Tejesh. K

Capstone project

Hr Data

x

**Table of content**

OVERVIEW OF PROBLEM…………………………………..2

PROBLEM STATEMENT…………………………………………2

INTRODUCTION TO BUSINESS INSIGHTS………………………..3

DEFINING PROBLEM STATEMENT………………………….3

NEED OF STUDY/PROJECT…………………………………..3

UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY………………….3

DATA REPORT………………………………………………4

DATA COLLECTION…………………………………………..4

FUTURE OF DATASET ………………………………………4

FUTURES DETAILS……………………………………………..4

EXPLORATORY DATA ANALYSIS…………………………….10

DATA DESCRIPTION……………………………………………10

ADDITION OF NEW VARIABLES…………………………………11

UNIVARIATE ANALYSIS ………………………………………….12

BIVARIATE ANALYSIS………………………………………………..15

MULTIVARIATE ANALYSIS………………………………………..

MODEL BUILDING……………………………………………28

MODEL VALIDATING………………………………………28

FINAL INTERPECTAION AND RECOMMENDATION…………35

**OVERVIEW OF THE PROBLEM**

**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 judgement about salary to be offered.

**GOALS AND OBJECTIVE**

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 judgements on selections are minimized. It is intended to have a robust approach and eliminate any discrimination in salary among similar employee profiles.

Dataset used: HR Data Capstone Project.

**INTRODUCTION TO THE BUSINESS PROBLEM**

**DEFINING PROBLEM STATEMENT**

The dataset consists of details of applicants who applied for various jobs under multiple departments in the company Delta Ltd. The applicants have given details pertaining to their Total Experience, Relevant Experience, Departments they belong to, Roles they have handled, Industry, Achievements including number of publications and certifications, International Degrees, Performance history, companies worked for, if the applicants currently have an offer, current Location, Preferred Location, Current CTC, and Expected CTC.

**NEED OF THE STUDY/PROJECT**

We required to analyze the data given by applicants and based on the analysis we are expected to determine the employee’s salary to be offered. The problem is a regression problem we are requested to use various regression models to predict a salary for any candidate. We are required to transform the variables appropriately to use it effectively for the regression model.

**UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY**

Removing the manual retaining in deciding salary for prospective candidates removes all forms of bias and errors pretraining to decide the CTC of an employee. This result can improve in:

* Improving employee’s satisfaction.
* Improves the employee’s possession.
* Improves parity in the organization across departments, designation hierarchies and levels.
* Helps in bringing about a sense of fairness in the job.

However, it has few cons.

* Eliminates the opportunity of hiring a worker with greater salary employees.
* Raises the consistent need for conducting market benchmark.
* Depends on accuracy and precision model development.

**DATA REPORT**

**DATA COLLECTION**

The data was collected randomly, probably through an Applicant Tracking System. Many of columns are not mandatory and depend on data availability. Most of the columns are having missing value because candidate are Freshers. The data dump attempts to collect data that might help in predicating a candidate’s salary if he/she is given an offer.

**FEATURES IN DATASET**

Table 1: Features in the Dataset

|  |  |  |
| --- | --- | --- |
| Total\_Experience | Total\_Experience\_in\_field\_applied | Department |
| Role | Industry | Organization |
| Designation | Education | Graduation\_Specialization |
| University\_Grad | Passing\_Year\_of\_Graduation | PG\_Specialization |
| University\_PG | Passing\_Year\_Of\_PG | PHD\_Specialization |
| University\_PHD | Passing\_Year\_Of\_PHD | Current\_Location |
| Preferred\_Location | Current\_CTC | Inhand\_Offer |
| Last\_Appraisal\_Rating | No\_of\_Companies\_Worked | Number\_of\_Publications |
| Certifications | International\_Degree\_Any | Expected\_CTC |

* There are 25000 rows and 29 features in dataset.
* There are 13 numerical features and 16 Categorical features.
* They are missing variables in Categorical dataset.
* Total missing values in the dataset 86853.
* There are 94038 unique values in the dataset across all the datasets.

Table 2: Features and Categories in dataset

|  |  |
| --- | --- |
| **Features** | **No. of Categories** |
| Department | 12 |
| Role | 24 |
| Industry | 11 |
| Organization | 16 |
| Designation | 18 |
| Education | 4 |
| Graduation\_Specialization | 11 |
| University\_Grad | 13 |
| PG\_Specialization | 11 |
| University\_PG | 13 |
| PHD\_Specialization | 11 |
| University\_PHD | 13 |
| Curent\_Location | 15 |
| Preferred\_Location | 15 |
| Inhand\_Offer | 2 |
| Last\_Apprasial\_Rating | 5 |

Table 3:Features and numbers of values in each categories

|  |  |  |  |
| --- | --- | --- | --- |
| **Department** | **Category** | **Industry** | Category |
| Marketing | 2379 | Training | 2237 |
| Analytics/BI | 2096 | IT | 2228 |
| Healthcare | 2062 | Insurance | 2219 |
| Others | 2041 | BFSI | 2207 |
| Sales | 1991 | Automobile | 2202 |
| HR | 1988 | Analytics | 2201 |
| Banking | 1952 | Retail | 2195 |
| Education | 1948 | Telecom | 2190 |
| Engineering | 1937 | Aviation | 2183 |
| Top Management | 1632 | FMCG | 2180 |
| Accounts | 1118 | Others | 2050 |
| IT-Software | 1078 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Designation** | **No.of Values** | | **Organization** | | **No.of Values** |
| HR | 1648 | | M | | 1574 |
| Others | 1647 | | J | | 1555 |
| Manager | 1628 | | P | | 1542 |
| Product Manager | 1626 | | H | | 1532 |
| Sr. Manager | 1617 | | A | | 1526 |
| Consultant | 1606 | | F | | 1505 |
| Marketing Manager | 1590 | | G | | 1504 |
| Assistant Manager | 1590 | | K | | 1503 |
| Data Analyst | 1575 | | I | | 1489 |
| Research Analyst | 1563 | | E | | 1488 |
| Medical Officer | 1047 | | B | | 1488 |
| Software Developer | 914 | | L | | 1484 |
| Web Designer | 882 | | C | | 1482 |
| Network Engineer | 862 | | N | | 1476 |
| Director | 772 | | D | | 1474 |
| CA | 715 | | O | | 1470 |
| Research Scientist | 537 | |
| Scientist | 52 | |
| **Education** | **No. of Values** | **Inhand\_Offer** | **No. of Values** | **Last\_Apprasial\_Rating** | **No. of Values** |
| PG | 6326 | N | 17418 | B | 5501 |
| Doctorate | 6285 | Y | 7582 | D | 4917 |
| Grad | 6209 |  |  | C | 4812 |
| Under Grad | 6180 |  |  | A | 4671 |
|  |  |  |  | Key\_Performer | 4191 |

