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. **Business Travel-** 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.

Project Increment - 1

Related Work (Background):

Numerous staff retention models were introduced as a result of hundreds of research publications investigating the various facets of turnover all through years. The original model, which receives the vast majority of research interest, was proposed in the year 1958 by Mr. Simon. Numerous efforts were numerous attempts to expand the idea since this model.

Personnel are crucial to the operation of any company; without workers, the enterprise would fail. In accordance with Bureau of Labor Data/ Statistics' 2006 report, companies today are discovering that employees stay with them for an average of twenty-three to twenty-four months.

Throughout the USA, the churn rate has increased over the last nine years. The only time that stands out, as would be anticipated, is during the Covid-19 pandemic's initial wave struck and people were unable to quit their jobs. Following this time, the phenomena experiences a significant escalation in its growth. Possible reasons include the willingness be employed for organizations with improved employment regulations, the need for a more fulfilling position and prospects for career progress, and safety worries associated with said COVID -19 epidemic.

Dataset:

To achieve the above mentioned objectives we are using a HR Employee Attrition Dataset from Kaggle. This dataset consists of the information of various attributes such as Age, Attrition, Business Travel, Education, Department, Daily Rate and Distance from home. The dataset has different kinds of data types such as int, char, float and boolean. Few columns in the dataset are further divided into different types. The column Age provides information about age of the employee, and it is int data type.

Attrition column is a Boolean data type which flags True if the attrition exists or else it flags False. The column Business Travel is String Data type which provides information about an employee and travel to the company. Department column provides information of the employee and the department he/she belongs to. It is a String datatype. There are few other columns which data types are int such as Distance from home, Employee Number and Employee Count. The Dataset is very huge and it might contain Nan/Null values. These unwanted values can be removed using Data Cleaning methods.

Detail design of Features:

The detailed design describes complete information of features used to build this model. In the feature, Environment Satisfaction we have categorized data into 4 different metrics such as low, medium, high and very high. It provides information on how good an employee is satisfied with his work environment. In the other feature, Job Involvement is an integer data type which again measures how good an employee has his/her involvement in the job. Similarly, Job satisfaction measures how satisfied an employee is with his/her job. Education feature provides what level of educational background an employee has. Work Life balance measures how an employee is maintaining his/her work life balance. With all the above features, the model is trained along with the trained dataset and used to predict the employee attrition rate.

Analysis:

Now let's see the type of observations which was obtained based on the feature selection and the data which is more required for the analysis which helps in future building the classification model.

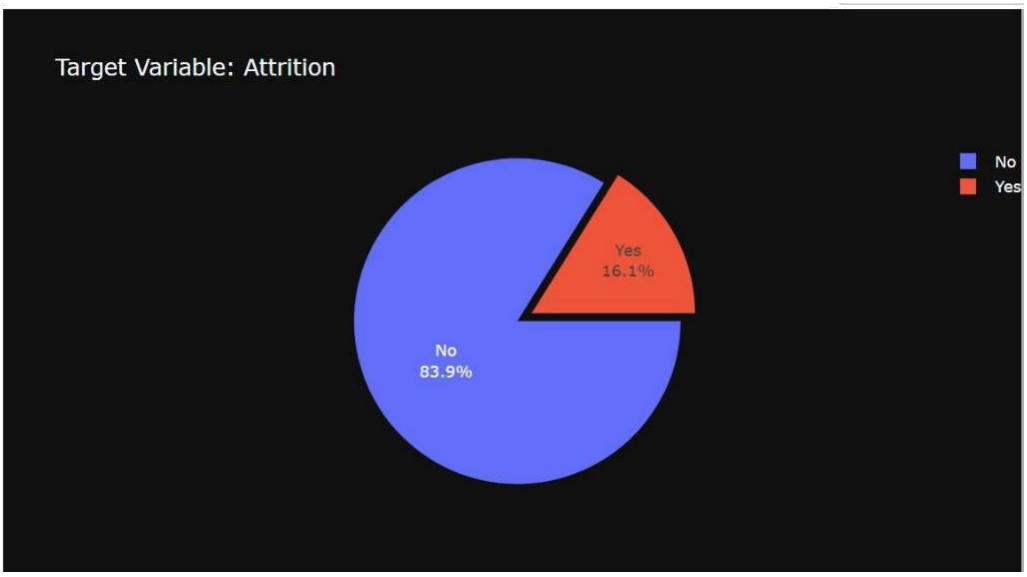

Fig No.: 1 - Pie Chart Analysis - Employee Attrition

We could observe the Attrition rate affecting the total employees, There are 16.1% of people who are accepting that the attrition is helping there growth where as the rest 83.9 % of people who declines that and say that the growth is not affecting.

