**PROJECT REPORT**

**Employee Attrition Analytics**

***Submitted By:***

**RATAN ERABATHINI**

**VIVEK PATTEBAHADUR**

**YADNESH KOLHE**

**SWAPNEEL INGLE**

***Course and Batch****: Data Science Program-DSP WD2-Thane*



**Acknowledgements**

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the DSP program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

**Certificate of Completion**

I hereby certify that the project titled “Employee Attrition Analytics” was undertaken and completed under my supervision by Group 1 from the batch of Data Science Program DSP-WD2-Thane

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Index**

|  |  |  |
| --- | --- | --- |
| **Sr. no.** | **Topic** | **Page no.** |
| 1 | Understand the problem and objective | 5 |
| 2 | Understand the data and develop some business sense | 6 |
| 3 | Exploratory Data Analysis | 8 |
| 4 | Data Cleaning and Preparation | 12 |
| 5 | Feature engineering | 13 |
| 6 | Model building | 14 |
| 7 | Prediction on Sample Data | 15 |
| 8 | Conclusion | 17 |
|  |  |  |

**Understand the problem and Objective:**

* Employee Attrition is a huge problem across industries and generally costs the company a lot for hiring, retraining, productivity and work loss for each employee who leaves.
* To understand weather an employee quits at the earliest, we need to build model to predict whether an employee would quit in the near future**.**

**Objective :**

The objective of this project is to build a data model which will

* Identify the factors influencing attrition
* Predict possible attrition
* Identify possible ways to retain high performers.

**Data Source :**

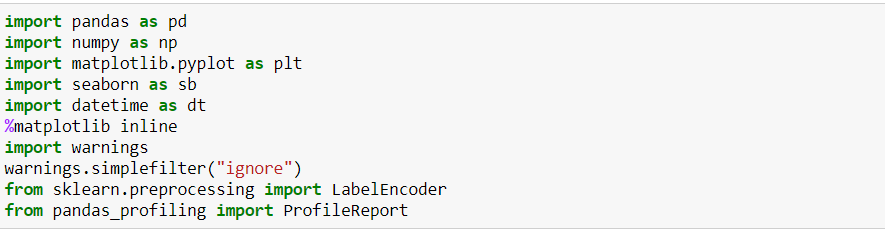
* We were provided with the three datasets along with the problem statement.
* The two datasets contained records of the employees about “Profit Center” , ”Employee Level” ,’’Employee Location” and other details of the employees for two years 2016-2018.
* The third dataset ‘Terminations’ contained all the details about resigned employees and the factors related to attrition.

**Tools and Techniques :**

* We have used Python using Jupiter notebook

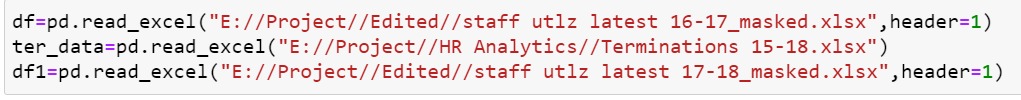
**Data Preparation and Understanding:**

To get deeper understanding of data or to apply different models to the datasets we need to prepare our data first, so that we can extract the data to read, finding actual predictors and perform visualization part also. Here we performed few steps to get the full knowledge

**1) Importing Modules**

**2) Read the Datasets :**

Now we are reading the datasets into object df, df1 and ter\_data.



**3) Concatenating & Merging the datasets**:

* We are selecting the particular rows or columns from data set  df & df1.
* After selection of particular rows or columns. we concatenating the data set & there is some duplicate values are shows. We tackled it and drop the particular rows
* After we concatenating our data set contains 13 features relating to 1111 observations.

**4) Creating New columns**:

* Using our dataset we created few useful columns,

1. **Attrition:-**

* we created Attrition columns based on current Status of employee & used as the target variable for the further analysis .
* Current states column contains Active, Resigned, Secondment, New joiner, Sabbatical.
* We mapping the Current states column values Active ,New Joiner & Sabbatical as “**No**” & Resigned, Secondment as “**Yes**” & creates new columns as Attrition.

1. **Tenure:-**

* Employee tenure is the length of the employee works for an employer or company.
* We are calculating tenure from column month year termination date subtracting to Joined date.