|  |  |  |  |
| --- | --- | --- | --- |
| **PG\_Specialization** | **No.of Values** | **University\_PG** | **No. of Values** |
| Mathematics | 1800 | Bhubaneswar | 1377 |
| Chemistry | 1796 | Delhi | 1368 |
| Economics | 1755 | Mangalore | 1367 |
| Engineering | 1674 | Mumbai | 1366 |
| Statistics | 1639 | Jaipur | 1359 |
| Others | 1629 | Guwahati | 1340 |
| Psychology | 1425 | Surat | 1329 |
| Zoology | 1424 | Lucknow | 1328 |
| Arts | 1410 | Pune | 1314 |
| Sociology | 1385 | Nagpur | 1313 |
| Botony | 1371 | Kolkata | 1306 |
|  | | Bangalore | 1287 |
| Ahmedabad | 1254 |
| **PHD\_Specialization** | **No. of Values** | **University\_PHD** | **No. Of Values** |
| Others | 1545 | Kolkata | 1069 |
| Chemistry | 1458 | Delhi | 1064 |
| Mathematics | 1378 | Mumbai | 1046 |
| Economics | 1343 | Guwahati | 1030 |
| Engineering | 1259 | Pune | 1011 |
| Statistics | 1236 | Surat | 1009 |
| Zoology | 1011 | Jaipur | 998 |
| Sociology | 989 | Lucknow | 995 |
| Psychology | 986 | Bangalore | 992 |
| Botony | 976 | Bhubaneswar | 988 |
| Arts | 938 | Mangalore | 964 |
|  | | Nagpur | 964 |
| Ahmedabad | 959 |
| **Curent\_Location** | **No. of values** | **Preferred\_Location** | **No. Of Values** |
| Bangalore | 1742 | Kanpur | 1720 |
| Jaipur | 1706 | Ahmedabad | 1715 |
| Bhubaneswar | 1704 | Guwahati | 1695 |
| Mangalore | 1697 | Mangalore | 1694 |
| Delhi | 1680 | Surat | 1693 |
| Ahmedabad | 1677 | Delhi | 1683 |
| Guwahati | 1672 | Chennai | 1680 |
| Chennai | 1669 | Kolkata | 1669 |
| Kanpur | 1664 | Jaipur | 1659 |
| Nagpur | 1663 | Pune | 1654 |
| Mumbai | 1658 | Bhubaneswar | 1653 |
| Lucknow | 1637 | Nagpur | 1650 |
| Pune | 1622 | Mumbai | 1617 |
| Kolkata | 1620 | Lucknow | 1612 |
| Surat | 1589 | Bangalore | 1606 |

|  |  |  |  |
| --- | --- | --- | --- |
| Graduation\_Specialization | No. of Values | University\_Grad | No. of Values |
| Chemistry | 1785 | Bhubaneswar | 1510 |
| Economics | 1774 | Delhi | 1492 |
| Mathematics | 1770 | Mangalore | 1490 |
| Zoology | 1730 | Mumbai | 1488 |
| Arts | 1721 | Jaipur | 1478 |
| Psychology | 1705 | Lucknow | 1457 |
| Sociology | 1697 | Guwahati | 1449 |
| Botony | 1674 | Pune | 1428 |
| Engineering | 1661 | Kolkata | 1426 |
| Others | 1660 | Surat | 1424 |
| Statistics | 1643 | Nagpur | 1420 |
|  | | Bangalore | 1394 |
| Ahmedabad | 1364 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Roles** | **No. of Values** | **Roles** | **No. of Values** |
| Others | 2248 | Bio Statistician | 1913 |
| Analyst | 1892 | Project Manager | 1850 |
| Team Lead | 1833 | Consultant | 1780 |
| Business Analyst | 1711 | Sales Executive | 1574 |
| Sales Manager | 1427 | Senior Research | 1236 |
| Financial Analyst | 1182 | CEO | 1149 |
| Scientist | 1139 | Head | 1108 |
| Associate | 767 | Data Scientist | 363 |
| Principal Analyst | 275 | Area Sales Manager | 134 |
| Senior Analyst | 128 | Researcher | 123 |
| Sr. Business Analyst | 114 | Professor | 33 |
| Research Scientist | 33 | Lab Executive | 25 |

The category under each of the feature and number of values shown in above tables. The essential to be consider if we decide to use one encoding for each of features then the dataset will become large, and we have reduced the dataset similar category given.

**EXPLORATORY DATA ANALYSIS**

**DATA DESCRIPTION**

Table 4: Data Description

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Numerical Feature** | **Counts** | **Mean** | **STD** | **MIN** | **25%** | **50%** | **75%** | **MAX** |
| Total Experience | 25000.0 | 1.249308e  +01 | 7.471398e  +00 | 0.0 | 6 | 12.0 | 1.9 | 25 |
| Total\_  Experience\_  In\_field  \_applied | 25000.0 | 6.258200e  +00 | 5.819513e  +00 | 0.0 | 1 | 5 | 1 | 25 |
| Current\_CTC | 25000.0 | 1.760945e  +06 | 9.202125 e+05 | 0.0 | 1.02 | 18025  67.5 | 2.44 | 3999  693.0 |
| No\_Of\_  Companies  \_Worked | 25000.0 | 3.482040e  +00 | 1.690335e+  00 | 0.0 | 2.0 | 3.0 | 5.0 | 6.0 |
| Number\_of\_  Publications | 25000.0 | 4.089040e  +  00 | 2.606612e+  00 | 0.0 | 2.0 | 4.0 | 6.0 | 8.0 |
| Certifications | 25000.0 | 7.736800e-  01 | 1.199449e  +00 | 0.0 | 0.0 | 0.0 | 1.0 | 5.0 |
| International\_  Degree\_  Any | 25000.0 | 8.172000e-  02 | 2.739431e-  01 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| Expected\_  CTC | 25000.0 | 2.2505155e+  06 | 1.160480e+  06 | 2037  44.0 | 1.360  6278  e+06 | 2252  136.5 | 3.051  354e+  06 | 5599  570.0 |
| Expected\_  hike | 25000.0 | 2.8758949  e+01 | 1.565230  e+01 | 14.999763 | 1.959999e+01 | 25.0 | 3.199996e+01 | 100.0 |

None of them shows any anomaly or show any many outliers in the dataset.