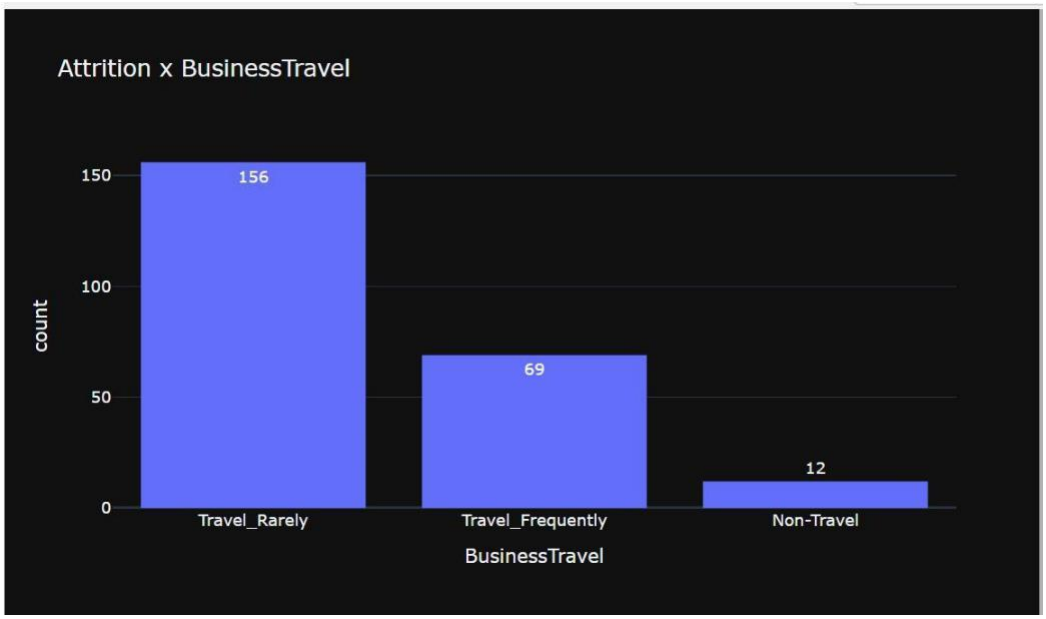

Fig No.: 2 - Bar Plot - Employment based on Business Travel

The Employment, which is affecting Travel, The people who travel more are the medium affected and people who migrate or move to new locations are the maximum affected users. That’s because they are not strong enough to make the decision which tends to increase the attrition rate. The people who are not even planning to travel are the least affected by attrition.

