Employee Churn Prediction

TABLE OF CONTENTS

1) Introduction of the business problem and Data Insights

Business Problem:

a) Scope: We aim to predict the employees who will churn in the near future.

b) Need of the study/project-The idea is to minimise human intervention to calculate future churn of an employee who is joining the company but automate it with data of the past and predict the future.

c)Understanding business/social opportunity- Helping business to understand what precautionary measures can be taken to prevent the employees from leaving the organization.

Data Dictionary:

Field	Description
Age	Age of the employee
Attrition	Attrition of employee yes/no
BusinessTravel	If business travelling is required frequently or rarely
DailyRate	Per day Income
Department	The work department of the employee
DistanceFromHome	Travelling Time
Education	**Education number of degrees**
EducationField	Field of education taken in
Employee Count	Number of employees in that age or row
Employee Number	employee ID
Environment Satisfaction	Satisfied with the work environment
Gender	Male/Female
HourlyRate	Per hour Income
JobInvolvement	Rating as to which the job input, they give
JobLevel	Career level
JobRole	What they are a scientist/researcher or others in the date
JobSatisfaction	Satisfied with the kind of job they are doing
MaritalStatus	Married/single/divorced
MonthlyIncome	Per month income
MonthlyRate	**Per month income**
NumCompaniesWorked	Previous work count
Over18	Age if under 18 or not
OverTime	If work needs Overtime due to volume
PercentSalaryHike	Hike percentage in the salary
PerformanceRating	Performance rating given
RelationshipSatisfaction	Relationship with peers satisfaction level
StandardHours	Standardized work hours
StockOptionLevel	If holds any stock of the company
TotalWorkingYears	Number of work years with our company and others/total experience
TrainingTimesLastYear	Total training for job given
WorkLifeBalance	Rating the work life balance the company offers
YearsAtCompany	Years only with our company
YearsInCurrentRole	Years in the job role they have been working in
YearsSinceLastPromotion	Time duration from the day of promotion to now
YearsWithCurrManager	Time duration working with current manager

- **The data looks like:**

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7

5 rows × 35 columns

Fig 1a-Data head

Inference:

- **The data has** `(1470, 35)` **1470 rows and 35 columns.**
- **The data has the columns like:**

```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')
```

Fig 1b-Data Columns

- **The data has the 1470 rows and 35 columns of which 26 columns have integer values and 9 are categorical or object data type. The memory that the data is occupying is 402.1+KB which should be minimized by dropping unnecessary columns.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

Fig 1c -Data Types

- There is no null value in the data.
- There is no duplicates in the data.

```
Number of duplicate rows = 0
```

- Percentage of missing values in the data is null.

```
Values of any data in the dataset is False
```

```
Age                         0.0
Attrition                   0.0
BusinessTravel              0.0
DailyRate                   0.0
Department                  0.0
DistanceFromHome            0.0
Education                   0.0
EducationField              0.0
EmployeeCount               0.0
EmployeeNumber              0.0
EnvironmentSatisfaction     0.0
Gender                      0.0
HourlyRate                  0.0
JobInvolvement              0.0
JobLevel                    0.0
JobRole                     0.0
JobSatisfaction             0.0
MaritalStatus               0.0
MonthlyIncome               0.0
MonthlyRate                 0.0
NumCompaniesWorked          0.0
Over18                      0.0
OverTime                    0.0
PercentSalaryHike           0.0
PerformanceRating           0.0
RelationshipSatisfaction    0.0
StandardHours               0.0
StockOptionLevel            0.0
TotalWorkingYears           0.0
TrainingTimesLastYear       0.0
WorkLifeBalance             0.0
YearsAtCompany              0.0
YearsInCurrentRole          0.0
YearsSinceLastPromotion     0.0
YearsWithCurrManager        0.0
dtype: float64
```

Fig 1d -Data Null Values

2. EDA and Business Implication

- **Descriptive summary for numeric or integer data types:**

	count	mean	std	min	25%	50%	75%	max
Age	1470.0	36.923810	9.135373	18.0	30.00	36.0	43.00	60.0
Attrition	1470.0	0.161224	0.367863	0.0	0.00	0.0	0.00	1.0
DailyRate	1470.0	802.485714	403.509100	102.0	465.00	802.0	1157.00	1499.0
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.00	7.0	14.00	29.0
Education	1470.0	2.912925	1.024165	1.0	2.00	3.0	4.00	5.0
EmployeeCount	1470.0	1.000000	0.000000	1.0	1.00	1.0	1.00	1.0
EmployeeNumber	1470.0	1024.865306	602.024335	1.0	491.25	1020.5	1555.75	2068.0
EnvironmentSatisfaction	1470.0	2.721769	1.093082	1.0	2.00	3.0	4.00	4.0
HourlyRate	1470.0	65.891156	20.329428	30.0	48.00	66.0	83.75	100.0
JobInvolvement	1470.0	2.729932	0.711561	1.0	2.00	3.0	3.00	4.0
JobLevel	1470.0	2.063946	1.106940	1.0	1.00	2.0	3.00	5.0
JobSatisfaction	1470.0	2.728571	1.102846	1.0	2.00	3.0	4.00	4.0
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.00	4919.0	8379.00	19999.0
MonthlyRate	1470.0	14313.103401	7117.786044	2094.0	8047.00	14235.5	20461.50	26999.0
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.00	2.0	4.00	9.0
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.00	14.0	18.00	25.0
PerformanceRating	1470.0	3.153741	0.360824	3.0	3.00	3.0	3.00	4.0
RelationshipSatisfaction	1470.0	2.712245	1.081209	1.0	2.00	3.0	4.00	4.0
StandardHours	1470.0	80.000000	0.000000	80.0	80.00	80.0	80.00	80.0
StockOptionLevel	1470.0	0.793878	0.852077	0.0	0.00	1.0	1.00	3.0
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.00	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.00	3.0	3.00	6.0
WorkLifeBalance	1470.0	2.761224	0.706476	1.0	2.00	3.0	3.00	4.0
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.00	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.00	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.00	1.0	3.00	15.0
YearsWithCurrManager	1470.0	4.123129	3.568136	0.0	2.00	3.0	7.00	17.0

Fig2a-Data description

Inference:

- The count column shows all columns as 1470 which means no missing or null values.
- Mean column represents all the values at an average under each column. Mean is more for the monthly rate column and least is for Attrition column which is 0.16. This is least because data has only 1 or 0.
- These columns have high variations in the minimum and maximum values as per the descriptive statistics: Age, Daily rate, Distance from home, Employee number, Hourly rate, Monthly income,

monthly rate, Number of companies, Percentage in salary hike, number of companies worked in, Last promotion and years with manager.

- The standard deviation represents how far each score lies from the mean. Monthly rate is having the maximum STD.
- Min represents the minimum value under each column. Minimum value is again maximum for monthly rate. Most columns are zero as they are ratings of 0-5, hence minimum is 0.
- 25% ,50% and 75% represent the Q1, Q2 and Q3. 25% of the data is less than the value in each column under 25%,50% and 75%.

- **Getting the summary statistics of object data type:**

	count	unique	top	freq
Attrition	1470	2	No	1233
BusinessTravel	1470	3	Travel_Rarely	1043
Department	1470	3	Research & Development	961
EducationField	1470	6	Life Sciences	606
Gender	1470	2	Male	882
JobRole	1470	9	Sales Executive	326
MaritalStatus	1470	3	Married	673
Over18	1470	1	Y	1470
OverTime	1470	2	No	1054

Fig2b-Data Description-Categorical

Inference:

- The count in each column is 1470 so no null values. Top represents maximum subheading and frequency states how many times the top column is present in the data.

✓ Attrition has 'NO' as maximum with an occurrence of 1233.
✓ Employees Travel_rarely with an occurrence of 1043 times.
✓ Above 18 years we have more employees with an occurrence of 1470 times.

- Number of subheadings under each categorical column:

```
BusinessTravel : ['Travel_Rarely' 'Travel_Frequently' 'Non-Travel']
Travel_Rarely        1043
Travel_Frequently     277
Non-Travel            150
Name: BusinessTravel, dtype: int64
_____________________________________________________________________
Department : ['Sales' 'Research & Development' 'Human Resources']
Research & Development    961
Sales                     446
Human Resources            63
Name: Department, dtype: int64
_____________________________________________________________________
EducationField : ['Life Sciences' 'Other' 'Medical' 'Marketing' 'Technical Degree'
 'Human Resources']
Life Sciences       606
Medical             464
Marketing           159
Technical Degree    132
Other                82
Human Resources      27
Name: EducationField, dtype: int64
_____________________________________________________________________
Gender : ['Female' 'Male']
Male      882
Female    588
Name: Gender, dtype: int64
_____________________________________________________________________
JobRole : ['Sales Executive' 'Research Scientist' 'Laboratory Technician'
 'Manufacturing Director' 'Healthcare Representative' 'Manager'
 'Sales Representative' 'Research Director' 'Human Resources']
Sales Executive              326
Research Scientist           292
Laboratory Technician        259
Manufacturing Director       145
Healthcare Representative    131
Manager                      102
Sales Representative          83
Research Director             80
Human Resources               52
Name: JobRole, dtype: int64
_____________________________________________________________________
MaritalStatus : ['Single' 'Married' 'Divorced']
Married     673
Single      470
Divorced    327
Name: MaritalStatus, dtype: int64
_____________________________________________________________________
Over18 : ['Y']
Y    1470
Name: Over18, dtype: int64
_____________________________________________________________________
OverTime : ['Yes' 'No']
No     1054
Yes     416
Name: OverTime, dtype: int64
_____________________________________________________________________
```

Fig2c-Values under Categorical columns

Converting Attrition column to numeric before exploratory data analysis (graphs and plots): 1 represents, set of terminated employees and 0 represent active employees

BEFORE CODING

	Age	Attrition
0	41	Yes
1	49	No
2	37	Yes
3	33	No
4	27	No

AFTER CODING

	Age	Attrition
1	49	0
3	33	0
4	27	0
5	32	0
6	59	0

Taken:

NO, as Zero = Active employees

Yes, as One = Terminated employees

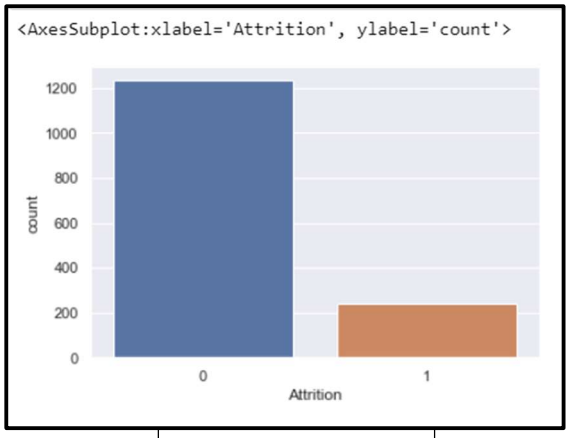


Fig2d-Attrition distribution

Inference:

- **Maximum people are active which is 0 and minimum is inactive which is 1.**
- **83.87% of the employees are active and 16.12% is inactive/terminated.**

```
No     1233
Yes     237
Name: Attrition, dtype: int64
```

- **Since the data has imbalanced target variable, the accuracy of model giving maximum Yes as Yes and No as No is** `0.8077858880778589`.