**REMOVING MISSING VALUES**

There are 86583 missing values in the dataset present. The missing values are present are shown below:

Table 5: Missing Values count shown in each feature

|  |  |  |  |
| --- | --- | --- | --- |
| **FEATURES** | **MISSING VALUES** | **FEATURES** | **MISSING VALUES** |
| PHD\_Specialization | 11881 | University\_Grad | 6180 |
| University\_PHD | 11881 | Passing\_Year\_of\_  Graduation | 6180 |
| Passing\_Year\_Of\_  PHD | 11881 | Designation | 3129 |
| University\_PG | 7692 | Department | 2778 |
| Passing\_Year\_Of\_  PG | 7692 | Role | 963 |
| PG\_Specialization | 7692 | Industry | 908 |
| Graduation\_Specialization | 6180 | Organization | 908 |
| Last\_Appraisal\_Rating | 908 |

Before we used to replace the values and then by imputing replace values we can fill the all the values in the dataset using integers in the dataset.

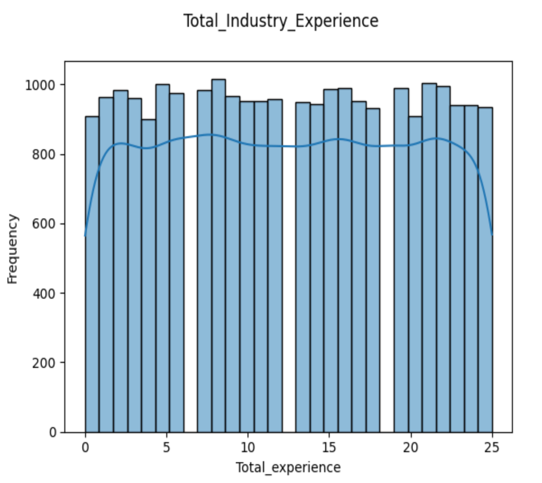
By using mean we can fill all missing values in the dataset.

**ADDITION OF NEW VARIABLES**

We have added new variables with the limit range of expected CTC with 100%.

Expected\_hike=[(Expected\_CTC-Current\_CTC/Current\_CTC) \*100%].

**UNIVARIATE ANALYSIS**

A picture containing text, screenshot, line, diagram

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, font, line

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, diagram, line

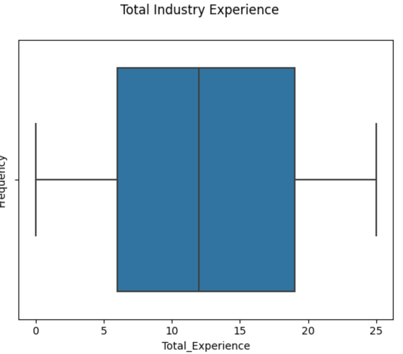
Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, line, diagram

Description automatically generated

Inference from the Histogram:

1. Total Experience is almost equally distributed across experience level.
2. Majority of applicant does not relevant experience in the applied field.
3. The Histogram showing distribution of applicant across currently earning CTC show dual peak 0.5 and 2.5.
4. Majority of applicants worked in 2 or 3 companies.
5. The number of Publications is also equally distributed among the applicant 0 publications.
6. Certifications are not available for majority applicants.
7. A huge majority of applicants do not have international degree.
8. Expected CTC distribution is left skewed in nature which suggest majority of expected CTC are towards lower range.
9. Applicants are expecting hike with 15-25%.

A picture containing text, screenshot, rectangle, diagram

Description automatically generatedA picture containing screenshot, text, diagram, rectangle

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generatedA picture containing screenshot, rectangle, diagram, line

Description automatically generatedA diagram of a number of relevant certifications

Description automatically generated with low confidence

A picture containing text, screenshot, rectangle, line

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generatedA picture containing screenshot, diagram, rectangle, line

Description automatically generated

Inference of BOX plot:

1. All box plot suggest hard outlier in the present dataset.
2. As shown in the histogram, the total experience across all levels is almost evenly distributed.
3. Like above point, relevant experience in the field applied average lower across levels.
4. Current CTC and Expected CTC is also evenly distributed across/levels.
5. Expected hikes range between 15 and 30%.
6. The boxplot of number of companies worked for lower salary, however the maximum is higher are 5.

**BIVARIATE ANALYSIS**

Checking the correlation between Expected hike with Tenure, Department, Designation, Industry, Role, and Education.

A picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, colorfulness, plot

Description automatically generated

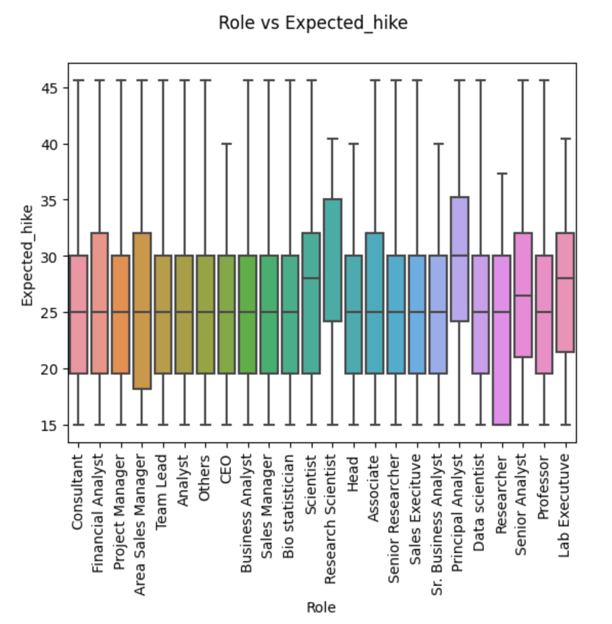
* We observe that a maximum of 25 year in present in the total Experience category.
* Applicants with experience between 4 and 8 years have higher salary expectations.

A picture containing text, diagram, line, plot

Description automatically generatedA picture containing text, screenshot, colorfulness, line

Description automatically generated

* Expectations by applicants from the IT Software department are higher than other department.

A picture containing text, screenshot, colorfulness, writing implement

Description automatically generated

* Research Scientist and Analysts are expecting higher hikes whereas a normal researcher’s expectations are lower than the average salary.
* Most in demand Role can be considered a research scientist and principal analyst.

A picture containing text, diagram, line, colorfulness

Description automatically generatedA picture containing text, screenshot, colorfulness, line

Description automatically generated

* Across Industry the expectations are equally distributed.
* Applicants in the automobile sector have largest variation in expectation.

A picture containing text, diagram, screenshot, rectangle

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

Description automatically generated

* Applicants who have a PG and Doctorate have higher salary expectations.