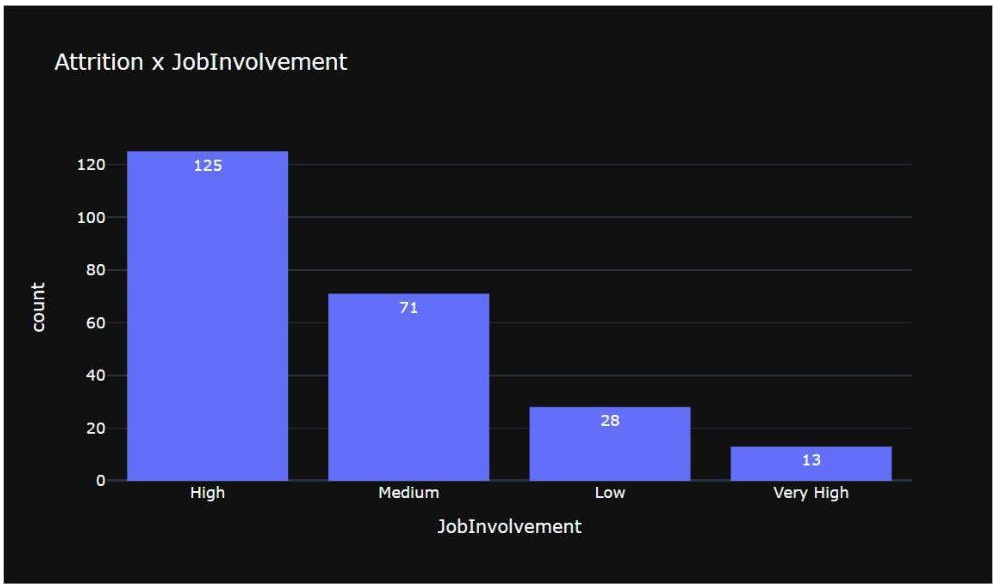

Fig No.: 3 - Bar Plot - Type of Job Involvement

The Attrition rate is based on the People who are interested in working on a particular type of job. Some might be more interested, some might be less interested, which can be found by observing the graph. The Highest Interest are of about 13 people followed by the second highest set of about 125 people. Which shows the job involvement also affects the attrition rate per person based on Involvement.

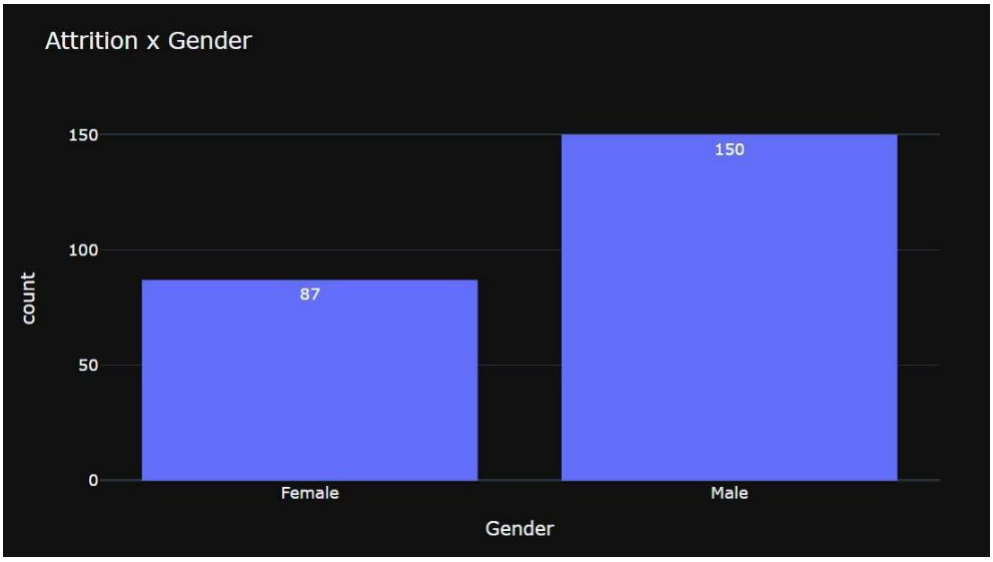

Fig No.: 4 - Bar Plot - Attrition based on Gender

The Comparision of data information based on the Gender based people is classsified, We could observe that male employees are more attracted towards the attrition, There might be several factors which affect this scenario. Let’s try to find more by doing the analysis on the specific set of factor in the later part. The Female is of about 87 out of 240 people. 36 % of people doesn’t accept it and the balance 64% Accept it.

