PROJECT REPORT

ON

EMPLOYEE ATTRITION

Submitted By:

Sherin Joseph

27/ NOV / 2021

# INTRODUCTION

The employee attrition project is conducted on the observational data of the employees in a company. The company uses the observational data to analyse and interpret where highest attrition occurs. They would like to know which category of employees are more subjected to attrition and the reasons behind employees leaving the firm. The meaningful insights derived from the data would enable the company to take necessary actions to deal with employees in a much efficient manner and reduce attrition rate in future.

# DATASET EXPLORATION

The first and foremost step in the data analysis project is to understand the data and then explore the data using different python codes and visualization libraries. The independent variables, dependent variable, their distribution, missing values, and other details related to the data is understood from the data exploration. The list of variables is as follows:

* Age : Age of the employees
* Attrition : Status of employees subjected to attrition or not
* BusinessTravel : Whether the employee have done business travel
* Department : Department of employee
* DistanceFromHome : Distance for employees’ home to company
* Education : Education of employees
* EducationField : Field of employee education
* EmployeeID : Unique ID of employee
* Gender : Gender of employee
* JobLevel : Level of job
* JobRole : Job role of employee
* MaritalStatus : Marital status of employee
* YearlyIncome : Annual income
* NumCompaniesWorked : Number of companies worked
* PercentSalaryHike : Previous salary hike percentage
* StockOptionLevel : Level of stock in the company share
* TotalWorkingYears : Total number of working years
* TrainingTimesLastYear : Number of trainings attended last year
* YearsAtCompany : Number of working years with the company
* YearsSinceLastPromotion : Number of years since last promotion
* YearsWithCurrManager : Number of years under same manager

Graphical user interface, text, application

Description automatically generated

There are 4410 observations and 24 columns in data. However, there are no duplicate observations in the dataset.

# FRAMEWORK

The Independent variables in this study are Age, BusinessTravel, Department, DistanceFromHome, Education, EducationField, Gender, JobLevel, JobRole, MaritalStatus, YearlyIncome, NumberOfCompaniesWorked, PercentSalaryHike, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsSinceLastPromotion and YearsWithCurrManager.

The Dependent variable for the analysis is **Attrition**.

# DATA PREPROCESSING

**Handling Missing Values:**

There were some missing values in the 'TotalWorkingYears'. They are replaced with the value in corresponding 'YearsAtCompany' field instead of 0 as though this company is the employees first company. The missing values in the 'NumCompaniesWorked' is replaced with the mode values corresponding to the 'JobLevel' and 'Education' of the respective employees.

**Excluding Irrelevant Columns:**

There are certain columns that do not have significance to the Attrition of the employees. They are Gender, JobRole, DistanceFromHome, YearlyIncome, PercentSalaryHike, EmployeeCount, Over18 and StandardHours. A chi-square summary table is created for this purpose.

Table

Description automatically generated

**Encoding:**

The relevant categorical variables are encoded using dummy variable encoding.

# DATA VISUALIZATION

The data visualization is the most important step because it gives the stakeholders a better understanding of the data as it is presented in the forms of charts and graphs rather than row data format. This enables to understand the hidden patterns, trends and draw meaningful insights from the data. We have utilized several plots such as pie charts, bar charts, grouped bar charts, histograms, grouped histograms, density plots, grouped density plots and box plots for this purpose.

# MODELLING

Our output variable is categoric and hence we have used different modelling techniques such as Logistic Regression, KNN, Random Forest, Adaboost and SVM. A comparison of these models is done and is as follows:

Table

Description automatically generated

From the different models, the Random Forest has the highest accuracy of 99% and hence can be used for the prediction purpose.

# CONCLUSIONS

The following conclusions are made from the data analysis:

* The level of attrition is significantly different across different age groups. Young employees leave company more than seniors.
* The Research and Development department had seen most attrition of its employees compared to other departments.
* The distance travelled by the employees to reach office does not play significant role in attrition.
* The people in field of Life Science and Medical field were more subjected to attrition than others.
* The attrition rate does not depend on the annual income of the employees.
* It is seen that the employees with high tenure years in company have less attrition compared to new staffs.
* The people who got promoted recently tend to leave company more than those who got it a long ago.
* The attrition in the employees who rarely travelled for business purpose was almost double the rate as compared to those who frequently had business travel.
* As the job level increases from junior to manager level, the attrition rate decreases.
* Employees who were single, left the company more compared to married and divorced employees.

# RECOMMENDATIONS

The following recommendations are done made so that the company could take necessary actions and reduce employee attrition in future.

* The employees in the Research and Development department especially from Life Science and Medical field should be given more attention so that they do not leave the company very soon.
* Company could provide the employees more benefits in terms of insurance, pension plans or company shares if they stay with company for more than 10 years. Thus, the chance of attrition can be reduced.
* If legal, company could make some agreements with employees when they get promoted that they would successfully carry out the new roles for a certain time period.