A picture containing text, screenshot, diagram, rectangle

Description automatically generatedA picture containing text, screenshot, rectangle, diagram

Description automatically generated

* Applicants who have an offer in hand have higher salary expectations.
* Expectations are irrespective of number of companies applicants have worked

A picture containing screenshot, text, diagram, rectangle

Description automatically generatedA picture containing text, screenshot, diagram, rectangle

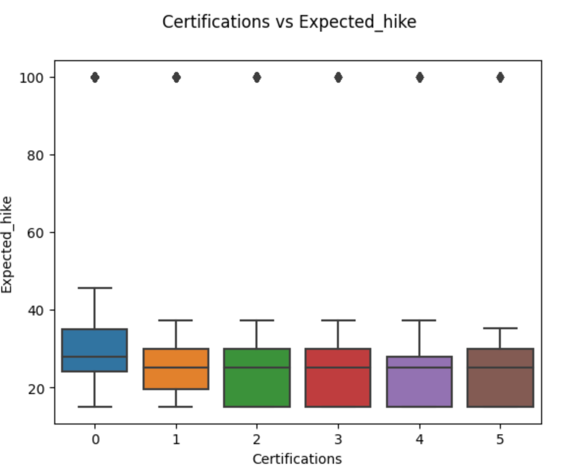
Description automatically generated

* Applicants with an International degree expected higher salary but not by a large extent.
* Applicants who were key performers expect a higher hike.

A picture containing text, diagram, screenshot, rectangle

Description automatically generatedA picture containing text, screenshot, colorfulness, plot

Description automatically generated

A picture containing text, screenshot, rectangle, colorfulness

Description automatically generated

* Certifications do not have a major effect on salary expectations.
* Number of Publications hardly influence salary expectations of applicants.

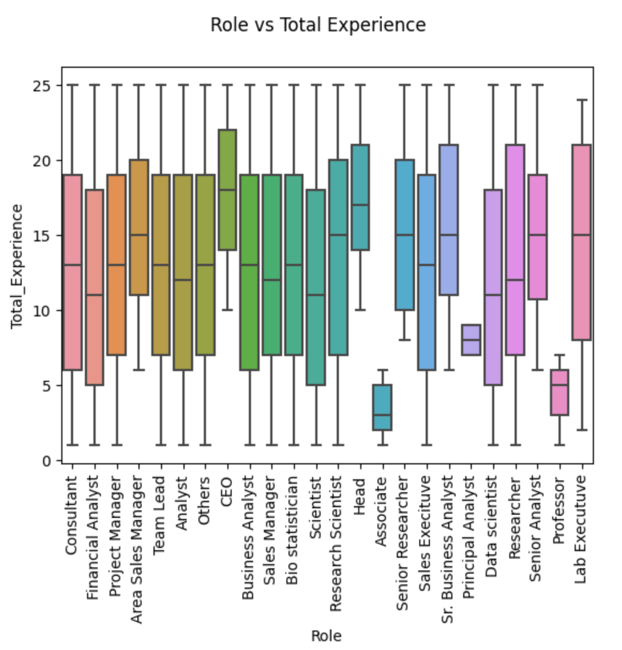
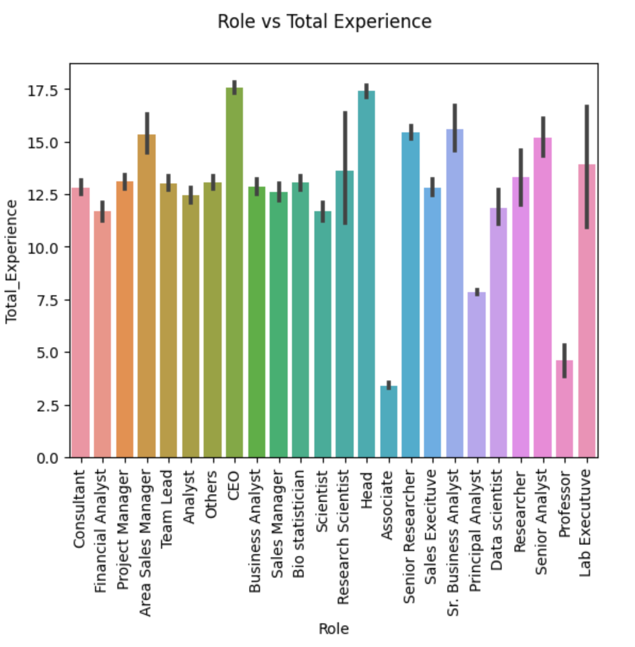
**Checking the correlation between Experience with Department, Designations, Industry, Role, and Educations.**

A picture containing text, diagram, line, plot

Description automatically generatedA picture containing text, screenshot, colorfulness, plot

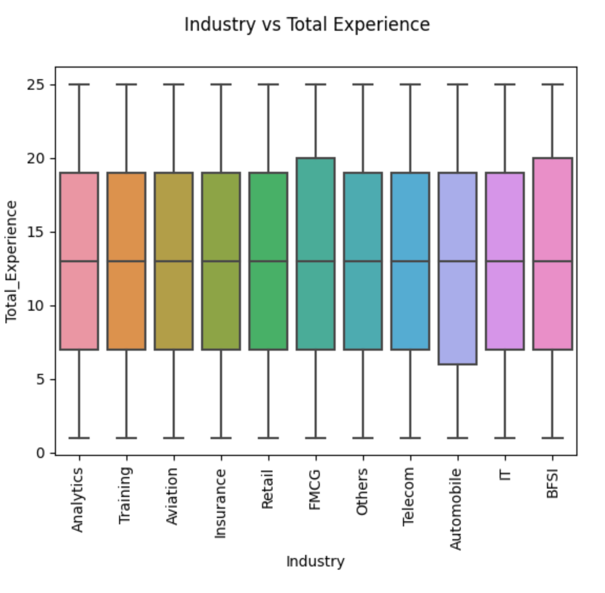
Description automatically generated

* It is obvious that the applicant in the top management has higher experience.