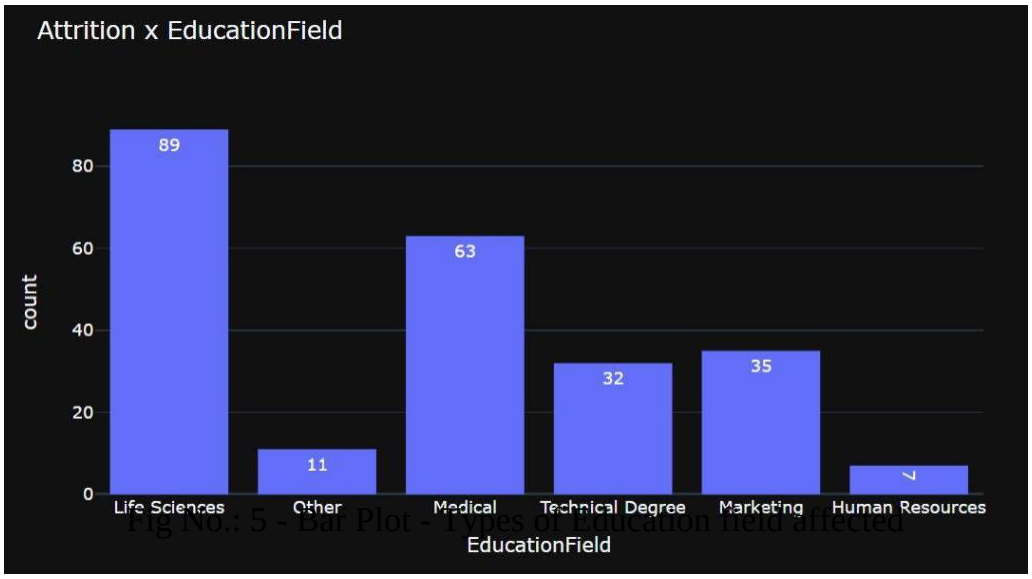

Fig No.: 5 - Bar Plot - Types of Education field affected

The Employment, which is affecting Travel, The people who travel more are the medium affected and people who migrate or move to new locations are the maximum affected users. That's because they are not strong enough to make the decision which tends to increase the attrition rate. The people who are not even planning to travel are the least affected by attrition.

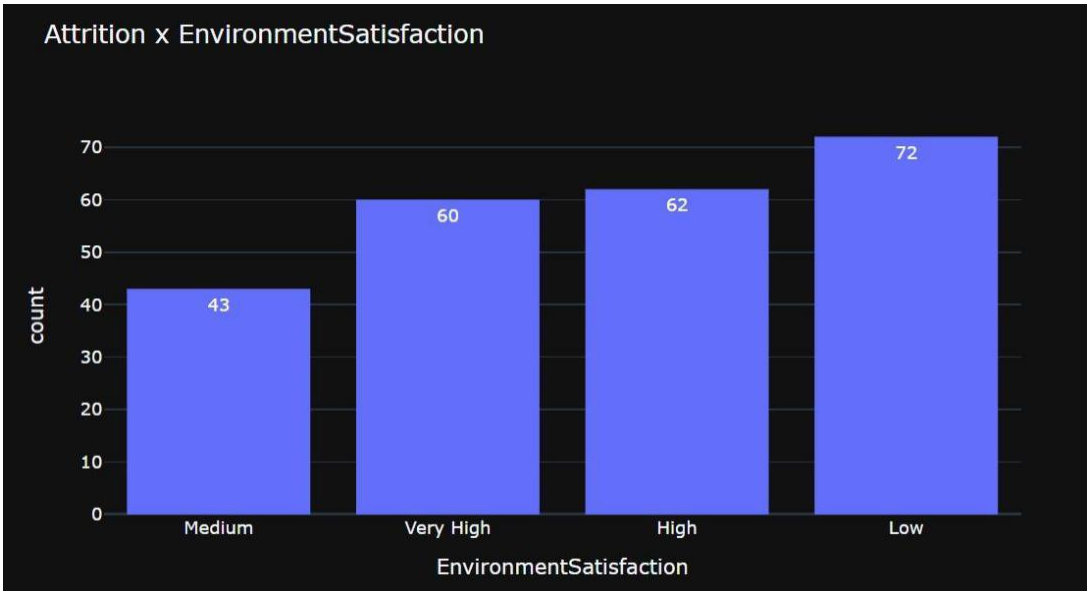

Fig No.: 6 - Bar Plot - Attrition rate based on Environment Satisfaction

The Environment where the people work also changes the thinking of each person based on the type he or she wanted to leave the company based on various other factors affecting them. The Lowest environment satisfaction people are the maximum no of employees

who tries to change the company and start to follow there interest. 72% accept that low satisfaction on the Environment makes them to leave the company .