- **The value of turnover rate is:**

```
Value of Turnover Rate is 0.1922141119221411
```

UNIVARIATE ANALYSIS

Fig2f-Univariate Analysis

Inference:

- Age has the highest value in the range 30-35 and lowest at 55-60. We can say maximum employees are middle aged employees.
- Daily Rate is highest in 530-540 lowest at 1000.
- Distance is maximum at 0-2 and minimum at 15.
- Education is maximum for 3 levels and least for 5. Assuming maximum is Graduate and least are PHD or Doctorates.
- Employee count and standard hours is in having only one value throughout so we can delete them.
- Performance rating maximum is 3 and minimum is 4. Considering no ratings is in between 3-4.
- Relationship Satisfaction we see that the values are 3 as highest rating and 1 as lowest.
- Stock Option level has 0 as maximum so majority employees don't hold stocks. 3 being the least.
- Working years is more at 10 and least as 40.
- Employee number makes no inference as we have no major information from it.
- Environment Satisfaction is maximum 3 and least is 1.
- Maximum hourly rate is 100 and least is 30.
- Jobinvolvment shows no one is giving 4 but 3 is the highest and 1 is the lowest.
- Job level: 1 is more and 5 is least.
- Job satisfaction states 4 is the maximum and least is 2.
- Monthly income falls more in 3000 slab and least in 15000.
- Monthly rate maximum falls in 5000 and least in 20000
- Number of companies being maximum at 1 and least at 5.
- Percentage salary hike is maximum at 5% and least at 23%.
- Training TimeslastYear states 2 trainings happened for maximum and 0 for the least.
- Work life balance has got a 3 as maximum and 1 as a least.
- YearAtCompany, we see 0-5 being the maximum and least being 30-40.
- YearsIn the Current role maximum is 2 and minimum is 16.
- Years since last Promotion maximum scored is 0. So indicates people do get promotion often.
- Years with current manager is more for 3 years least for 15 and beyond.

BI_VARIATE ANALYSIS

❖ **ANALYSING AGE:**

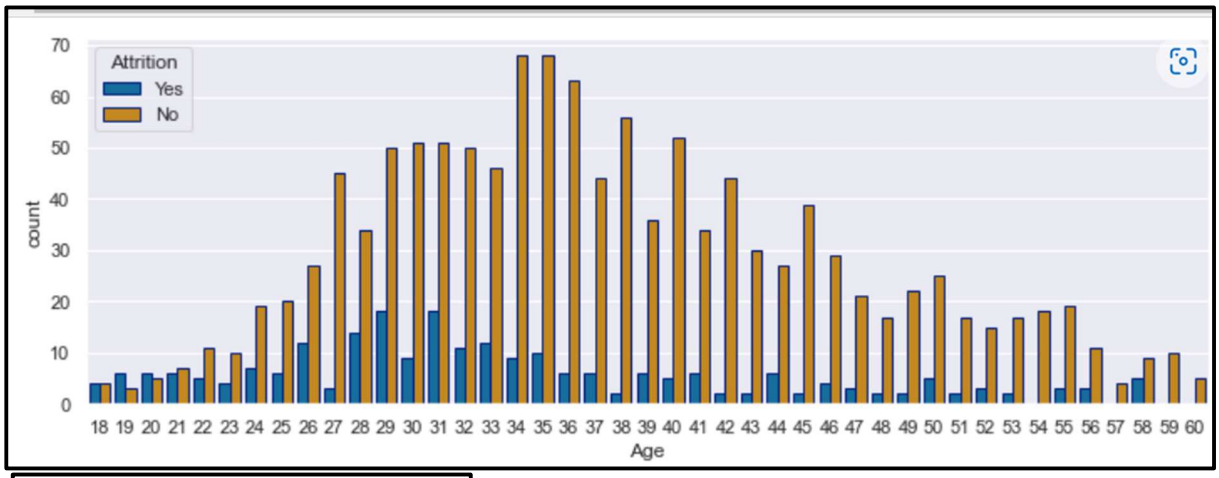


Fig2g-Age

Inference:

- We see maximum age is 34-35 who are active employees. Attrition is maximum for 29 and 31 years.
- Human resources have the highest count if we go age wise. Lets see the department which is having maximum attrition.

❖ **ANALYSING ATTRITION TO DEPARTMENT AND GENDER WISE:**

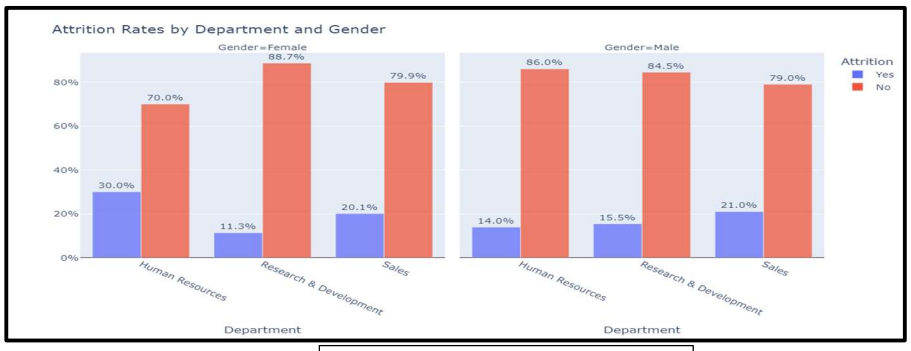


Fig2h- Department-Gender

Inference:

- Research and Development team has the maximum Active employees for Female and for Male we see Human Resources having maximum employees as active.
- Human Resources team has the maximum number of attritions in females and Sales has the maximum attrition for males. Least attrition is in least at 26.8% in Research and Development team for both males and females combined together.

❖ **ANALYSING ATTRITION TO JOBROLE AND GENDER:**

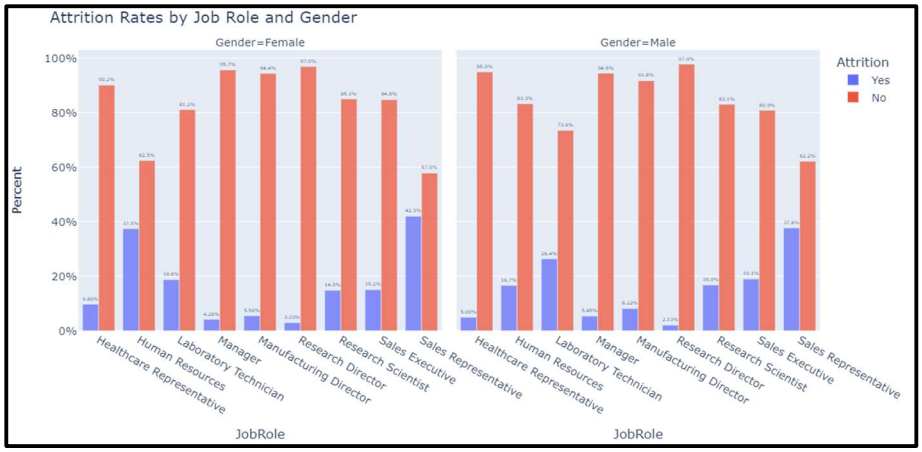


Fig2h- Job Role -Gender

Inference:

- We can see that the least attrition is in Research Director of 2.13%(male) and 3.03% (female).
- Maximum attrition is from Sales Representative for male (37.8%) and female (42.1%).
- Retention is maximum for Research Director and Manager both male and female.

❖ **ANALYSING ATTRITION TO GENDER:**

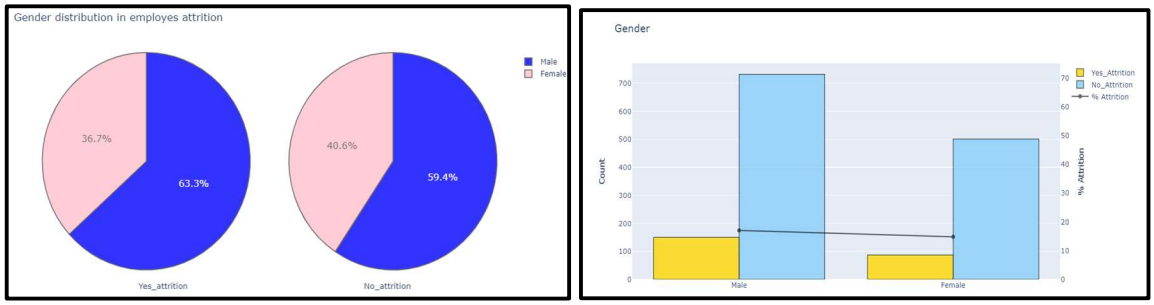


Fig2i- Attrition-Gender

Inference:

- **Female count is lower than male which the bar plots in the left represent.**
- **The attrition count in female is less and male is more.**
- **Females have an attrition of 36.7% whereas males have an attrition of 63.3% which is too high.**

❖ ANALYSING ATTRITION TO TRAVEL FREQUENCY:

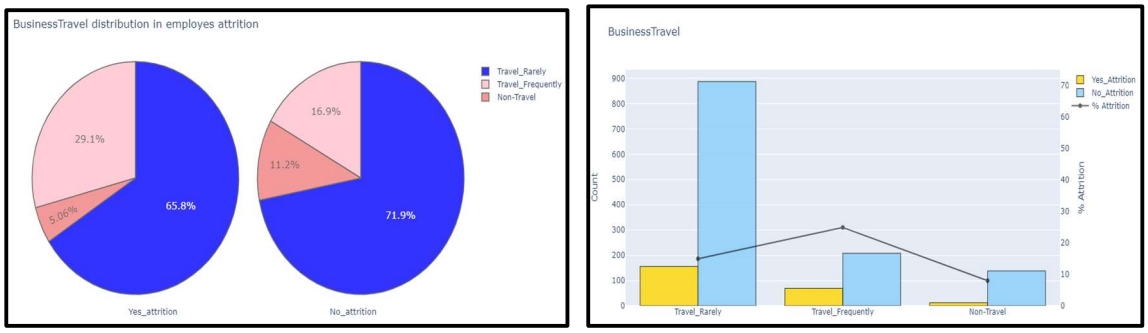


Fig2j- Attrition-Travel

Inference:

- **Employees who rarely travel have the highest count whereas non-Travels the least.**
- **We see attrition is more for rarely travel which is 65.8% and least for non-Travel.**

❖ ANALYSING ATTRITION TO OVERTIME:

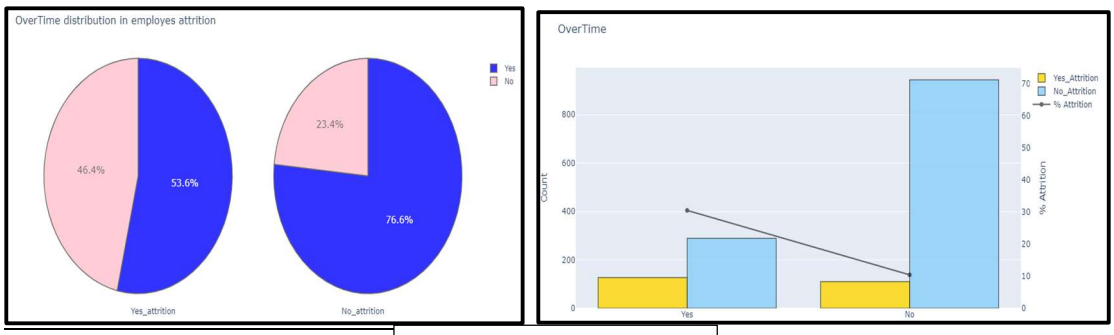


Fig2k- Attrition-Overtime

Inference:

- **Employees who do an overtime is lesser in count than who work in the work hours.**
- **The attrition is more for who work overtime that is 53% of the attrition. The retention rate is high here as No attrition has a good count comparatively to other attributes when compared to Attrition.**

INFERENCE:

- **The Attrition rate is 53.6% which is contributed mainly by laboratory technician.**
- **Research Director has the least attrition.**

❖ ANALYSING ATTRITION DEPARTMENT WISE:

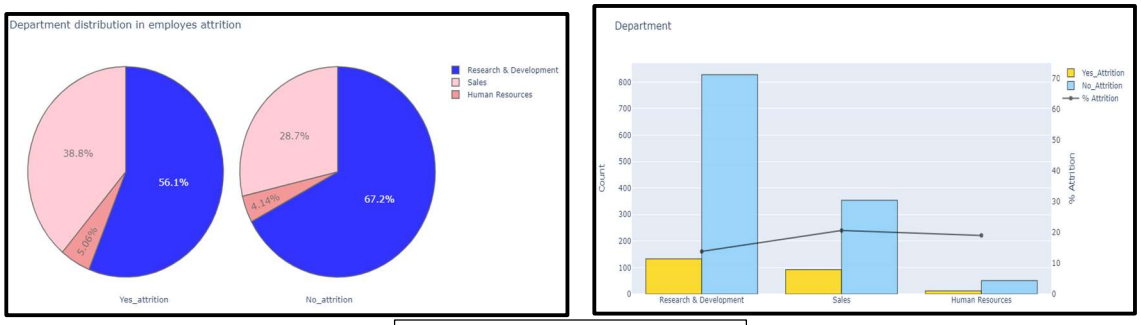


Fig2m- Attrition-Department

INFERENCE:

- **The attrition is maximum for Research and Development team with a 56.1%. Least is for Human Resources.**
- **The retention is higher for Research and Development department itself as the count is higher than other departments.**

❖ ANALYSING ATTRITION MARITAL STATUS WISE:

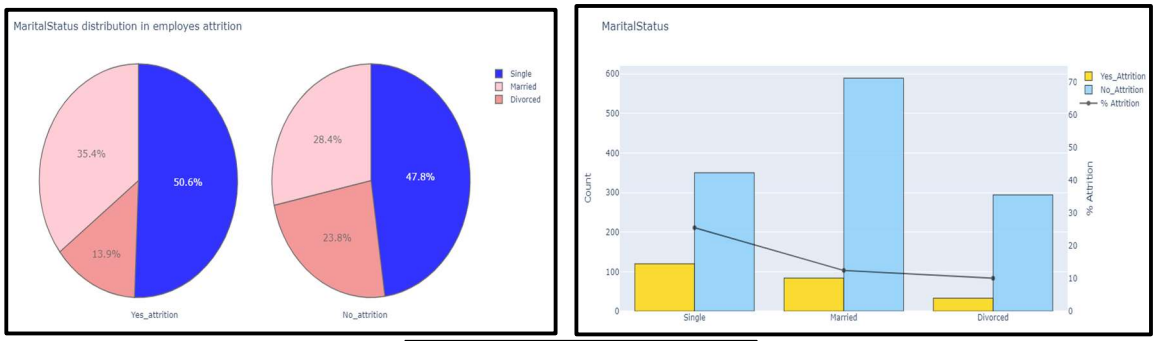


Fig2n- Attrition-Marital Status

INFERENCE:

- **The maximum attrition is from the status single which is 50.6%. The count of married people is the highest.**
- **The least attrition is for Divorced status employees.**

❖ ANALYSING ATTRITION EDUCATION FIELD WISE:

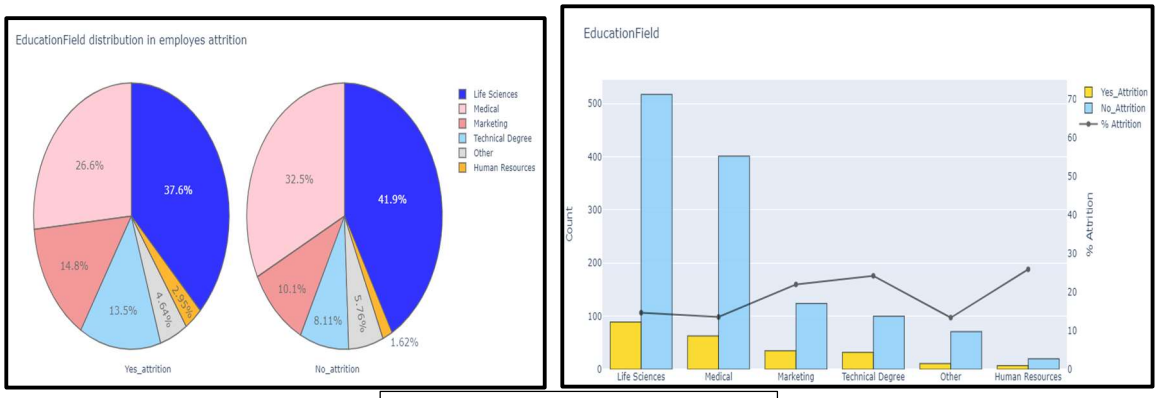


Fig2m- Attrition-Education Field

Inference:

- **The attrition is the maximum for Life Sciences and the least attrition is for Others and Human Resources.**
- **The attrition rate for life sciences is 37.6% least being of others at 4.64%.**

❖ ANALYSING ATTRITION TO NUMBER OF EDUCATION WISE:

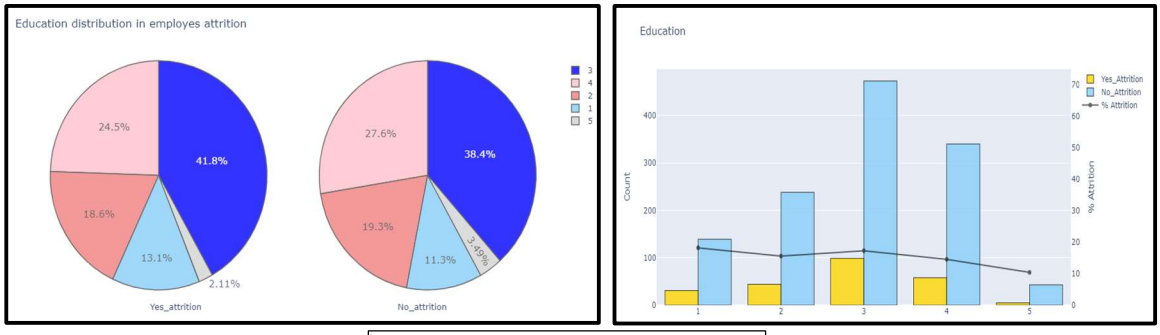


Fig2n- Attrition-Education Wise

INFERENCE:

- **The attrition is the highest for the employees with an education degree of 3.The least is for 5.**
- **The count for employees with education degree of 3 is the maximum.**

❖ ANALYSING ATTRITION TO ENVIRONMENT SATISFACTION:

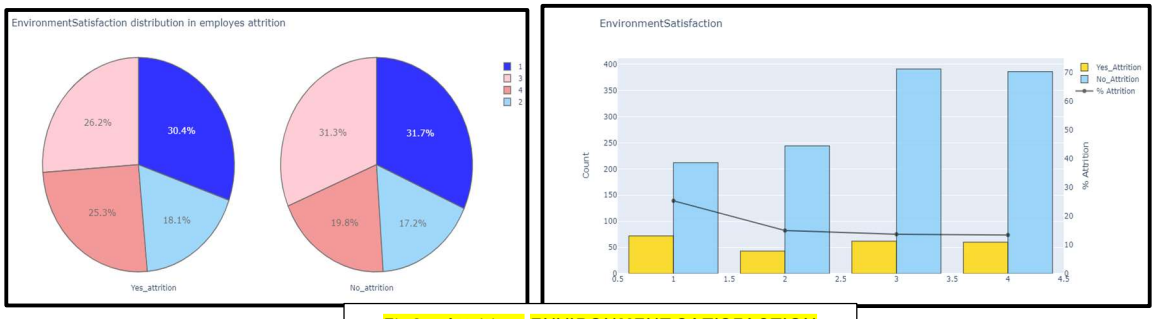


Fig2o- Attrition- ENVIRONMENT SATISFACTION

INFERENCE:

- **The maximum attrition is for employees with satisfaction of 1 contributing 30% of the attrition rate.**
- **The minimum is for satisfaction of 2 with a 18.1% attrition rate.**

❖ ANALYSING ATTRITION TO JOB INVOLVEMENT:

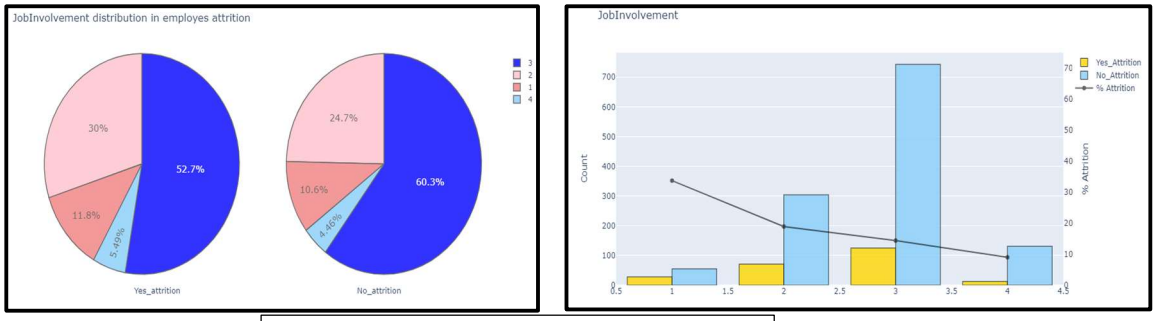


Fig2p- Attrition- Job Involvement

INFERENCE:

- **The attrition for job involvement rating of 3 is more. The least is for the rating 4.**
- **Here higher the rating more satisfied is the employee with the job involvement.**

❖ ANALYSING ATTRITION TO JOB LEVEL:

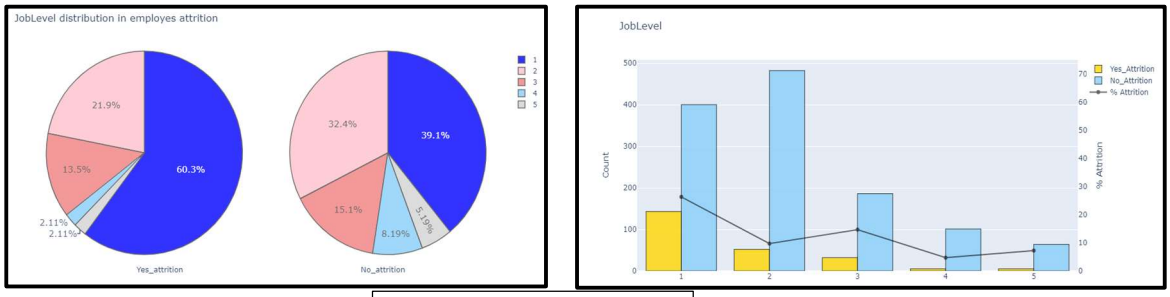


Fig2q- Attrition- Job Level

INFERENCE:

- **The attrition is more for Job level 1 and the least is for job level 5.**
- **Here the maximum count with no attrition is for job level 2.**
- **We can say that 5 is of a higher job level than 1.**

❖ ANALYSING ATTRITION TO JOB SATISFACTION:

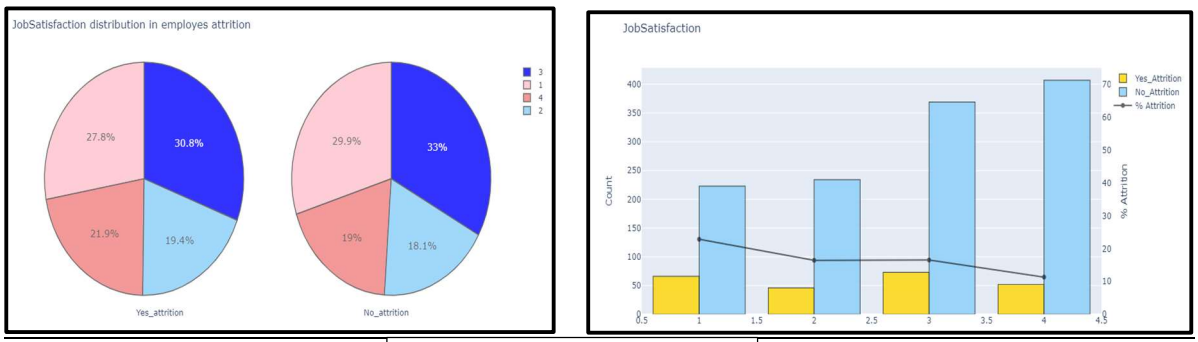


Fig2r- Attrition- Job Satisfaction

INFERENCE:

- **The higher attrition is for 3 job satisfaction rating and the least is for 2.**
- **There is more count for job satisfaction with a score of 4.**

❖ ANALYSING ATTRITION TO PERFORMANCE RATING:

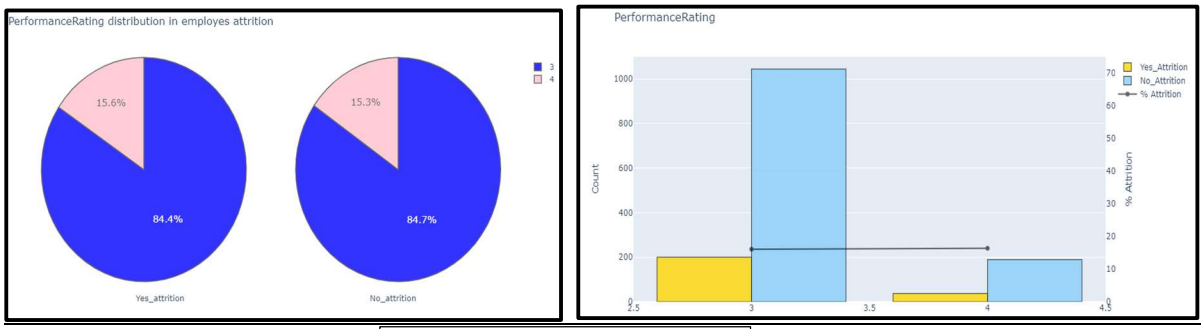


Fig2s- Attrition- Performance Rating

INFERENCE:

- **The performance rating is more for 3 and hence the attrition is more for people with a rating of 3. There is a 84.4% of total attrition from performance rating by employees with 3 ratings.**
- **The count for 4 is very less.**

❖ ANALYSING ATTRITION TO RELATIONSHIP SATISFACTION:

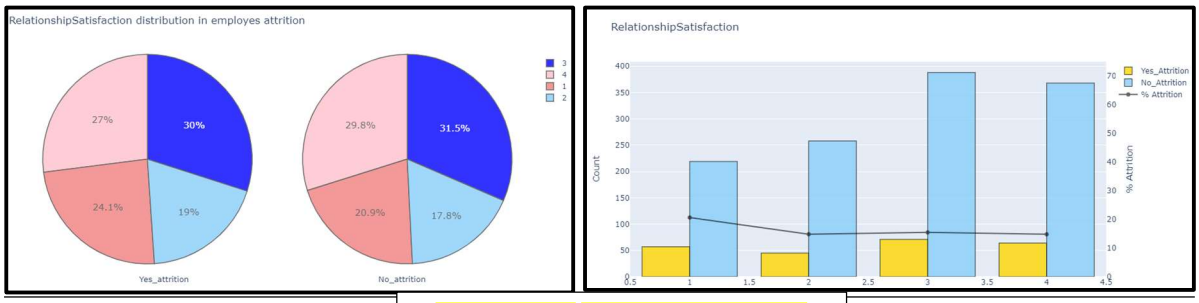


Fig2s- Attrition- Relationship Satisfaction

Inference:

- **Employees with a relationship of 3 has the maximum contribution to attrition which is 30% and least is for 2.**
- **The count of people with 3 rating is more.**

❖ ANALYSING ATTRITION TO RELATIONSHIP STOCK OPTION LEVEL:

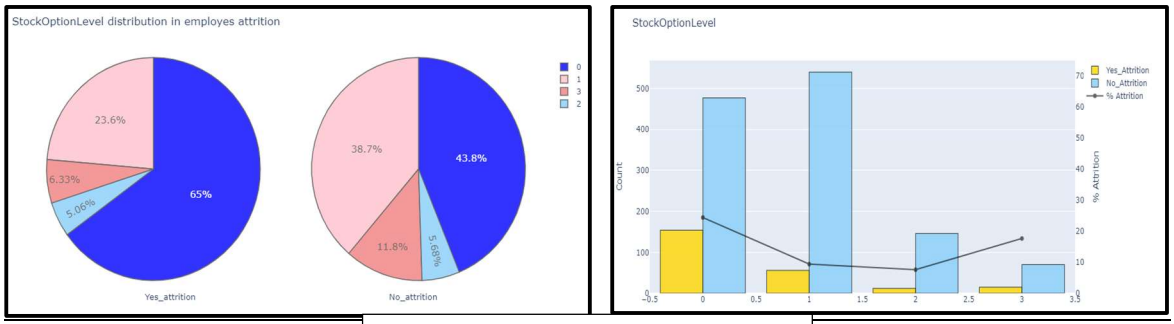


Fig2t- Attrition- Stock option level

Inference:

- **Stock option level of 0 has more attrition and the least is for employees with 2 stocks.**
- **The count of people with stock of 1 is more.**

❖ ANALYSING ATTRITION TO RELATIONSHIP SATISFACTION:

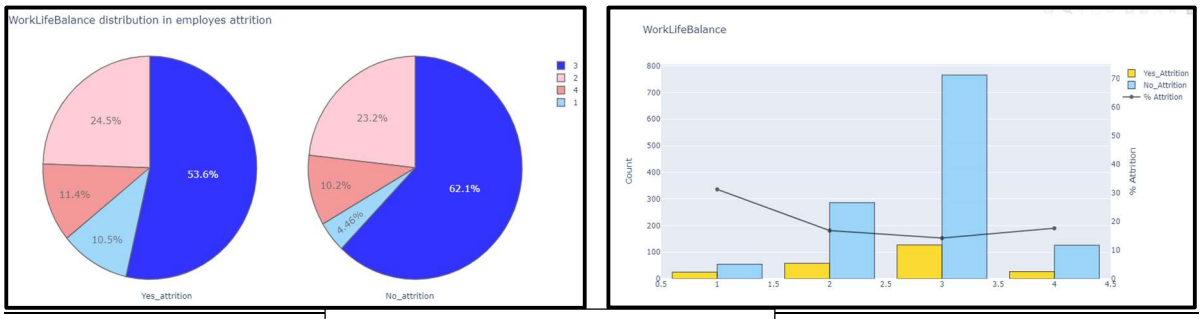


Fig2q- Attrition- Relationship Satisfaction

Inference:

- **Work life balance with a rating of 3 has maximum attrition and the least is with a rating of 1.**
- **The work life balance of rating 3 contributes 53.6% which is half and remaining ratings consists of the other 50%. This needs to be resolved as 50% of the employees aren't getting work life balance.**

❖ ANALYSING ATTRITION TO DEPARTMENT AND GENDER:

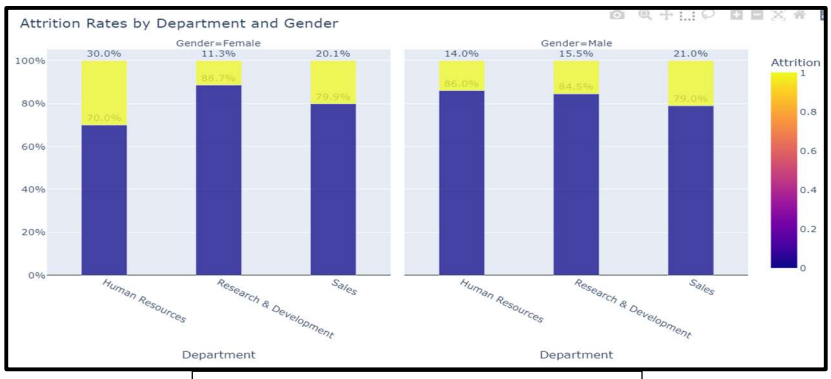


Fig2r- Attrition- Department and Gender wise

Inference:

- **Maximum attrition is for Human Resources and Female.**
- **Research and Development has minimum attrition in Female and Human Resources in Male.**

❖ ANALYSING ATTRITION TO DEPARTMENT AND GENDER:

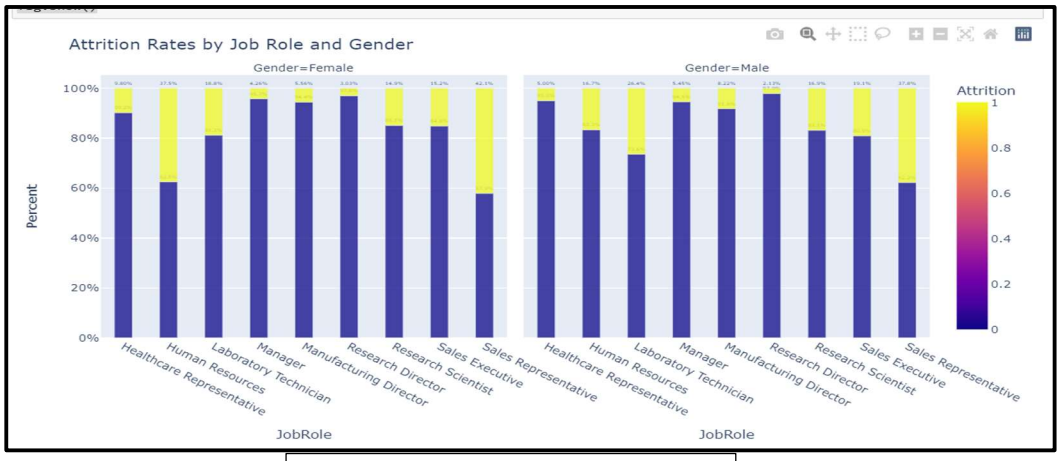


Fig2s- Attrition- Job Role and Gender wise

Inference:

- Maximum attrition is for Sales representative for both male and female.
- Laboratory technician in male also has attrition more than other departments.
- Least is for Research Director.

MULTIVARIATE ANALYSIS

Fig2t Pair Plot

Inference:

- The data is highly related amongst each other.
- One factor seems to influence the other.