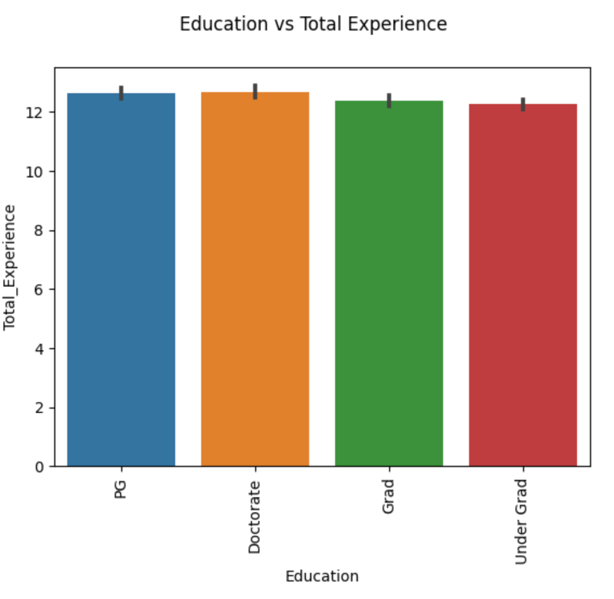


* Applicants who are Associate or Professor have least experience which show that these are entry level jobs.
* Moreover, Principal analyst seems to be a new job role introduced.

A picture containing text, screenshot, colorfulness, line

Description automatically generated

* Experience across industries are almost equally distributed with a little more than average in the FMCG and BFSI sectors.
* A picture containing diagram, rectangle, square, screenshot

  Description automatically generated
* There are hardly any difference in the distribution of experience and education levels.

**Count Plot of CURRENT LOCATION and Preferred Location.**

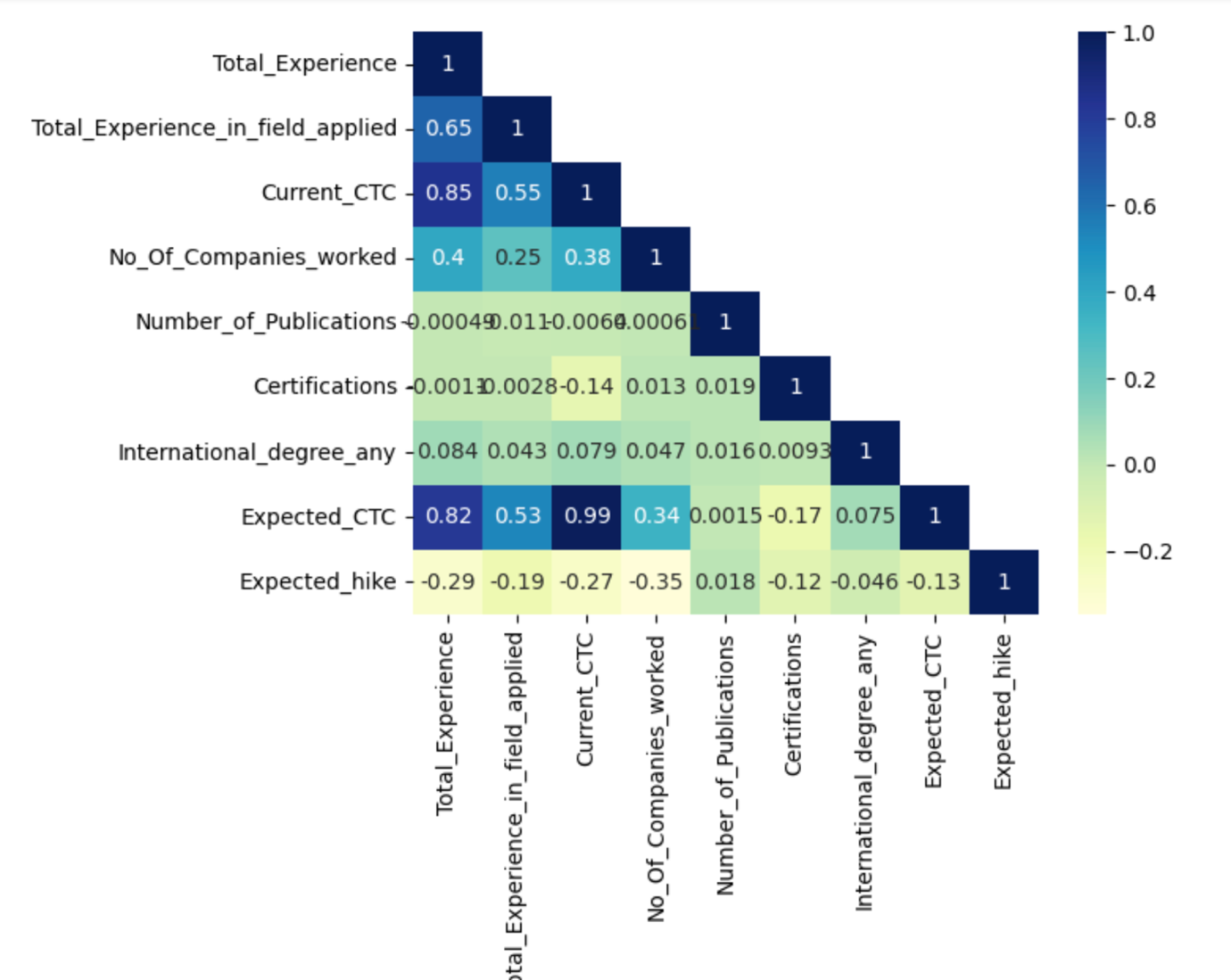
A picture containing text, screenshot, colorfulness, plot

Description automatically generatedA picture containing text, screenshot, colorfulness, line