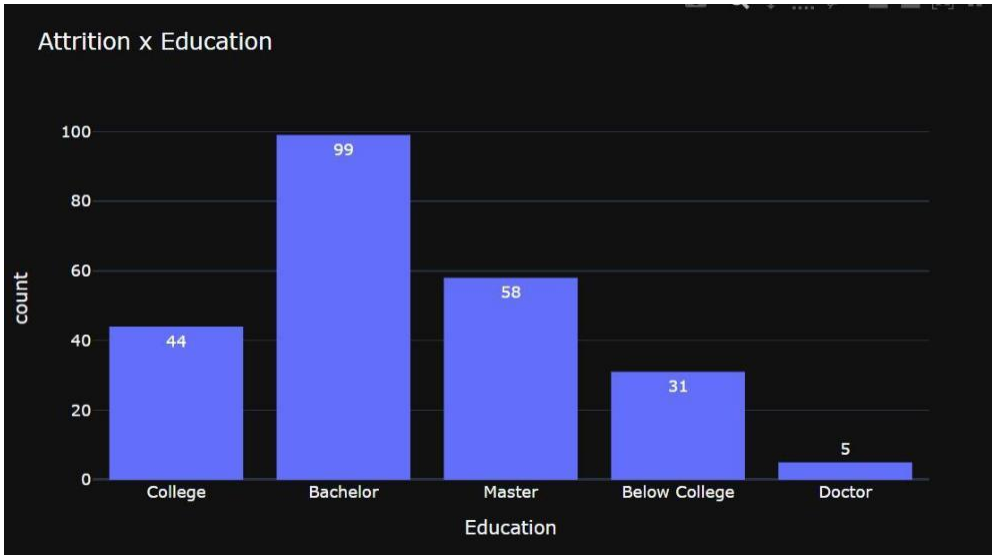

Fig No.: 7 - Bar Plot - Types of Education Degree

The Maximum affected degree completed employees who accept the attrition is right as they change the company after the first few years of experience and plan to pursue there higher degree or plan to change the company so that there salary hike is increased, on the other hand we could observe the highest degree obtained employees are unable to move as they might be overqualified or factors tending towards them would make them not to move out of the company. As they don't have many factors which tends them to move.

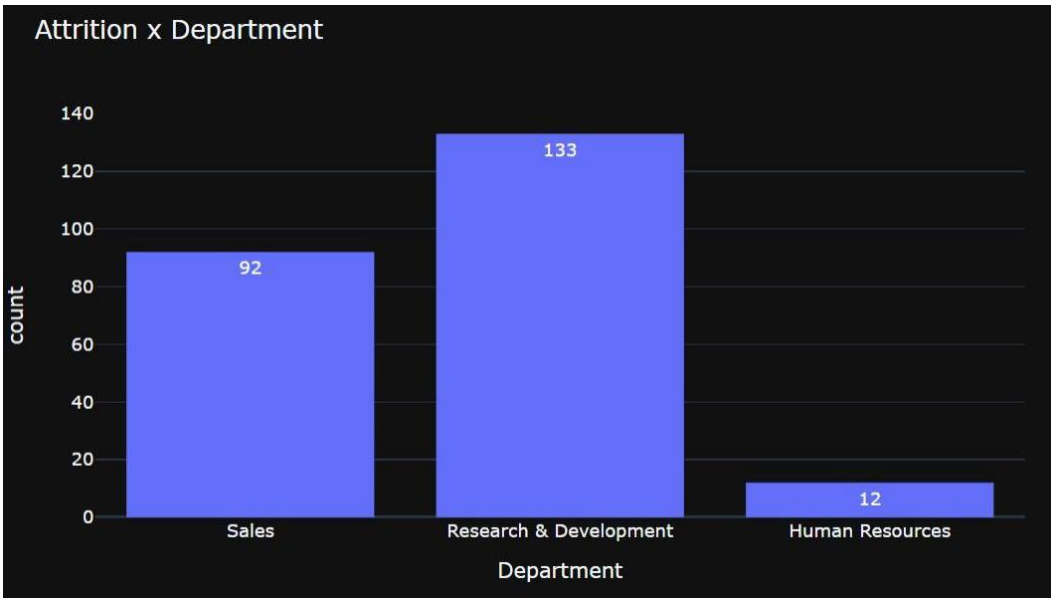

Fig No.: 8 - Bar Plot - Attrition based on Department

Based on the Fig. No.:8 We could observe that the R&D is the department with maximum attrition rate when compared with the others. This could because R&D , needs to research a lot of information with respective to the scenario which makes people who are interested to work in there field. When compared with the HR department the attrition rate is low 12 .This shows the employees in the HR department don't like to change as their work is monotonous and fixed, which also makes some happy with their current job.