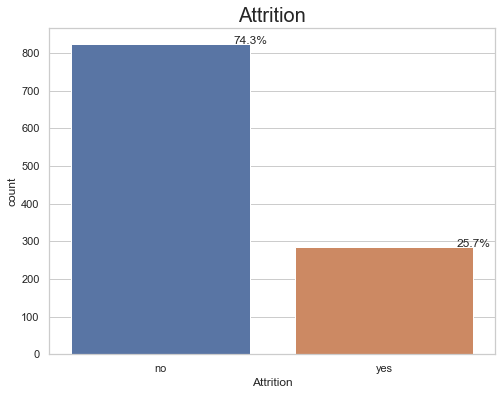
1. **Appraisal :-**

* Employee appraisal process is crucial for organizations to boost employee productivity and improve their outcomes.
* We have two year of data set. date set contain employee position columns .so we compared employee position of last year and present year and created a new column as appraisal
* In appraisal columns we created levels as promotion, demotion ,resigned and not promoted.

**Exploratory Data Analysis:**

1. **Attrition of employee :-**

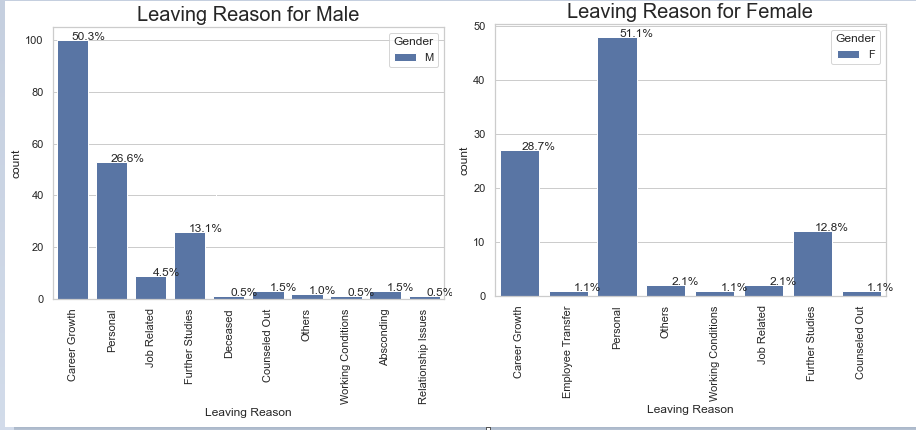
As per below graph, the employee attrition is approximate 25 %



1. **Attrition : Gender vs Leaving Reasons**

As per below graph analyzed the below pointers:

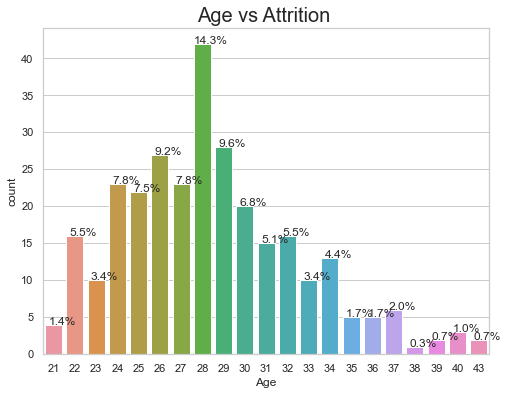
* 50.3 % Men’s & 28.7% female’s lefts for the Career growth .
* 51.7 % Women’s & 26.6% male’s lefts for Personal reasons.



1. **Age vs Attrition :**

As per below graph we analyzed the below pointers

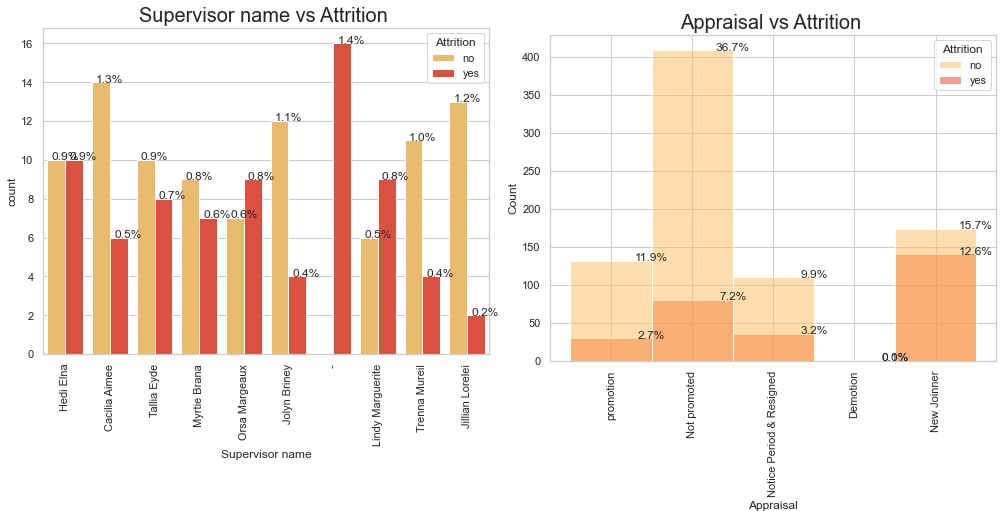
* 65 to 75 % employees lefts the company below age of 30 .
* Max employee left the company at the age of 28 i.e 14.3%