Description automatically generated

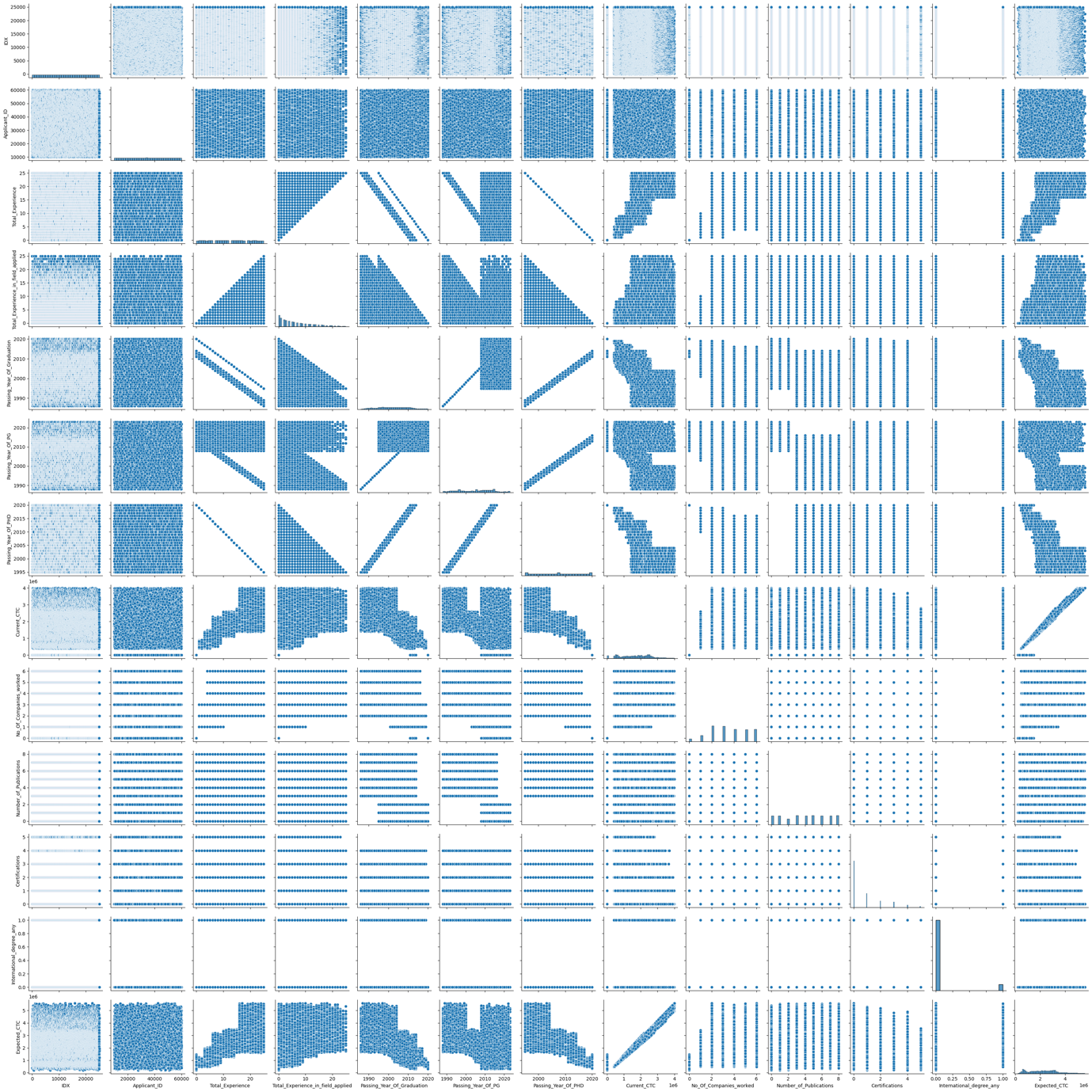
Majority of applicants are based on the Bangalore, Bhubaneswar, Mangalore, Jaipur.

Majority of applicants prefer Ahmedabad, Guwahati, Mangalore, and Kanpur.



* There are clear correlation between Current\_CTC and Expected\_CTC.
* There are also a relation between Expected\_CTC and Total\_Experience of an applicants.
* No further Correlations are visible.

Multi-variate analysis

* 

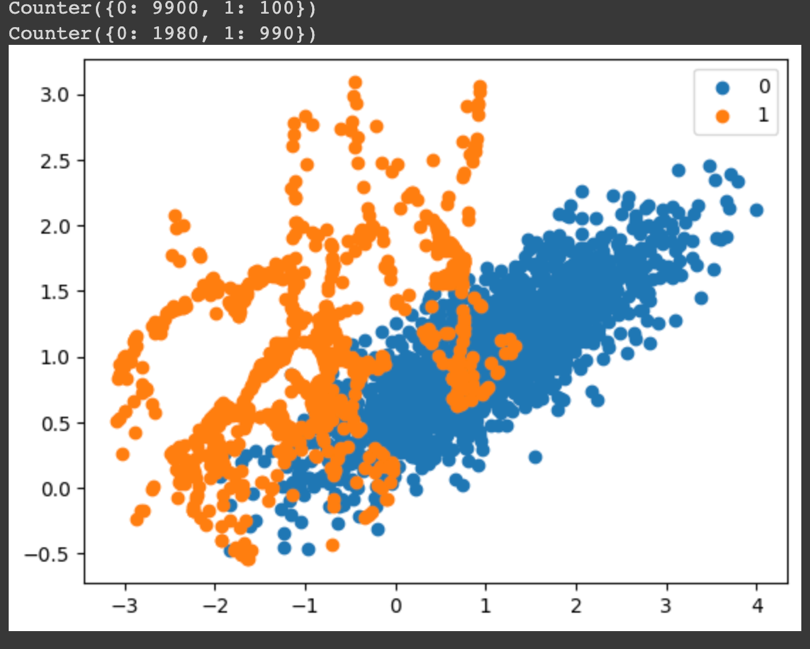
**VARIABLE TRANSFORMATION**

The Object type variable/features are converted into category type.

Then we use label encoded to encode the data as shown below:

|  |  |  |
| --- | --- | --- |
| **DEPARTMENT** | **DEPARTMENT\_ENCODED** | |
| Accounts | 12 | |
| IT-Software | 11 | |
| HealthCare | 10 | |
| Marketing | 9 | |
| Education | 8 | |
| Analyst/BI | 7 | |
| Others | 6 | |
| Engineering | 5 | |
| Sales | 4 | |
| Banking | 3 | |
| Top Management | 2 | |
| HR | 1 | |
| NAN | 0 | |
| **INDUSTRY** | **INDUSTRY\_ENCODED** | |
| Analytics | 1 | |
| Training | 2 | |
| Aviation | 3 | |
| Insurance | 4 | |
| Retail | 5 | |
| FMCG | 6 | |
| Others | 7 | |
| Telecom | 8 | |
| Automobile | 9 | |
| IT | 10 | |
| BFSI | 11 | |
| NAN | 0 | |
| **ORGANSIZATION** | **ORGANSIZATION\_ENCODED** | |
| A | 1 | |
| B | 2 | |
| C | 3 | |
| D | 4 | |
| D | 5 | |
| E | 6 | |
| F | 7 | |
| G | 8 | |
| H | 9 | |
| I | 10 | |
| J | 11 | |
| L | 12 | |
| M | 13 | |
| N | 14 | |
| O | 15 | |
| P | 16 | |
| **ROLE** | **ROLE\_ENCODED** | |
| Consultant | 0 | |
| Financial Manager | 1 | |
| Project Manager | 2 | |
| Area Sales Manager | 3 | |
| Team Lead | 4 | |
| Analyst | 5 | |
| Others | 6 | |
| CEO | 7 | |
| Business Analyst | 8 | |
| Sales Manager | 9 | |
| Bio Statistician | 10 | |
| Scientist | 11 | |
| Research Scientist | 12 | |
| Head | 13 | |
| Associate | 14 | |
| Senior Researcher | 15 | |
| Sales Executive | 16 | |
| Sr. Business Analyst | 17 | |
| Principal Analyst | 18 | |
| Data Scientist | 19 | |
| Researcher | 20 | |
| Senior Analyst | 21 | |
| Professor | 22 | |
| Lab Executive | 23 | |
| **Designation** | **Designation Encoded** |
| HR | 1 |
| Others | 2 |
| Manager | 3 |
| Product Manager | 4 |
| Sr. Manager | 5 |
| Consultant | 6 |
| Marketing Manager | 7 |
| Assistant Manager | 8 |
| Data Analyst | 9 |
| Research Analyst | 10 |
| Medical Officer | 11 |
| Software Developer | 12 |
| Web Designer | 13 |
| Network Engineer | 14 |
| Director | 15 |
| CA | 16 |
| Research Scientist | 17 |
| Scientist | 18 |
| NAN | 0 |
| **GRADUATION\_SPECIALIAZATION** | **ENCODED** |
| Chemistry | 0 |
| Economics | 1 |
| Mathematics | 2 |
| Zoology | 3 |
| Arts | 4 |
| Psychology | 5 |
| Sociology | 6 |
| Botony | 7 |
| Engineering | 8 |
| Others | 9 |
| Statistics | 10 |
| **UNIVERISTY\_GRAD** | **UNIVERSITY\_GRAD ENCODED** |
| Bhubaneswar | 0 |
| Delhi | 1 |
| Mangalore | 2 |
| Mumbai | 3 |
| Jaipur | 4 |
| Lucknow | 5 |
| Guwahati | 6 |
| Pune | 7 |
| Kolkata | 8 |
| Surat | 9 |
| Nagpur | 10 |
| Bangalore | 11 |
| Ahmedabad | 12 |
|  |  |
|  |  |
| **EDUCATION** | **EDUCATION\_ENCODED** |
| PG | 1 |
| Doctorate | 2 |
| Grad | 3 |
| Under Grad | 4 |
| **UNIVERSITY\_PG** | **UNIVERSITY\_P1G\_ENCODED** |
| Bhubaneswar | 0 |
| Delhi | 1 |
| Mangalore | 2 |
| Mumbai | 3 |
| Jaipur | 4 |
| Guwahati | 5 |
| Surat | 6 |
| Lucknow | 7 |
| Pune | 8 |
| Nagpur | 9 |
| Kolkata | 10 |
| Bangalore | 11 |
| Ahmedabad | 12 |
| **INHAND\_OFFER** | **INHAND\_OFFER\_ENCODED** |
| N | 0 |
| Y | 1 |
| **LAST\_Apprasial\_Rating** | **Encoded** |
| Key Performance | 0 |
| A | 1 |
| B | 2 |
| C | 3 |
| D | 4 |

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



The imbalance data is cleared using the smote method to now we can use the Machine Learning Technique to observe the data in the set

Any business insight using clustering? A picture containing diagram, rectangle, display, screenshot

Description automatically generated

* Hierarchical Clustering is an unsupervised learning method for clustering data points. The algorithm builds cluster by measuring the dissimilarities between data.
* By using K-means Clustering n-Cluster are 6 in data value is :- 583063.6
* After using this I just train and split the data set into 70 and 30 model as train data is 70% and test is 30%.

**MODEL BUILDING INTERPRETATION**

To forecast the Expected\_CTC for applicants who are applied for joining for various roles in the given data set Delta.LTD, we can create various regression models such as Linear Regression, Decision Tree, KNN, Ridge, LASSO, and some various models in this model building exercise. The major goal of the problem is to minimize human judgement regrading salary that offered by providing candidates with salary estimation at the time of joining based on their job, title, location, Salary, Experience, skills, and profile. It is critical to give employees a fair salary package that matches the market requirement’s while also being what honestly diverse.

**VARIOUS MODEL BUILDING**

Simple Linear Regression,

Lasso,

Ridge,

Decision TREE,

KNN, Random forest

Ensemble Modeling Technique

Bagging,

adaBoosting

Gradient Boosting

Linear Regression

OLS regression Results for the Predictive data you see the data with the values.

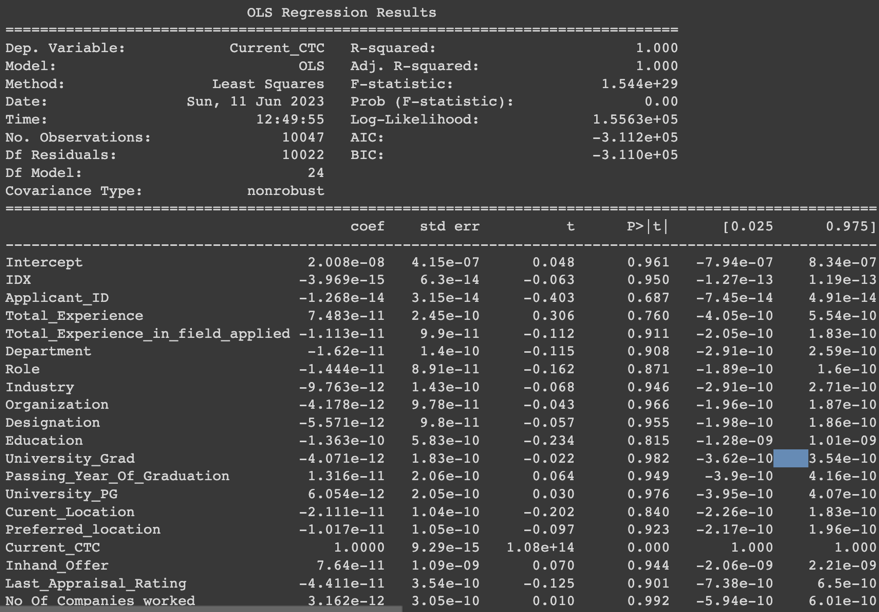
The R-Squared value is 1.000.

The summary of the OLS model data shows the Coefficient, Standard Error, P>|t| and soon

The standard errors assume that the covariance of matrix’s of the errors is correctly specified in the data description.

It is used by the Linear Regression to find the Covariance of matric and Fitted values in the dataset the P> | t | is decreased to 0 and then vif model should be built in the data set.

