, which maximize your model's predictive accuracy.

We have done model tuning with 

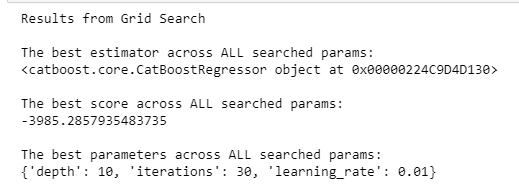
The training and Testing data shape rows is to columns that has been considered is:



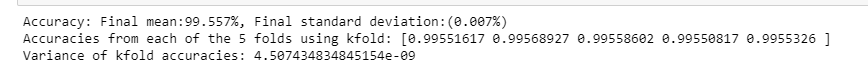
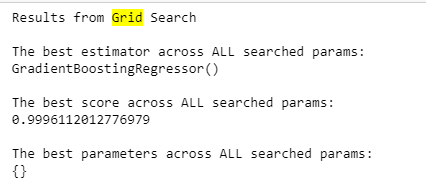
The learning rate taken into consideration for the model is as:

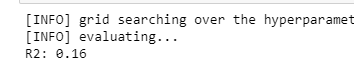
The best fit for the model will be at learning rate 0.01:

The other best fit applications for the model are:



The final accuracy of the tuned model is Grid search has been applied:





The model has an accuracy of 99.57% with a R2 of 0.16 which will be our best model. Since R2 is higher we can conclude that model has good corelation between parameters during tuning.

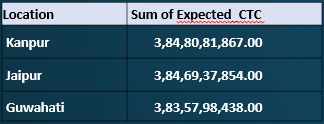
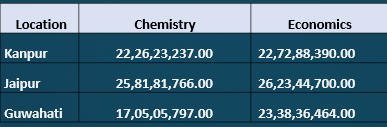
Interpretation of the most optimum model and its implication on the business

MODELS WITH SCORES:

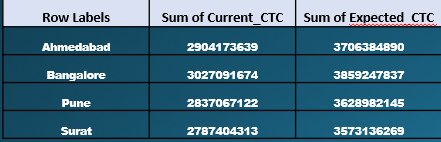
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Train Score | Test Score | MAPE | RMSE | MSE |
| LINEAR REGRESSION | 0.996 | 0.995 | 0.035 | 73,942.57 | 5,46,75,04,339.81 |
| DECISION TREE | 0.9997 | 0.9995 | 0.810 | 4,698.76 | 2,20,78,320.12 |
| RANDOM FOREST | 0.9999 | 0.9999 | 0.8153 | 5,429.01 | 2,94,74,159.18 |
| GRADIENT BOOSTER REGRESSOR | 0.9997 | 0.9996 | 0.034 | 74,143.02 | 5,49,71,87,564.10 |

* Decision tree is the best model of all since the lower the MSE the better is the model.
* However, the MAPE states Gradient Booster as the best model since 20%-50% is better than 80%.
* Also, the accuracy score is more for Gradient booster to that of Decision tree.
* Since our base model has also given accuracy of 99.7%, we go ahead with DECISION TREE although Gradient Booster is a good model. Linear is chosen because it is easy to apply whereas Gradient Boosting uses derivates which can complicate model later.
* Location wise we can see Ahmedabad the top most paid and Surat the least. We see the education in top 3 earning cities is maximum in Chemistry and Economics. Existing employees in this case can be educated more in the respected EDUCATION and given a different role rather than hiring from outside at high pay scale.

6) Final Interpretation / Recommendation

* Since we know highest dependency of Expected CTC is on Current CTC, we need to also decide a percentage as per that. Many employees leave companies once the appraisal is done so a bar needs to be fixed for such employees to 20% or 25% as profitable to company.
* From EDA we see ID#27518 having the maximum CTC. He is a Doctorate and PHD in Organisation A from PUNE.
* Relevant experience in the applied field is ‘ZERO’ in maximum cases so employees who aim or have upskilled themselves in the company should be given internal transfers as it will be profitable to retention plans as well as employee welfare.
* Education should have more data and more importance as wrong man at the wrong place can mess up a process.
* KEY PERFORMERS are the highest paid CTC if we see from EDA and all are Doctorate, Sales being maximum paid.
* Passing years of 1997 and 1990 are getting paid more so we can say doctorates with experience of 8-18 years are being paid the most.
* Location wise we can see Ahmedabad the top most paid and Surat the least. More investigation needs to be done as to why people from Surat are less paid. If they are well qualified and educated hiring from there would be beneficial to company, as good resource at low cost would be received.



* For PHD\_specialization we again see Post Graduates not eligible for PHD and null columns have maximum salaries. We have listed the top six.

1)anyone with Post Graduate is not Elligible for PHD

2)anyone doing under graduate is not applicable for PHD

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Sum of Current\_CTC** | **Sum of Expected\_CTC** |
| NOT APPLICABLE | 8586164272 | 10615927964 |
| (blank) | 7124359270 | 9114879409 |
| NOT ELLIGIBLE | 3631319721 | 4626091219 |
| Others | 3011632385 | 3842847540 |