1. **Supervisor name vs Attrition & Appraisal vs Attrition:**

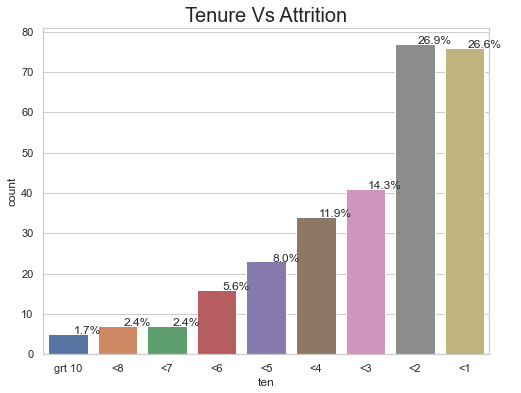
As per below graph analyzed the below pointers :

* Most Employee are leaving under supervisor name Orasa Margeaux and lindy marguerite with highest rate of attrition compared to others.
* If employee got Demotion , 40-45% chances he/she left the company.



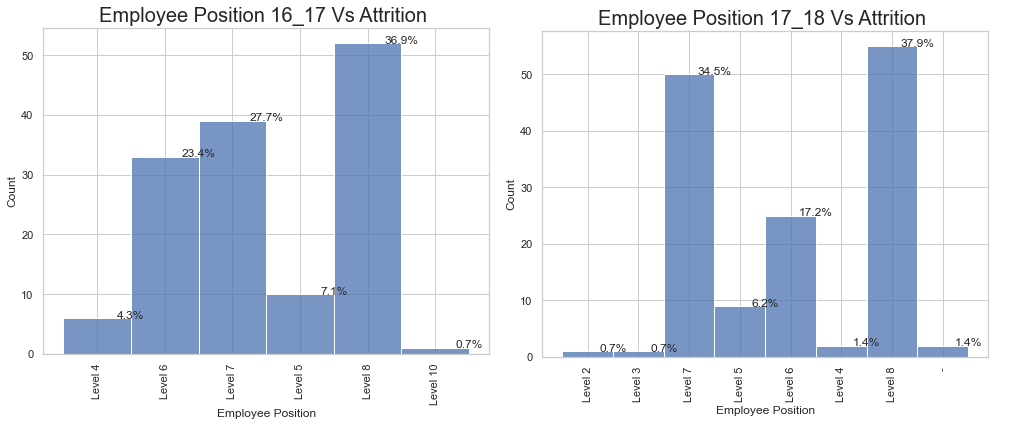
1. **Tenure vs Attrition :**

* Most of the Employee are leaving the company who worked less than 2 years.
* 26.6% of the Resigned Employee falls under less than 1 year of working in company.
* 26.9% of the Resigned Employee falls under less than 2 year of working in company.

****

1. **Employee Position 2016-17 & 2017-2018 vs Attrition :**

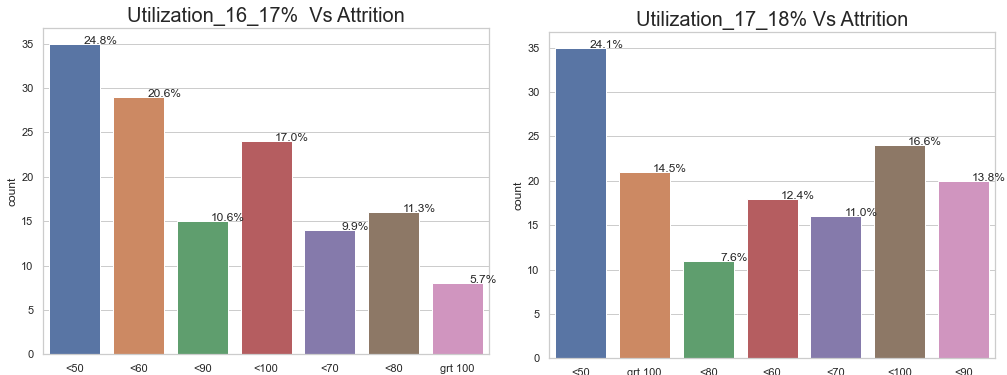
* In 2016-17 and 2017-2018, **Level 8** and **level 7** showing Most of the Employees leaving the Company.
* **36.9% and 27.7%** of the Resigned Employees falls on Employee position of level 8 and level 7 Respectively in 2016-17.
* **37.9% and 34.5%** of the Resigned Employees falls on Employee position of level 8 and level 7 Respectively in 2017-18.

