CSCE 5310: Methods in Empirical Analysis
Project Proposal

GitHub Link : https://github.com/adityapujari98/Employee_Attrition_Analysis
Dataset Link : https://www.kaggle.com/datasets/whenamancodes/hr-employee-attrition

Project Title: Employee Attrition Analysis

Team Members :
1. Aditya Pujari - AditiyaPujari@my.unt.edu (11491374)
2. Brinda Potluri - BrindaPotluri@my.unt.edu (11526591)
3. Nitin Dunday Mohan - NitinDundayMohan@my.unt.edu (11515126)
4. Sai Tarun Gunda - saitarungunda@my.unt.edu (11516657)

Idea description:

The primary purpose is to determine whether the existing employee will quit the organization or not. Based on exploratory data analysis, correlation analysis, and binary classification, which can be anticipated. The purpose of implementing the techniques is that we have a huge dataset and need to select features based on our requirement.

Goals and Objectives:

The goals and objectives of this project are as follows.

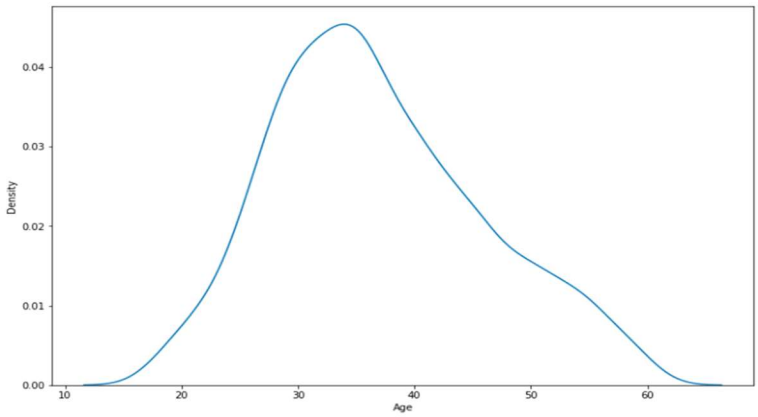
1. Exploratory Data Analysis
2. Model Building
3. Model Evaluation

Exploratory Data Analysis:

In the first step of achieving goals and objectives of this project, we'd be performing Exploratory Data Analysis to discover the patterns and understand the hypothesis of the data. To analyze the data furthermore we'd also plot graphs. As a part of Exploratory Data Analysis we would be performing Univariate Analysis, Bivariate Analysis and Correlation Analysis.

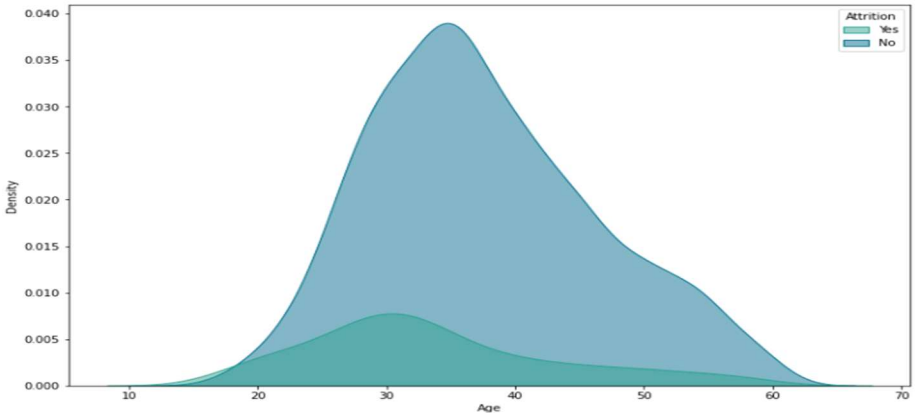
Univariate Analysis:

In Univariate Analysis, we would be considering a single variable in the data in order to discover the patterns of the data. In this analysis, it takes the data and summarizes and finds the patterns in the data. For this project we have observed the density of the variable age. In the below plot you can observe the density is higher in the age groups 30 - 40 which is around 0.05. Density in the age groups less than 15 years and more than 60 years is almost equal to zero. Age groups around 25 years and 45 years are the second highest in terms of density. In this way we have discovered the patterns of the data using univariate analysis.



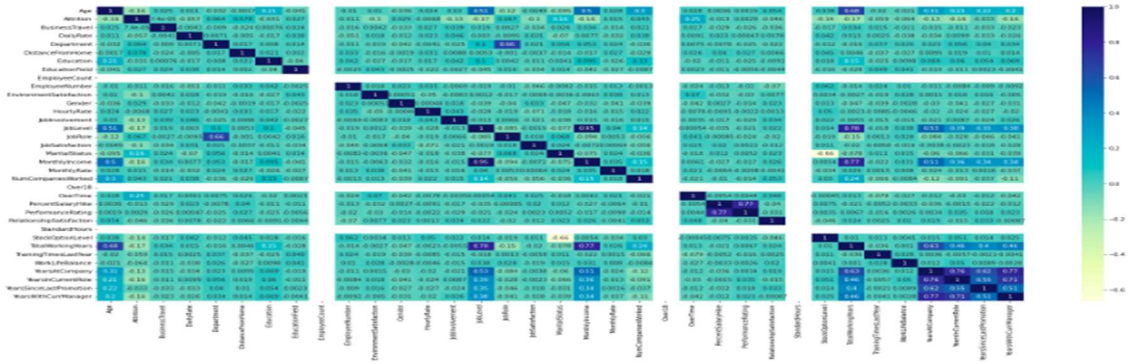
Bivariate Analysis:

In the Bivariate Analysis, a statistical analysis is made on two different variables. Where one variable is dependent on another variable. In this project, we have applied Bivariate Analysis on the variables Age and Attrition. In the below plot you can observe the age group around 40 years has the least density in attrition and the age group around 30 years has the highest density attrition.



Correlation Analysis:

Correlation Analysis is used to test relationships between quantitative variables or categorical variables. It measures how things are related. This correlation is of three types. They are Positive Correlation, Negative Correlation and No Correlation. In this project we have observed a negative correlation.



Model Building:

Once we understand the patterns of the data, we would be building the model. We have chosen Bagging Classifier, Random Classifier and Adaboost Classifier to build the model.

Bagging Classifier:

In Bagging classifier, a meta-estimator fits the base classifier into every subset of the dataset. In the next step, it aggregates the single prediction which eventually forms the final prediction. This meta-estimator is utilized to decrease the variance of the black-box estimator. Later the randomization is introduced into the construction procedure which makes an ensemble out of it.

Random Classifier:

Random classifier is an ensemble method for classification. This is a supervised learning algorithm. It consists of decision trees. For classification the model classifies the data using the random trees. The data would be getting selected as random samples by the model and each tree would choose the best prediction using the random trees.

Adaboost Classifier:

Adaboost classifier is a meta-estimator, it starts with a classifier on the original dataset and then fits extra copies of the classifier on the same dataset. Here the weights of instances which are incorrectly classified are adjusted such that subsequent classifiers focus more on difficult cases.

Model Evaluation:

In Model Evaluation we would be calculating three different measures. They are Precision, Recall and F-1 Score. Also, we would be calculating Accuracy, Macro Average and Weighted average. With all these metrics we would be evaluating the models Bagging Classifier, Random Classifier and Adaboost Classifier.

Motivation

The project was an initiative, since the Employment market is dull and this shows the current picture of how the market would treat an employee. The economy would suffer if unemployment rose. We might investigate the employee's transition from one business to another. The transition of an employee to a new or current function may also be examined.

Significance -

Employers with lower churn rates might spend less on recruiting and related expenses, which can boost the company's profit margin. Having a workforce that is consistently made up of relatively less experienced workers restricts the company because of high employee turnover.
Recruiting and teaching/ training new staff takes a lot of time, effort, and money, and turnover may hurt the company's performance. Numerous issues are brought on by significant personnel turnover, including high expenses, expertise loss, and reduced performance. The cost of replacing an employee can be anything from 16% and 213% of their annual compensation. Organizations in the United States spend about $1 trillion a year on turnover costs.

Literature Survey:

These days, among the biggest problems an Hr department has is employee attrition. Employees expect a variety of comfort levels from the company where they work, including the employer's reputation within the market, pay, growth prospects, working conditions, colleagues, present role's marketability, and most importantly, risk and ensure with the company.

There has been a significant lot of effort done in this field, and each individual or team has developed some solutions. Staff turnover in small-scale scale to medium-scale scale businesses is influenced by a variety of factors, including the work environment, the type of work, company ideology, remuneration, and career advancement (SMEs).

Some researchers came to the conclusion that there was a substantial relationship between organizational commitment and employee turnover, as well as between work satisfaction and the high level of incremental variation in attrition reason.

Features:

Considering the following features,

1. **Age** - Represents the age of the employee
2. **BusinessTravel**- The Employee travel requirement
3. **Department** - Employee working under the Department
4. **EducationField** - The Employee is working Education
5. **Gender** - Represents the Male or Female
6. **HourlyRate** - Represents the Employee pay per hour basis
7. **JobRole** - Represents the the type for job role
8. **MonthlyIncome** - Represents the monthly salary
9. **OverTime** - Represents the employee has worked over time

10. **PercentSalaryHike** - The Employee salary hike
11. **TotalWorkingYears** - Total years an employee has worked in their lifetime.
12. **YearsAtCompany** - Total years spent in the current company
13. **YearsInCurrentRole** - Total years spent in the current role.

Expected outcome:

In this project we would be building a model which takes the random dataset in the form input and classifies the data if the data has employment attrition or not. With this we can analyze and understand the ongoing needs of an employee.

References

- ➔ https://www.business.com/hr-software/employee-attrition/
- ➔ https://www.linkedin.com/pulse/analyzing-employee-attrition-mike-west/
- ➔ https://towardsdatascience.com/using-ml-to-predict-if-an-employee-will-leave-829df149d4f8