Implementation:

The Term Implementation which shows what we had tried to execute and the results of those can be shared and presented here. The data which is based on the attrition of employees, we have used python and data visualization tools to predict and classify the model , how many people are towards the attrition.

```
# Creating a function to plot histograms
def barplot(i):
    fig = px.histogram(Attrition, x = Attrition[i], template = 'plotly_dark',
                       title = f'Attrition x {i}', text_auto = 'd3-format')
    fig.show()
```

```
# Creating visualizations for categorical values
barplot('BusinessTravel')
barplot('Department')
barplot('Education')
barplot('EnvironmentSatisfaction')
barplot('EducationField')
barplot('Gender')
barplot('JobInvolvement')
barplot('JobSatisfaction')
barplot('WorkLifeBalance')
barplot('PerformanceRating')
barplot('JobRole')
barplot('MaritalStatus')
barplot('RelationshipSatisfaction')
barplot('OverTime')
```

Fig No.: 9 - Python code to visualize the bar plots based on the categorical values

```
# Transforming bicategoric variables into binary values
X_train['OverTime'].replace({'Yes': 1,
                                        'No':0}, inplace=True)
X_test['OverTime'].replace({'Yes': 1,
                                        'No':0}, inplace=True)
X_train['Gender'].replace({'Male': 1,
                                        'Female':0}, inplace=True)
X_test['Gender'].replace({'Male': 1,
                                        'Female':0}, inplace=True)
```

Fig No.: 10 - Python code to convert the categorical set of information as binary values

```
print(f"{Attrition.shape[0]} emplyees in the dataset left the company.\n")
print("Let's try to find some more information about them!")
```

```
237 emplyees in the dataset left the company.
```

Fig No.: 11 - Python code - Employees Left the company

```
# Unencoding Categorical Features
col = ['EnvironmentSatisfaction','JobInvolvement','JobSatisfaction','RelationshipSatisfaction']

for i in df['Education']:
    df['Education'].replace({1:'Below College',2:'College',3:'Bachelor',4:'Master', 5:'Doctor'},
                        inplace = True)

for i in df['PerformanceRating']:
    df['PerformanceRating'].replace({1:'Low', 2:'Good',3:'Excellent',4:'Outstanding'},
                                    inplace = True)

for i in df['WorkLifeBalance']:
    df['WorkLifeBalance'].replace({1: 'Bad', 2:'Good', 3:'Better', 4:'Best'},
                                    inplace = True)

for i in df[col]:
    df[i].replace({1:'Low', 2:'Medium',3:'High', 4:'Very High'},
                    inplace = True)
```

Fig No.: 8 - Python code - classifying the set of categories into sub

Preliminary Results:

Obtained Results which show majority of people had left the company based on the factors affecting them. 30 % of Employees are leaving as the working environment is not that satisfactory.

Majority of Men Employees which are about 64% tend towards leaving the company as they grow in the career and the higher degree pursing interest. The Employee experience also makes them to move out of the companies because lack of salary for that particular job role.

Life Science based Sector has the majority of people leaving as it is similar to the technological stack where the talent of the particular resource is most required than compare to the level of handling in the type of role job. Almost 80 % of people who had polled shows the attrition rate which affects is based on the type of field each employee is working with.

The Overall results which show not much are interested to move out but there are forced to move based on the environment, situations and other factors. This is based on the given data we were able to analyze the information . In future we are planning to build an machine learning classifier model to predict the attrition rate.

Project Management:

The Project Management is a function which is like an integrated set of inputs from each team member. The overall working of the project and the tasks can be found and placed as a placeholder so that the entire project flow is recorded and helps in finding out the issues. The Stakeholders can just go through the overview of the project and check how the team works towards the project.

Implementation status report

Work completed

Description: We are going to implement employee attrition prediction using machine Learning algorithms starting by cleaning or preprocessing the data followed by exploratory data analysis, training and testing the model.

Responsibility (Task, Person):

Aditya Pujari:

- Searching for an appropriate dataset that most suits our purpose.
- Preprocessing the dataset for missing values, outliers and normalizing.
- Writing documentation or report.