****

1. **Utilization 2016-17 & 2017-2018 vs Attrition :**

As per below graph , analysed the below points for only Resigned data:

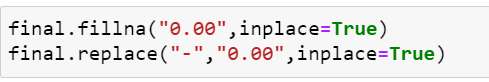
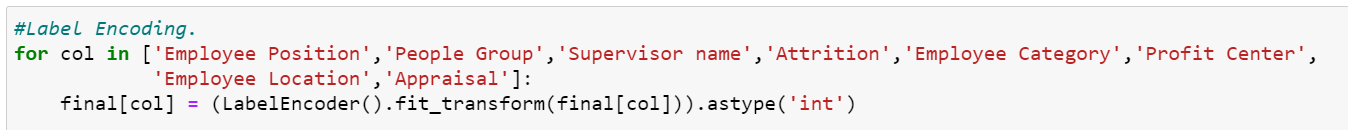
* Most Employee are leaving below 50% & 60% of Utilization in year 2016-17.
* Most Employee are leaving below 50% of utilization in year 2017-18.
* Analysing the following graph, Employee having greater than 90% of utilization must be Retained

****

**Data Cleaning and Preparation :**

1. We are removing the columns which are not important for our employee attrition analysis. The columns that are removed are listed below

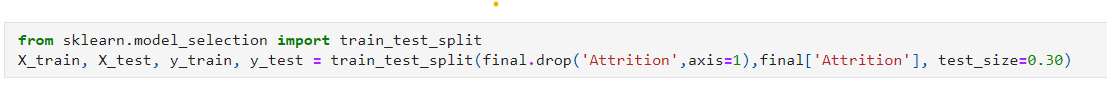
* Employee No
* Employee Number
* Join Date
* Last Update Date
* Emp Ref.
* Employee Name
* Latest Available Rating
* YEAR of Birth
* Termination Date
* Current status
* People group

1. We have some garbage data which is available as ‘NaN’ and ‘-‘. We have replaced them as ‘zero’
2. We done a Label Encoding for categorical columns

**Feature engineering**

Feature engineering has two goals primarily:

* Preparing the proper input dataset, compatible with the machine learning algorithm requirements
* Improving the performance of machine learning models

****

* After pre-processing, we split our data into training and testing dataset. From a total of 1111 observations.
* Splitting the dataset into training and testing dataset w.r.t. attrition column.
* 70% observation for the training dataset.
* 30% observation for the testing dataset

**Model building :**

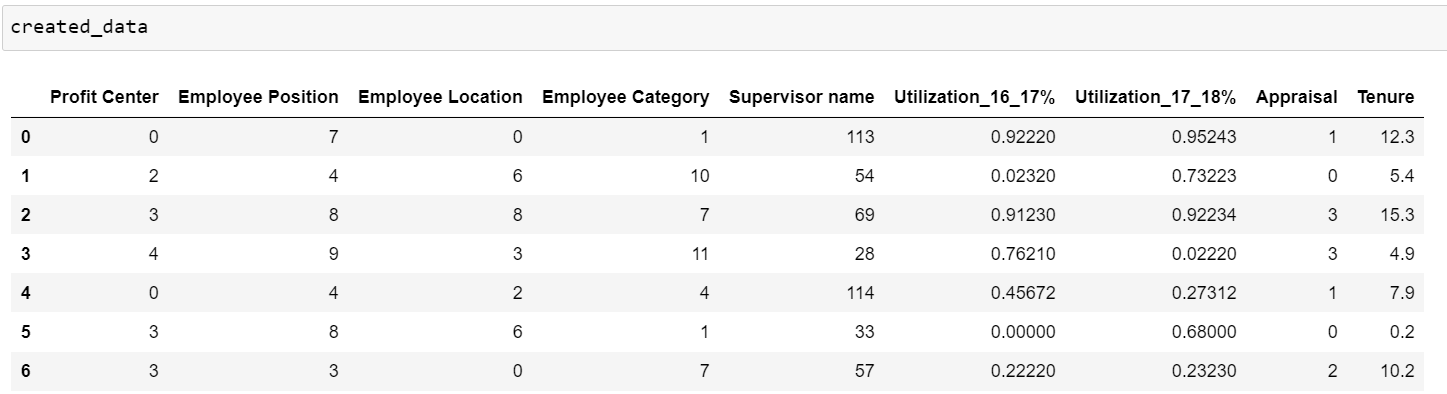
The following table shows which of data models performed the best :