HEAT MAP ANALYSIS

To know the above inferences better let's do a collinearity check in order to know exactly why values are high and low as inferred from histogram:

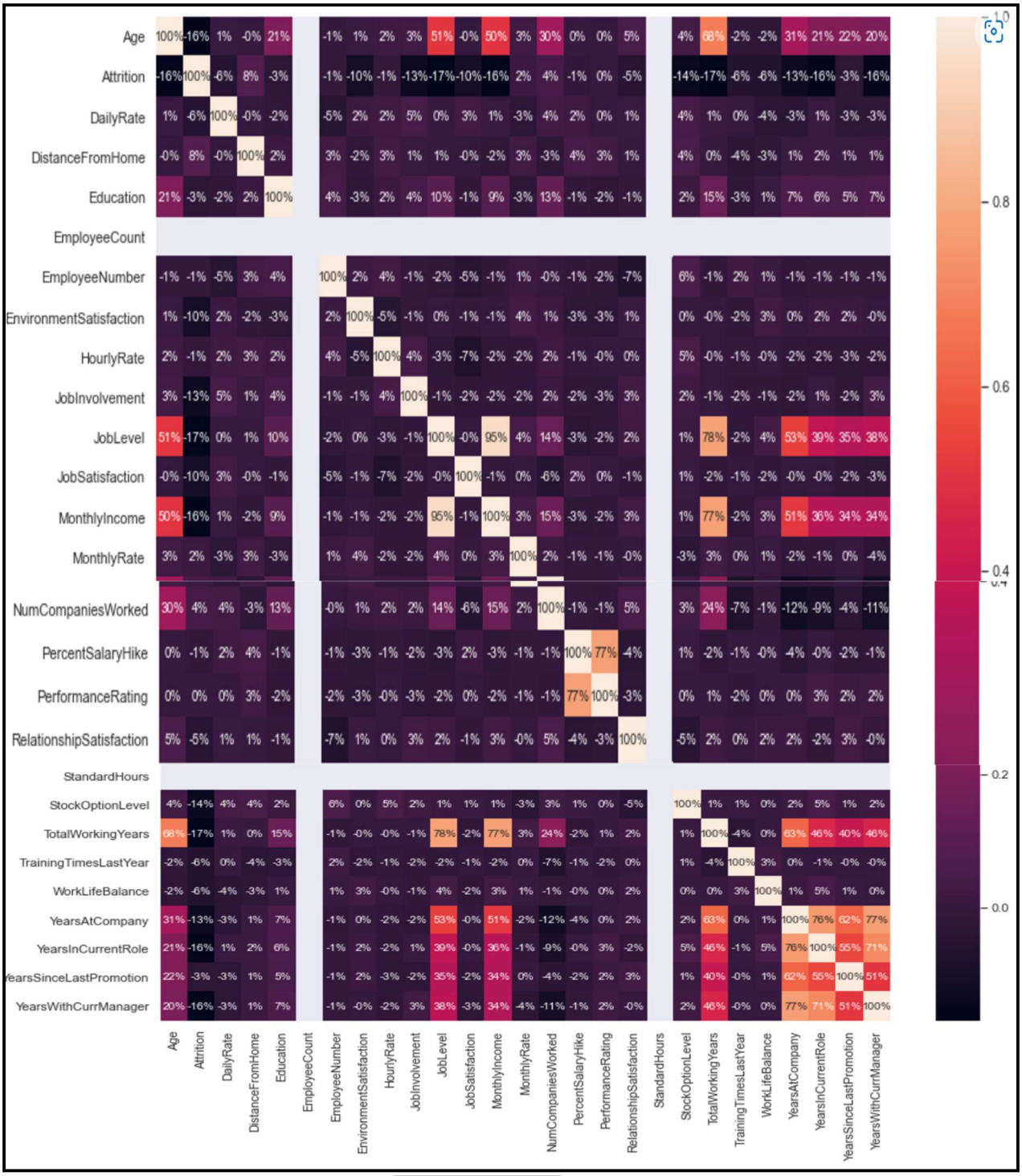


Fig2u Heat Map

Inference:

- We see total working years to Age is corelated at 68%. This is true as more the age more is the year of experience.
- Total number of working years is corelated to Job level. The more the experience the higher the level/position in the job. This also relates to Total working years being 77% corelated to Monthly income. More the number of working years, more is the salary and higher is the job level.
- Job level and monthly income is red in 50% collinearity as the age for job was maximum at 30-35years.
- Monthly income is also high for those with more working years hence corelated.
- Percentage hike to performance rating is corelated as better performance more hike.
- Years in the company is corelated to years in the current role, years since last promotion and years with current manager as more the number of years more will be all these 3 parameters applicable to the employee.

3) Data Cleaning and Pre-processing

- Dropped columns like Standard Hours, Employee count and Over18 as it was having the same values in all rows. Employee Number is not helpful since its just an ID with no unique interpretation.

```
(1470, 31)
```

- Result for outliers states total attributes with outliers are 11:

```
No. of outliers in Attrition: 237
No. of outliers in MonthlyIncome: 114
No. of outliers in NumCompaniesWorked: 52
No. of outliers in PerformanceRating: 226
No. of outliers in StockOptionLevel: 85
No. of outliers in TotalWorkingYears: 63
No. of outliers in TrainingTimesLastYear: 238
No. of outliers in YearsAtCompany: 104
No. of outliers in YearsInCurrentRole: 21
No. of outliers in YearsSinceLastPromotion: 107
No. of outliers in YearsWithCurrManager: 14


No of attributes with outliers are : 11
```

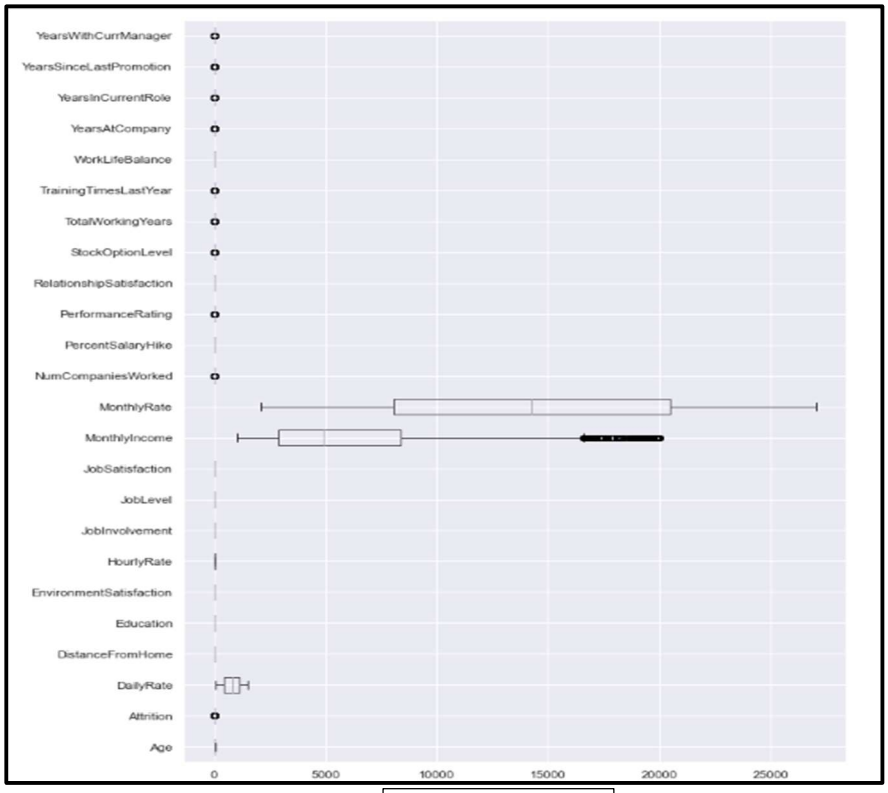
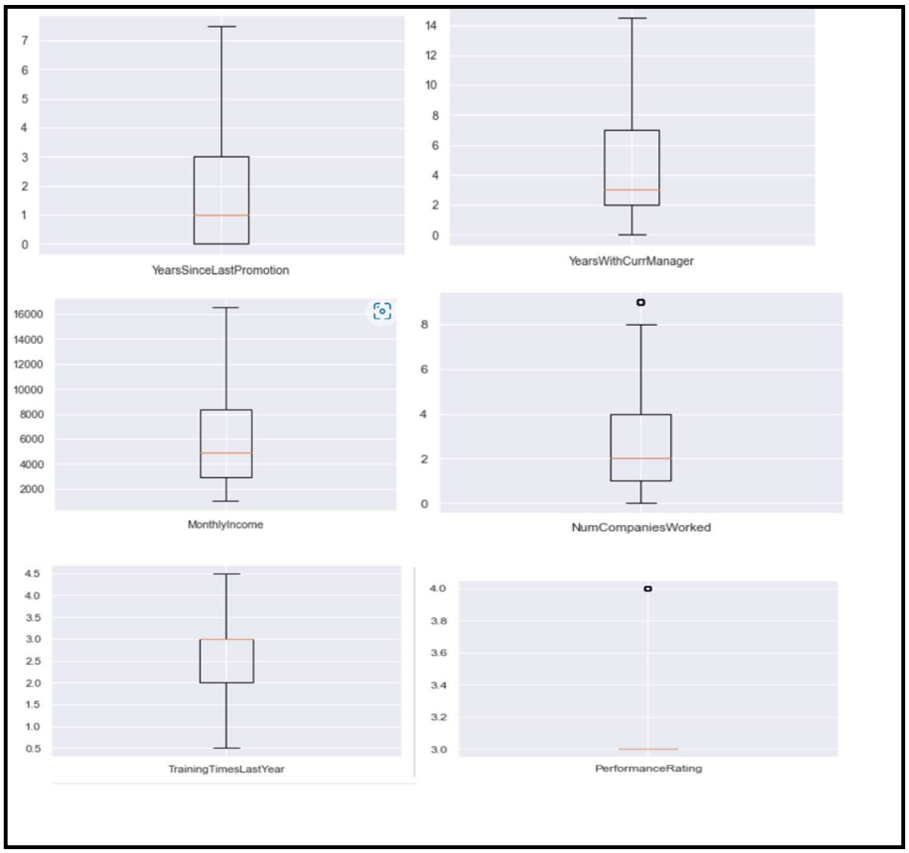


Fig 3a - Outliers

- **Removing outliers to avoid fitting issues of the model.**



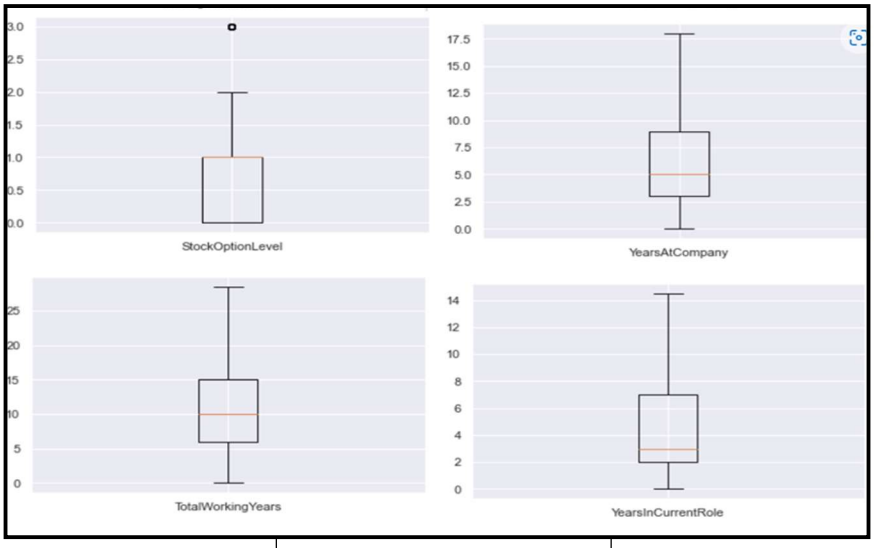


Fig 3b – Outliers Treatment

Inference:

- **The field wise co-relation of attributes to Attrition. The data has outliers still but less than what it had. The outliers present aren't that corelated with target variable so we can move ahead with these outliers.**
- **We have removed outliers only of the fields that had it and did not treat the entire data on it.**

Corelation to Attrition

```
Attrition                  1.000000
DistanceFromHome           0.077924
NumCompaniesWorked         0.043494
MonthlyRate                0.015170
PerformanceRating          0.002889
HourlyRate                -0.006846
PercentSalaryHike         -0.013478
Education                 -0.031373
YearsSinceLastPromotion   -0.033019
RelationshipSatisfaction  -0.045872
DailyRate                 -0.056652
TrainingTimesLastYear     -0.059478
WorkLifeBalance           -0.063939
EnvironmentSatisfaction   -0.103369
JobSatisfaction           -0.103481
JobInvolvement            -0.130016
YearsAtCompany            -0.134392
StockOptionLevel          -0.137145
YearsWithCurrManager      -0.156199
Age                       -0.159205
MonthlyIncome             -0.159840
YearsInCurrentRole        -0.160545
JobLevel                  -0.169105
TotalWorkingYears         -0.171063
Name: Attrition, dtype: float64
```

Fig 3c – Corelation to Attrition

OUTLIERS

```
No. of outliers in Attrition: 237
No. of outliers in NumCompaniesWorked: 52
No. of outliers in PerformanceRating: 226
No. of outliers in StockOptionLevel: 85


No of attributes with outliers are : 4
```

Fig 3d – Outliers left post treatment

- **ENCODING the data:**

The categorical columns have been done one hot encoding:

	Age	Attrition	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	...	JobRole_Labor Techn
0	41	1	1102	1	2	2	94	3	2	4	...	
1	49	0	279	8	1	3	61	2	2	2	...	
2	37	1	1373	2	2	4	92	2	1	3	...	
3	33	0	1392	3	4	4	56	3	1	3	...	
4	27	0	591	2	1	1	40	3	1	2	...	