Figure 1: OLS model



A screenshot of a computer

Description automatically generated with medium confidence

VIF (Variance Inflation Factor):

VIF is measure of amount of multicollinearity in regression analysis. Multicollinearity exist when there is correlation between independent variable in multiple regression model. The best VIF model is > 10 as an indicator of multiple collinearity but choosing more threshold of 5 and 2.5 even they can choose by dropping here Current\_CTC Created the VIF model below

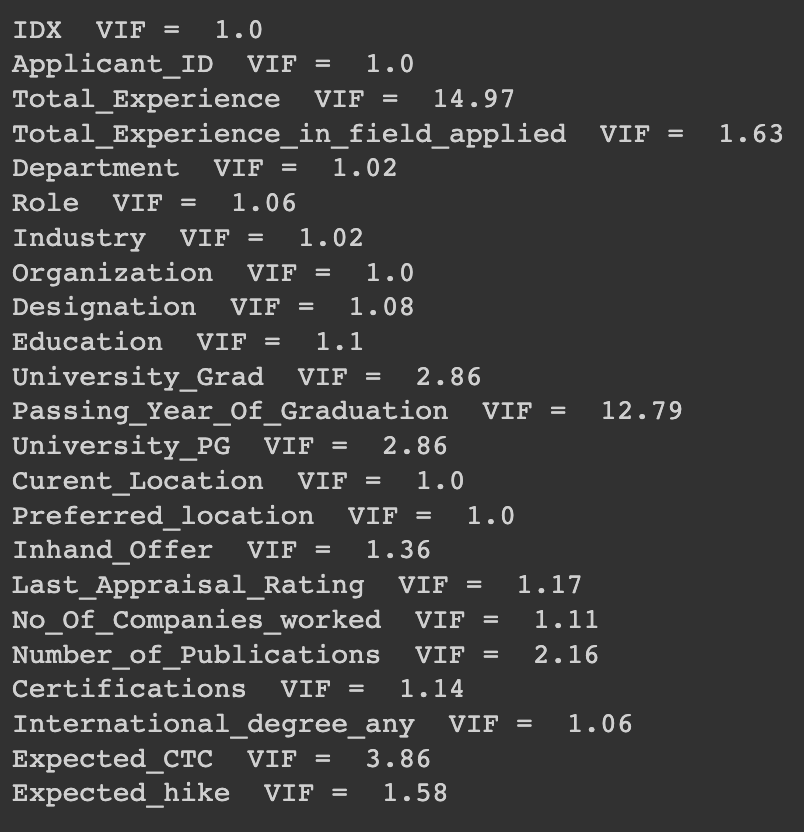


Figure 2:VIF MODEL !

Here below I just dropped of the Current CTC and Passing year of Graduations it can decreased the values show and I just took the below 5 and best model in the data set.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 3:VIF MODEL AFTER DROPING SOME DATA

After that I used the OLS model to check the P factor and it become all zero in the covariance matrix.

A screenshot of a computer screen

Description automatically generated with low confidence

Figure 4:OLS MODEL AFTER THE DROPPING

Simple Linear regression method

A picture containing line, screenshot, plot, diagram

Description automatically generated

R-squared and Adj R-squared both are same- 0.997

Both test and Train data performing well

|  |  |  |
| --- | --- | --- |
| Factors to compared | Simple Linear Regression Train | Simple Linear Regression for Test |
| R-Squared | 0.997 | 0.997 |
| Adj. R- Squared | 0.997 | 0.997 |
| RMSE | 48863.63 | 48587.75 |
| MAE | 19522334.58 | 3.50864 |
| MAPE | 4.2307 | 4.306 |

Decision Tree model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Train RMSE | TEST RMSE | Train Mae | Test MAE |
| Decision Tree | 2137559.094 | 4.35 | 1952334.86 | 3.506 |

R-Squared value shown in decision tree model was good model.

**Ridge model**

it is used for the model tuning methods that is used to analyze any data suffer from multicollinearity.

For any type of regression models machine learning models, the usually regression equation form which basically written as:

Y=XB+e.

The below shows the ridge model coefficient.

A screenshot of a computer

Description automatically generated with low confidence

Figure 5:RIDGE MODEL

LASSO model:

LASSO model score for the train dataset are 74.0%

And test data set are 75.0%.A picture containing text, font, screenshot

Description automatically generated

Figure 6:LASSO MODEL

And the RFM (Random Forest Classifier).

The RFM prams in the dataset.

Figure 7: RFM PARAMS

A screen shot of a computer code

Description automatically generated with low confidence

And then predicted the data using the Covariance matrix on the dataset.

A picture containing screenshot, text, colorfulness, display

Description automatically generated

Figure 8:Covariance matric.

The above show the predicted and True label in the dataset

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | Actual Values | Fitted Values | Residuals |
| 0 | 2718037 | 1.926036e+06 | 15418.944273 |
| 1 | 2609383 | 2.017288e+06 | 21293.027021 |
| 2 | 2869809 | 2.200829e+06 | 6717.286073 |
| 3 | 2133605 | 1.825179e+06 | -41227.930976 |
| 4 | 2686600 | 2.126835e+06 | 36288.526805 |

R-Squared: -99.6% after the fitted values and predicted the values.

MAPE for test data set 4.31.

**Model Tuning:**

By using Bagging Classifier, we tune the model using dTree methods as base estimator the predicted value for the Bagging classifier.

A screenshot of a graph

Description automatically generated with medium confidence

Figure 9:MODEL TUNING

ADDA Boosting Classifiers and Gradient boosting classifiers it shows the same values in the covariance matrix in the data.

The AUC curve values after fitted and predicted values after tuning the model is 52.01% for the given data set.

BUSINESS Insights and recommendations:

* According to our analysis the majority of the variables, such as Origination, Graduation Specialization, University Grad, PG Specialization, PHD Specialization etc., do not exhibit any variation with the target variable (Expected\_CTC), and there is no clear relationship between then and target.
* Therefore, we suggest the company to collect other variables from applicants, such as education background, work experience, etc. we observed in our data that passing year for UG, PG, doctoral degrees are request by applicants. Rather than requesting passing year for all year, we can ask for passing year according to greatest level of education an applicants may process.
* Company should focus on threads of market salary for different industries, roles, and designation so every applicant will get an unbiased salary which he/she truly deserve.
* As we saw in our analysis fresher applicant have zero total experience and zero current ctc. As these being the important predictor target. So, we need to build a separate model for such applicants.
* The hiring company should confirm with applicant that they have earned their degrees within expected time frame. Candidate with incomplete degrees’ or backlogs will receive lesser ctc.
* Recent graduate should be recruited by the company even if they are asking for lower starting salaries.
* Company needs to implement new HR initiative that meet candidate’s requests and help the business grow get their employees paid according to their budget.
* Applicants with higher number of companies work experience business should look for those applicant who worked lees number of company but have experienced well and perform well in those companies.