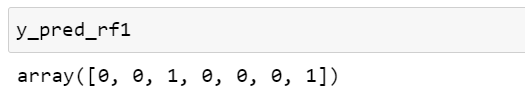
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Precision** | **Recall** | **F1-score** | **Accuracy** |
| Logistic Regression | 0.76 | 0.70 | 0.73 | 0.85 |
| KNN | 0.90 | 0.86 | 0.88 | 0.93 |
| Random Forest | 1 | 0.98 | 0.99 | 0.99 |
| Decision Tree | 0.98 | 0.95 | 0.96 | 0.98 |
| SVM(Classification) | 0.97 | 0.61 | 0.75 | 0.88 |

**RANDOM FOREST**

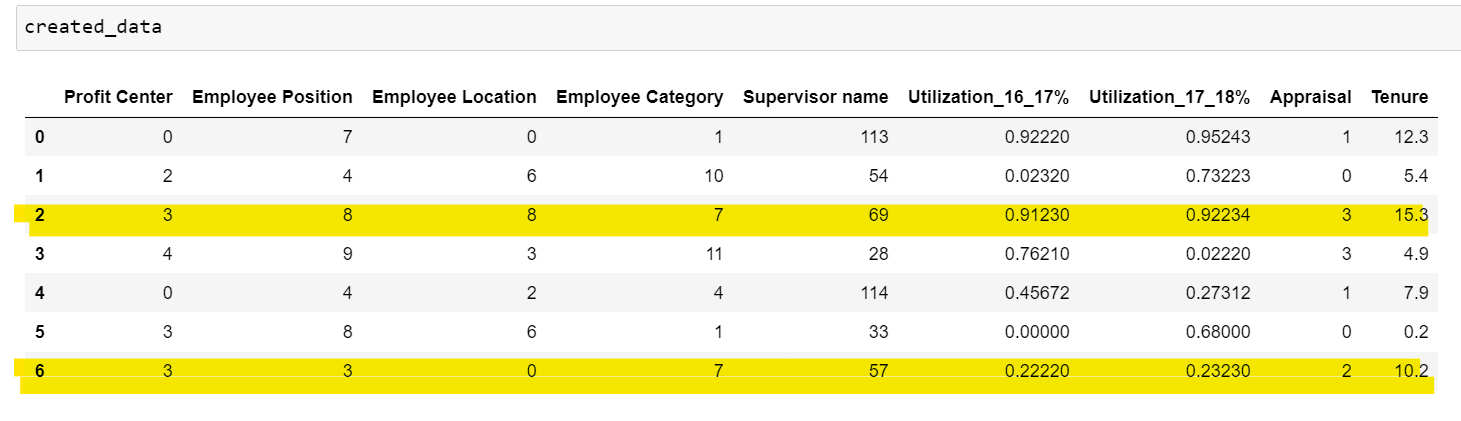
* We applied Random Forest on the final dataset with best parameters,
* Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
* Random forest gives us more than 99% accuracy.

**Prediction on Sample Data:**

* We have created a Dummy data for testing the model for Predicting Possible Future Attrition.
* We have assigned random values for the columns as shown in the below table.
* We have taken Random forest as our best model to predict the attrition on the created Dummy data.
* We have predicted the possible outcomes of Employee Attrition on created Random Forest model.
* Following is the result Showing “0” as Active Employee and “1” as Resigned Employee for Future attrition on Dummy data.



* As per above predicted outcome following index no 2 & 6 showing future attrition shall be resigning.
* Index no 2, Employee has above 90% Utilization so we can Retain the Employee on the basis of Employee best performance.
* Index no 6, Employee has only 23% Utilization so if the Employee is leaving, than the Company must not consider to retain, as Employee has poor performance.



**Conclusion:**

From the plots showing employee positions and leaving reason, we can conclude that the maximum resigned employees falls in low level positions as well as their common leaving reasons is career growth and Personal reason. Maximum Employee having experience less than 2 years are prone to leave their job. Maximum employee left the company at the age between 25 to 30.

There are some key indicators which helps to predict that the Employee is Profitable for the company or not. So on the basis of 80% or Above Utilization of the Employee are retainable.