5 rows × 45 columns

Fig 3e – Encoded Data Heading

- **Post data encoding we checked multi-collinearity which states the below columns are high contributors. The are relate more amongst themselves and should be removed. However major chunk will get removed if dropping these.**

```
There are 8 columns to remove :

['MonthlyIncome',
 'PerformanceRating',
 'TotalWorkingYears',
 'YearsInCurrentRole',
 'YearsWithCurrManager',
 'BusinessTravel_Travel_Rarely',
 'Department_Sales',
 'JobRole_Sales Executive']
```

threshold = 0.75 is taken into consideration

- **Variance_inflation_factor is highest for Department of Research and development and Performance rating. Lest is for Job Manufacturing Director and Overtime done yes.**

	Features	VIF_values
0	Age	34.973932
1	DailyRate	5.087776
2	DistanceFromHome	2.327308
3	Education	9.661867
4	EnvironmentSatisfaction	7.287676
5	HourlyRate	11.573748
6	JobInvolvement	15.584312
7	JobLevel	52.232701
8	JobSatisfaction	7.173085
9	MonthlyIncome	45.778779
10	MonthlyRate	5.087879
11	NumCompaniesWorked	2.772352
12	PercentSalaryHike	44.938461
13	PerformanceRating	159.643509
14	RelationshipSatisfaction	7.419002
15	StockOptionLevel	3.522498
16	TotalWorkingYears	16.531527
17	TrainingTimesLastYear	7.980208
18	WorkLifeBalance	16.226150
19	YearsAtCompany	18.558064
20	YearsInCurrentRole	8.284030
21	YearsSinceLastPromotion	2.588216
22	YearsWithCurrManager	8.250131
23	BusinessTravel_Travel_Frequently	2.902465
24	BusinessTravel_Travel_Rarely	8.069918
25	Department_Research_and_Development	93.398828
26	Department_Sales	52.927738
27	EducationField_Life_Sciences	39.745908
28	EducationField_Marketing	11.901847
29	EducationField_Medical	30.665316
30	EducationField_Other	6.236998
31	EducationField_Technical_Degree	9.481653
32	Gender_Male	2.558163
33	JobRole_Human_Resources	4.717404
34	JobRole_Laboratory_Technician	3.855170
35	JobRole_Manager	4.152928
36	JobRole_Manufacturing_Director	2.121463
37	JobRole_Research_Director	2.648094
38	JobRole_Research_Scientist	4.242715
39	JobRole_Sales_Executive	17.998928
40	JobRole_Sales_Representative	5.584305
41	MaritalStatus_Married	3.378382
42	MaritalStatus_Single	4.385166
43	OverTime_Yes	1.440002

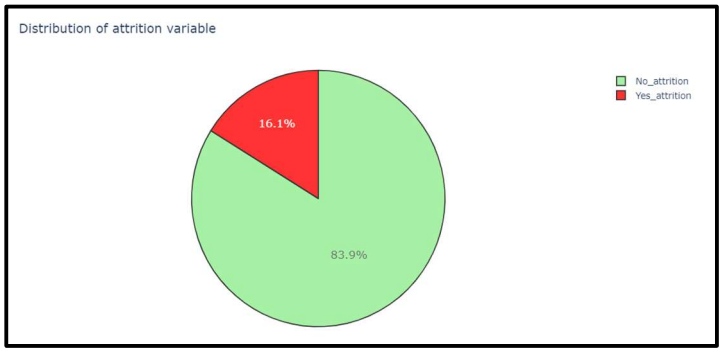
Fig 3f – VIF data

4) Model Building

- Since our target variable is an attribute with 'Yes' or 'No' we have coded them as YES=1 and NO = 0

```
0    1233
1     237
Name: Attrition, dtype: int64
```

- The spread of our target variable is as:



- The train and test set looks like:

  X_train: (1029, 44)
  X_test: (441,44)
  y_train: (1029,)
  y_test: (441,)

- **The train and test have been split into 70:30 ratio.**

MODEL 1:

```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

- ❖ The accuracy result is 85.7% which is good for a model on test data.
- ❖ Performance of the model on train and test data is:

Train Data

```
1.0
[[853   0]
 [  0 176]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       853
           1       1.00      1.00      1.00       176

    accuracy                           1.00      1029
   macro avg       1.00      1.00      1.00      1029
weighted avg       1.00      1.00      1.00      1029
```

Test Data

```
0.8049886621315193
[[333  47]
 [ 39  22]]
              precision    recall  f1-score   support

           0       0.90      0.88      0.89       380
           1       0.32      0.36      0.34        61

    accuracy                           0.80       441
   macro avg       0.61      0.62      0.61       441
weighted avg       0.82      0.80      0.81       441
```

Inference:

- The model shows a precision, recall and f1-score of 1. Accuracy is also 1 which is impossible for a model to have.
- This clearly means as data is imbalanced, we see over fitting in train data.
- We need to balance the data and go ahead with predicting other models.
- The test data gives an accuracy of 80.4% which indicates a difference of 20% between train and test models. This will lead to my model performing poorly with future unknown data.
- Train and Test precision, recall and F1-score also shows a huge difference which means my predictions have been maximum false. Model predicted No attrition when it was actually attrition.

MODEL IMPROVEMENT

- Since the model is bias towards NO ATTRITION, we will Over sample our data and bring up the data of ATTRITION to near about NO ATTRITION.
- METHOD USED: SMOTE (Synthetic minority oversampling technique) -Since data is small and we don't have millions of records we will proceed with SMOTE.
- SMOTE IS ONLY APPLIED ON TRAINING DATA TO AVOID OVERFITTING.
- New sample looks like: X_train, y_train

```
(1458, 44)
(1458,)
```

60:40

```
(1706, 44)
(1706,)
```

70:30

MODEL 2:

```
RandomForestClassifier
RandomForestClassifier(random_state=1)
```

- The model gives us an accuracy of 86.8% which is a very good score.
- Tried a 70:30 split in the data post SMOTE.
- Performance of the model on train and test data is:

Train Data

```
1.0
[[853   0]
 [  0 853]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       853
           1       1.00      1.00      1.00       853

    accuracy                           1.00      1706
   macro avg       1.00      1.00      1.00      1706
weighted avg       1.00      1.00      1.00      1706
```

Test Data

```
0.63718820861678
[[239 141]
 [ 19  42]]
              precision    recall  f1-score   support

           0       0.93      0.63      0.75       380
           1       0.23      0.69      0.34        61

    accuracy                           0.64       441
   macro avg       0.58      0.66      0.55       441
weighted avg       0.83      0.64      0.69       441
```

Inference:

- The model shows a precision, recall and f1-score of 1 on train data. Accuracy is also 1 which is impossible for a model to have.
- This clearly means that Random Forest is unable to learn the new model from SMOTE balancing.
- The test data gives an accuracy of 63.71% which indicates a difference of 40% between train and test models. This clearly means my model is a failed model and cannot be implemented.
- Train and Test precision, recall and F1-score also shows a huge difference which means my predictions have been maximum false. Model predicted No attrition when it was actually attrition.

From this we can conclude that we need to use better and faster classification models like KNN, ADA boost, Gaussian as these are quick learners and penalise bad learners of the previous data.

Trying Random Forest with a 60:40 split to see variations in the result:

Train Data

```
1.0
[[729   0]
 [  0 729]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       729
           1       1.00      1.00      1.00       729

    accuracy                           1.00      1458
   macro avg       1.00      1.00      1.00      1458
weighted avg       1.00      1.00      1.00      1458
```

Test Data

```
1.0
[[504   0]
 [  0  84]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       504
           1       1.00      1.00      1.00        84

    accuracy                           1.00       588
   macro avg       1.00      1.00      1.00       588
weighted avg       1.00      1.00      1.00       588
```

Inference:

- The model gives us 100% accuracy in both train and test which is unacceptable.
- However, the 60 and 40 splits for the data seems practical as model was over sized and had less attrition so test on attrition should be more.

MODEL 3:

```
AdaBoostClassifier
AdaBoostClassifier(n_estimators=100, random_state=1)
```

- We have taken default value of N_estimator which is 100.
- To obtain a deterministic behaviour during fitting, random_state has to be fixed to 1 to maintain last and first value of the data.
- The model gives us an accuracy of 84.03% which is a good model.

Train Data

```
0.9320987654320988
[[688  41]
 [ 58 671]]
              precision    recall  f1-score   support

           0       0.92      0.94      0.93       729
           1       0.94      0.92      0.93       729

    accuracy                           0.93      1458
   macro avg       0.93      0.93      0.93      1458
weighted avg       0.93      0.93      0.93      1458
```

Test Data

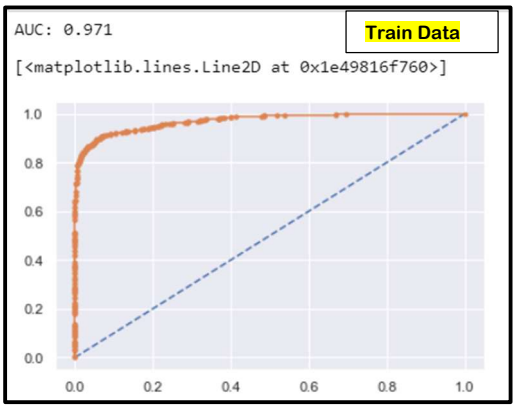
```
0.8894557823129252
[[483  21]
 [ 44  40]]
              precision    recall  f1-score   support

           0       0.92      0.96      0.94       504
           1       0.66      0.48      0.55        84

    accuracy                           0.89       588
   macro avg       0.79      0.72      0.74       588
weighted avg       0.88      0.89      0.88       588
```

Inference:

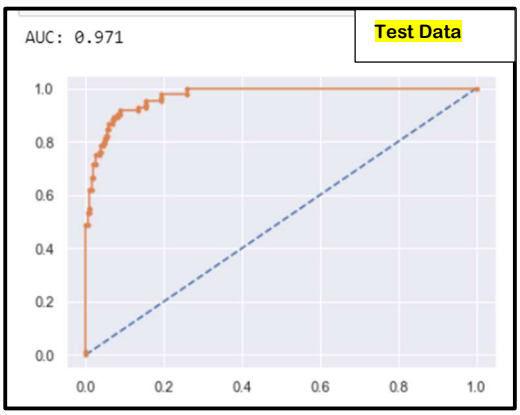
- The train and test model seems to be a good model as train accuracy (93.20%) to test accuracy (88.94%) is 3% at par.
- We have 688 observed zeros and 671 observed ones in our train data which gives us 94% accuracy. False negative is 58 and false positive is 41 in the train data.
- Out of 729 (671+58) we correctly identified 671 which means our model is a good predictor of 0 as 0 and 1 as 1. Recall which is why gives us 92% as 671 out of 729 were predicted correctly.
- 92% is the capability of the model to predict a category correctly.
- The test data is also having values near to train data in the row of 0 where precision is 92%, recall is 96%.
- The model is a good model.

AUC ROC CURVE

Train Data



Test Data



Inference:

- Both AUC for test and train data is near to one so it is a good model.
- The ROC curve or the red line is also near to Y axis which means it is closer to true positive rate. The predictions made by the model is true maximum times. However, in test data as we saw in the confusion matrix the prediction of false is more than true so the ROC curve is wavy in the test ROC CURVE.

MODEL 4:

```
▾ LogisticRegression
LogisticRegression()
```

- The model gives us an accuracy of 66.8% which is low for a model.

Train Data

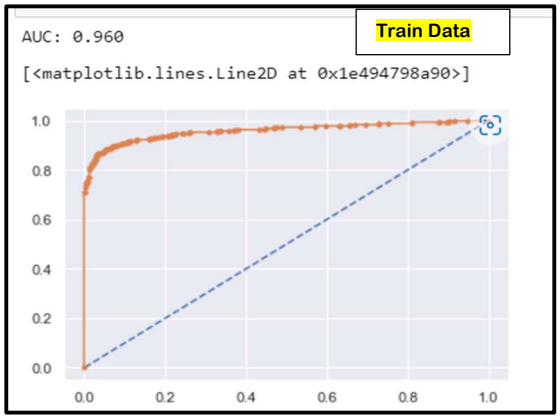
```
0.7139917695473251
[[518 211]
 [206 523]]
              precision    recall  f1-score   support

           0       0.72      0.71      0.71       729
           1       0.71      0.72      0.71       729

    accuracy                           0.71      1458
   macro avg       0.71      0.71      0.71      1458
weighted avg       0.71      0.71      0.71      1458
```

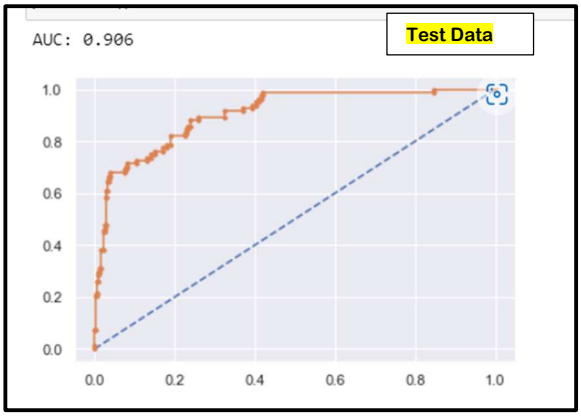
Test Data

```
0.6683673469387755
[[347 157]
 [ 38  46]]
              precision    recall  f1-score   support

           0       0.90      0.69      0.78       504
           1       0.23      0.55      0.32        84

    accuracy                           0.67       588
   macro avg       0.56      0.62      0.55       588
weighted avg       0.80      0.67      0.71       588
```

Inference:

- The model shows an accuracy of 71.4% in train data and 66.83% in test data which is said to be a good model, since difference isn't more than 10% between two.
- The model with a precision value of 71% means it predicts 71% of the 0 as 0 and 1 as 1.
- The recall score indicates 72% of the total values predicted the model as predicted true as true and false as false.
- The train and test data seems fairly different for f1 score, recall score and precision.





Inference:

- Both AUC for test and train data is near to one so it is a good model with 6% difference.
- The ROC curve or the red line is also near to Y axis which means it is closer to true positive rate. The predictions made by the model is true maximum times. However, in test data as we saw in the confusion matrix the prediction of false is more than true so the ROC curve is wavy in the test ROC CURVE.

MODEL 5:

```
GaussianNB
GaussianNB()
```

- The model gives us an accuracy of 64.28% which is too low for a model.

Train Data

```
0.7510288065843621
[[460 269]
 [ 94 635]]
              precision    recall  f1-score   support

           0       0.83      0.63      0.72       729
           1       0.70      0.87      0.78       729

    accuracy                           0.75      1458
   macro avg       0.77      0.75      0.75      1458
weighted avg       0.77      0.75      0.75      1458
```

Test Data

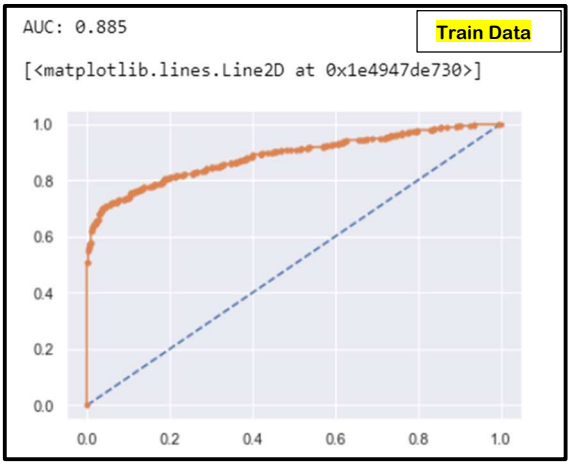
```
0.6428571428571429
[[327 177]
 [ 33  51]]
              precision    recall  f1-score   support

           0       0.91      0.65      0.76       504
           1       0.22      0.61      0.33        84

    accuracy                           0.64       588
   macro avg       0.57      0.63      0.54       588
weighted avg       0.81      0.64      0.70       588
```

Inference:

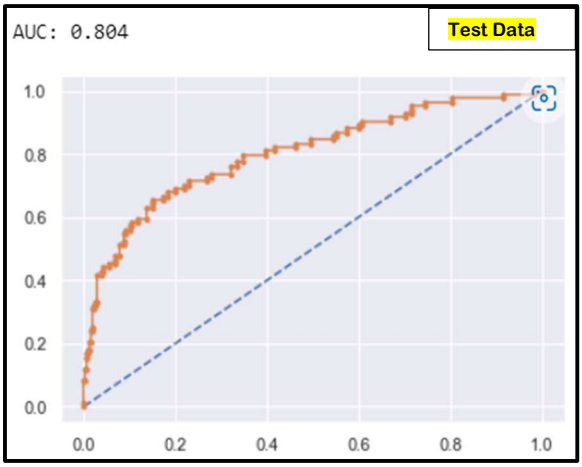
- The model shows an accuracy of 75.10% in train data and 64.28% in test data which is said to be a good model, since difference is little more than 10% between two.
- The model with a precision value of 70% means it predicts 70% of the 0 as 0 and 1 as 1.
- The recall score indicates 87% of the total values predicted the model as predicted true as true and false as false.
- The train and test data seems fairly different for f1 score, recall score and precision. Hence, we will not rely on this model as chances of predicting No attrition is less and Attrition is more.

AUC ROC CURVE

Train Data



Test Data



Inference:

- AUC for train data is 88.5% and test data is 80.4% which is perfect a model.
- The test data seems to have a wavy red curve for ROC which is a bit far from True positive rate.
- The Train seems to be perfect.

- **MODEL 6:**

```
XGBRFClassifier
XGBRFClassifier(base_score=0.5, booster='gbtree', callbacks=None,
XGBRFClassifier
               colsample_bylevel=1, colsample_bytree=1,
               early_stopping_rounds=None, enable_categorical=False,
               eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
               importance_type=None, interaction_constraints='', max_bin=256,
               max_cat_to_onehot=4, max_delta_step=0, max_depth=6,
               max_leaves=0, min_child_weight=1, missing=nan,
               monotone_constraints='()', n_estimators=100, n_jobs=0,
               num_parallel_tree=100, objective='reg:squarederror',
               predictor='auto', random_state=0, reg_alpha=0,
               sampling_method='uniform', scale_pos_weight=1, ...)
```

- The model gives us an accuracy of 80.95% which is too low for a model.

Train Data

```
0.9622770919067215
[[699  30]
 [ 25 704]]
              precision    recall  f1-score   support

           0       0.97      0.96      0.96       729
           1       0.96      0.97      0.96       729

    accuracy                           0.96      1458
   macro avg       0.96      0.96      0.96      1458
weighted avg       0.96      0.96      0.96      1458
```

Test Data

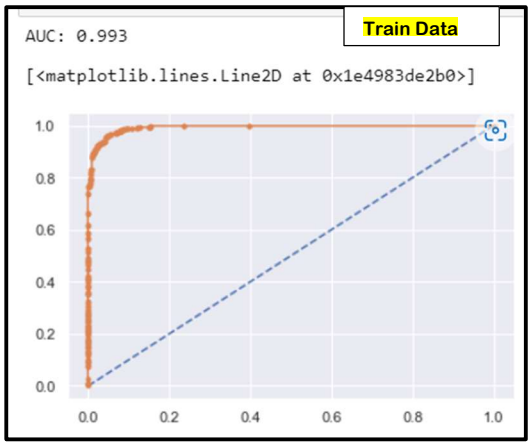
```
0.8095238095238095
[[452  52]
 [ 60  24]]
              precision    recall  f1-score   support

           0       0.88      0.90      0.89       504
           1       0.32      0.29      0.30        84

    accuracy                           0.81       588
   macro avg       0.60      0.59      0.59       588
weighted avg       0.80      0.81      0.81       588
```

Inference:

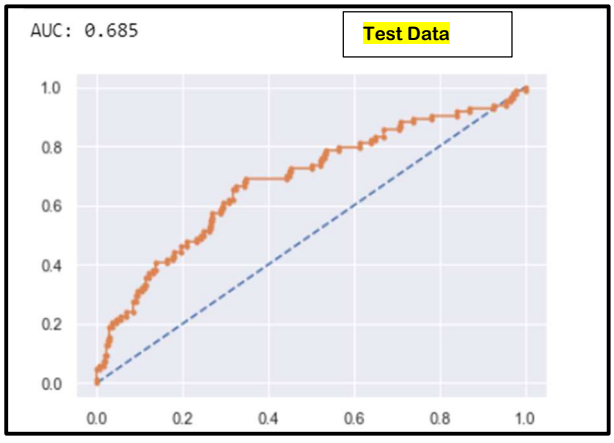
- The model shows an accuracy of 96.23% in train data and 80.95% in test data which is said to be a good model, since difference not more than 10% between two.
- The model with a precision value of 96% means it predicts 96% of the 0 as 0 and 1 as 1.
- The recall score indicates 97% of the total values predicted the model as predicted true as true and false as false.
- The difference between train and test precision is also nearby hence this can be a model we can apply.

AUC ROC CURVE

Train Data



Test Data



Inference:

- AUC for train data is 99.3% and test data is 68.5% which is far more than 10% difference.
- We won't go ahead and apply this model as test data is having unpredictability (red line is squiggling) so wrong predictions can be made. The blue line in test data is the AUC curve which is almost aligning with ROC curve hence false prediction is having high chances.

MODEL 7:

```
▾ KNeighborsClassifier
KNeighborsClassifier()
```

- The model accuracy is 63.23% which is too less.

Train Data

```
0.8257887517146777
[[524 205]
 [ 49 680]]
              precision    recall  f1-score   support

           0       0.91      0.72      0.80       729
           1       0.77      0.93      0.84       729

    accuracy                           0.83      1458
   macro avg       0.84      0.83      0.82      1458
weighted avg       0.84      0.83      0.82      1458
```

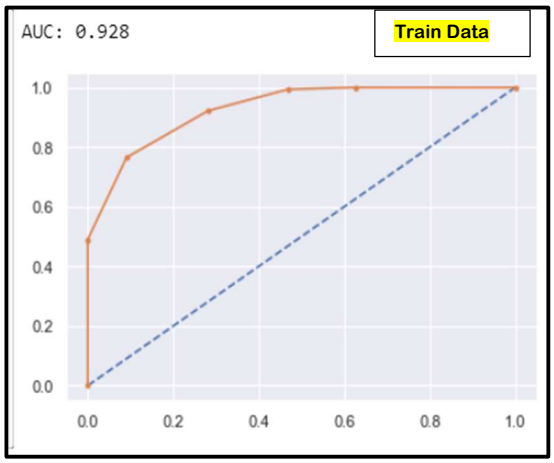
Test Data

```
0.6326530612244898
[[334 170]
 [ 46  38]]
              precision    recall  f1-score   support

           0       0.88      0.66      0.76       504
           1       0.18      0.45      0.26        84

    accuracy                           0.63       588
   macro avg       0.53      0.56      0.51       588
weighted avg       0.78      0.63      0.68       588
```

Inference:

- The model shows an accuracy of 82.57% in train data and 63.26% in test data which is said to be a bad model, since difference more than 10% between two.
- The model with a precision value of 77% means it predicts 77% of the 0 as 0 and 1 as 1.
- The recall score indicates 93% of the total values predicted the model as predicted true as true and false as false.
- The train and test data seems widely different for f1 score, recall score and precision. Hence, we will not rely on this model as chances of predicting No attrition is less and Attrition is more.

AUC ROC CURVE



Inference:

- AUC for train data is 92.8% and test data is 59.9% which is more than 10% difference.
- We won't go ahead and apply this model as test data is having unpredictability (red line is squiggling) so wrong predictions can be made. The blue line in test data is the AUC curve which is almost aligning with ROC curve hence false prediction is having high chances.

MODEL INFERENCE

Model	Accuracy	Train	Test	AUC_Train	AUC_Test
AdaBoost	84.03%	93.02%	88.94%	97.10%	97.10%
Logistic Regression	66.80%	71.39%	66.84%	96.00%	90.60%
Gaussian NB	64.28%	75.10%	64.28%	88.50%	80.40%
XGBRF	80.95%	96.23%	80.95%	99.30%	68.50%
KNN	63.23%	82.57%	63.26%	92.80%	59.90%

Inference:

- Highest Accuracy is of Adaboost and XGBRF model.
- Comparing the train and test for all models we see Adaboost again as the best. Logistic Regression is also having a good score but since accuracy matters, we select Adaboost.
- The precision, recall and F1 was also near about to each other in train and test for Adaboost hence we again confirm this model being implemented.
- AUC and ROC for train and test were near to true positive prediction with little variance in both results.

Adaboost is the model we will implement.

Lets see the variables which are having more importance in this model:

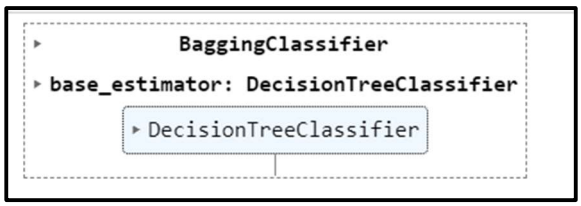
	Imp
JobSatisfaction	0.097632
Department_Research_and_Development	0.085118
StockOptionLevel	0.074661
YearsAtCompany	0.067489
EnvironmentSatisfaction	0.055992
TotalWorkingYears	0.054309
JobLevel	0.048060
JobInvolvement	0.047688
MonthlyIncome	0.039701
YearsInCurrentRole	0.035953
TrainingTimesLastYear	0.034155
RelationshipSatisfaction	0.028674
BusinessTravel_Travel_Rarely	0.028516
YearsWithCurrManager	0.027493
Age	0.024940
EducationField_Medical	0.023492
EducationField_Life_Sciences	0.022823
YearsSinceLastPromotion	0.017820
MaritalStatus_Married	0.017569
Department_Sales	0.016355
JobRole_Research_Scientist	0.016286
WorkLifeBalance	0.015437
MonthlyRate	0.015244
MaritalStatus_Single	0.015006
DistanceFromHome	0.014955
HourlyRate	0.010420
DailyRate	0.009104
NumCompaniesWorked	0.007421
PercentSalaryHike	0.006330
Gender_Male	0.005814
Education	0.005498
JobRole_Sales_Executive	0.005332
EducationField_Marketing	0.004531
OverTime_Yes	0.003184
PerformanceRating	0.003072
BusinessTravel_Travel_Frequently	0.003049
JobRole_Manufacturing_Director	0.002575
JobRole_Laboratory_Technician	0.002047
JobRole_Manager	0.001935
EducationField_Technical_Degree	0.001491
JobRole_Sales_Representative	0.001270
JobRole_Human_Resources	0.000854
EducationField_Other	0.000469
JobRole_Research_Director	0.000235

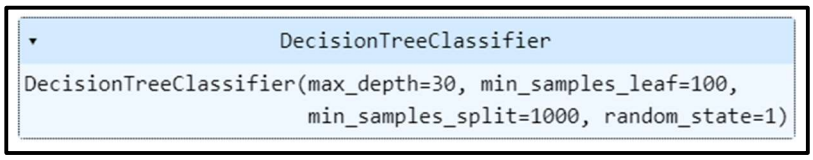
Inference:

- Job satisfaction contributes majority to Attrition.
- Least attrition is contributed but job_role_Director.
- The model is more impacted by job satisfaction however none of them will we remove and perform modelling again as all are important contributors.

5) Model Bagging and Parameter Tuning

Bagging helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model.





- Max_depth is the maximum depth of the tree which has been taken as 30. The max depth is in the data until all nodes are broken till pure or is less than min_samples_split

- N_estimator is the number of trees in the forest with random state 1 which is only to maintain the purity of data.

- Min_Sample_leaf-The minimum number of samples required to be at a leaf node.

- Min_samples_split=The minimum number of samples required to split an internal node. The important features of the training and the testing model is:

Train Data

```
1.0
[[729   0]
 [  0 729]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       729
           1       1.00      1.00      1.00       729

    accuracy                           1.00      1458
   macro avg       1.00      1.00      1.00      1458
weighted avg       1.00      1.00      1.00      1458
```

Test Data

```
0.858843537414966
[[480  24]
 [ 59  25]]
              precision    recall  f1-score   support

           0       0.89      0.95      0.92       504
           1       0.51      0.30      0.38        84

    accuracy                           0.86       588
   macro avg       0.70      0.62      0.65       588
weighted avg       0.84      0.86      0.84       588
```

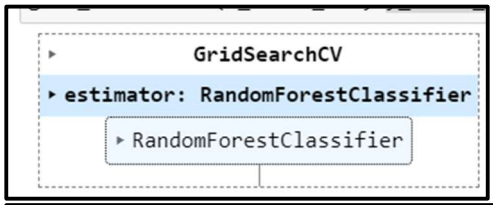
Inference:

- We tried to drop relevant feature in Adabooster and perform bagging which didn't prove well.
- The test and train data has a relevant accuracy of 100% to 85%.
- Precision is also 100% which is comparatively near to test data of 89%.
- However, bagging is not working well with decision tree and we should take up some other model to reduce variance in test and train recall, precision and f1-score.

  Concluding to say we will use all parameters as in the ADABOOST without removing any feature because an accuracy of 100% means over fitting most of the time and is impossible to run a love model with 100% accuracy.

Parameter Tuning

Since we saw decision tree having a draw back in bagging let's try parameter tuning.




```
best_grid
RandomForestClassifier(max_depth=10, max_features=6, min_samples_leaf=10,
                       min_samples_split=50, n_estimators=300, random_state=1)
```

- Accuracy scores: We see 91% of accuracy in train and 82% in test data.(70:30 split)

0.9087791495198903	0.8299319727891157
Train Data	Test Data

- Accuracy scores: We see 91% of accuracy in train and 85% in test data.(60:40 split)

0.9108367626886146	0.8571428571428571
Train Data	Test Data

- The random forest earlier gave us 100% in train and 80% in test but post applying the best grid we see a fair result which is 90% train and 82.9% in test.
- The variables of importance for the random forest is from descending to ascending impacting our model is:

	Imp		
JobSatisfaction	0.099338		
StockOptionLevel	0.098116		
Department_Research_and_Development	0.064425		
JobLevel	0.059732		
JobInvolvement	0.057320		
MonthlyIncome	0.048689		
BusinessTravel_Travel_Rarely	0.041196		
EnvironmentSatisfaction	0.038900		
RelationshipSatisfaction	0.037890		
Age	0.033495		
MaritalStatus_Married	0.033114		
TotalWorkingYears	0.031895		
EducationField_Medical	0.029013		
YearsInCurrentRole	0.027399		
YearsAtCompany	0.026941		
TrainingTimesLastYear	0.026544		
YearsWithCurrManager	0.023339	JobRole_Manufacturing_Director	0.006814
EducationField_Life_Sciences	0.023325	OverTime_Yes	0.006035
YearsSinceLastPromotion	0.021463	JobRole_Sales_Executive	0.005670
WorkLifeBalance	0.016003	JobRole_Laboratory_Technician	0.003091
MonthlyRate	0.015221	EducationField_Technical_Degree	0.002896
Education	0.013594	EducationField_Marketing	0.002605
DistanceFromHome	0.013512	JobRole_Manager	0.002563
MaritalStatus_Single	0.012847	BusinessTravel_Travel_Frequently	0.002465
JobRole_Research_Scientist	0.012468	PerformanceRating	0.002110
DailyRate	0.012352	EducationField_Other	0.001140
HourlyRate	0.012349	JobRole_Human_Resources	0.000999
Department_Sales	0.009910	JobRole_Sales_Representative	0.000783
PercentSalaryHike	0.007766	JobRole_Research_Director	0.000150
Gender_Male	0.007450		
NumCompaniesWorked	0.007072		

6) Recommendations

- As we see more attrition at the age of 31-35 and males have more attrition so reason can be asked. Some may be for personal choice and some for growth. It would be easier to retain if the next survey has a column called reason for leaving.
- Attrition rate in overtime is more so a deep dive into increasing FTE or dividing work of the employee should be done. May be ways to manage time properly can also help the employee to complete work on time.
- People are tending to switch to a different job at the start of their careers, or at the earlier parts of it. Once they have settled with a family or have found stability in their jobs, they tend to stay long in the same organization- only going for vertical movements in the same organization.
- Where business travel is present more attrition is seen so company can think of ways to minimize travel. Travel allowances should be revised if that is the case if employee attrition is happening. Employee might travel more than the allowance covers.
- Research and Development team has maximum attrition for laboratory technician. More employee engagement and leadership connect should be done to understand the exact issue behind the attrition.
- Since single status employees have more attrition, we can utilise them more and engage them in a productive way that they do not look for switch. As we see age is 31-35, we have young crowd who can be given trainings and opportunities to explore.
- Stock option and salary have an important role in the employees benefit interest with the company. More employees should be given stock option and salary hike to drive retention as to loyalty.
- With increase in salary also work life balance should be accounted for. As we see over time and work life balance has attrition more employees should be trained to do a specific task or timelines should be made flexible for work delivery.
- People should move around in the organization as under the same manager and department sees an attrition at a higher level. More exploring opportunities inside lesser will an employee look for growth outside. Also, the work environment issue resolves since they get new people to work with.
- Most attrition rate is for education with 3 rating. People can be given higher education benefits or opportunities with organizational tie ups with Educational Institutions.