Brinda Potluri:

- Implemented exploratory data analysis for data visualization so that it would be easy to understand the data by plots.
- Writing documentation or report.

Nitin Dunday Mohan:

- Model Visualizations and prediction on the Attrition rate of the dataset Behavior. The EDA Analysis is based on how much the employees are leaving the company
- Writing documentation or report.

Sai Tarun Gunda:

- Implemented hyper parameter tuning for accurate predictions.
- Writing documentation or reports.

Contributions (members/percentage)

Aditya Pujari	25%
Brinda Potluri	25%
Nitin Dunday Mohan	25%
Sai Tarun Gunda	25%

Work to be completed

Description: Building Model to find the classification model and make the efficiency of the model by tuning the hyper parameters and deploying it on Heroku instance.

Responsibility (Task, Person):

Aditya Pujari:

- Trying to implement few more visualizations

Brinda Potluri:

- Try to build and work on the performance of the model

Nitin DM:

- The Model Prediction is checked and evaluation metrics and confusion matrix is obtained. Training and testing different machine learning models for building proper predictive model for employee attrition prediction.

Sai Tarun Gunda:

- The Data can be projected in Heroku and displayed as an application.

Issues/Concerns :

The dataset had unnecessary features so choosing between factors that impact employee attrition and doesn't impact was an issue.

Dataset was I cleaned and imbalanced. Cleaning including replacing the bill values and balancing the dataset was a concern. As if this is not done properly then it would effect the final model predictions.

Project Increment – 2

Video Link: https://drive.google.com/drive/folders/1w_XbJLwWFbe0NEgcG-hJpUHvi33hNH3L

GitHub Link : https://github.com/adityapujari98/Employee_Attrition_Analysis

Dataset Link : https://www.kaggle.com/datasets/whenamancodes/hr-employee-attrition

Introduction

Motivation:

Due to the current employment market situation and how bad it tells a lot about how an employee would be treated by his/ her company; our project is just an initiative. We are well aware of what impact or effect would be on the economy if unemployment keeps increasing. We can research about the employee's path leading to one company from another. Examining an employee's shift to a different or existing role is another option.

Significance:

It has been observed that companies invest a lot of capital to hire employees, and reducing employee attrition will directly help in saving a lot of money for the company. The organization is constrained by having a staff that is continually comprised of employees with less experience due to a relatively high rate of employees leaving the company.

It's not just about recruiting new people but the expenses that follow it, for instance after hiring an employee the company needs to put in a lot of effort and time in training or coaching them. There are several other issues that come along with employee attrition like a decrease in performance, and investment, along with losing expertise. After proper research, the range of expenses of replacing one employee with another (new hire) will be between 16 percent to 213 percent of their total yearly compensation. In particular, the US spends nearly one trillion dollars in one year on overall turnover expenses.

Objectives:

We have divided and explained our main objectives below: Firstly, we would understand our data thoroughly by doing Exploratory Data Analysis. This will be very helpful in understanding the underlying patterns like what factors are affecting employee churn, to what extent, and how many such factors are there. So

that proper actions can be taken to reduce their employee attrition. To achieve this, we have performed Correlation analysis followed by Univariate and Bivariate analysis.

Univariate, Bivariate, and Correlation Analysis:
When we are analyzing just one variable with the target then it is known as univariate analysis. Similarly, when we are analyzing 2 variables it is known as bivariate analysis. In case we are considering more than two then it is called multivariate analysis. For understanding the relationships between our categorical or numerical variables data we use correlation analysis. The outcome of this analysis would be either positive, negative, or no correlation. Where positive means the value of one variable increase or decreases with the other. A negative correlation means the value of one variable decreases with the increase in another variable's value. No correlation means that there is no such relationship between the two variables. We will discuss our analysis in detail in our plots.

Model Building:
After performing Exploratory Data Analysis, we will then focus on model building. We chose the Ada-Boost classifier, Gradient Boosting, and Random classifier for our project.

Ada-Boost Classifier:
An Ada-Boost classifier also called a meta-estimator starts by trying to fit the classifier towards our initial dataset. It then after that will fit multiple copies of our classifier onto that very dataset, but with the weights of our instances that were incorrectly classified being changed so that later classifiers would concentrate more on challenging cases.

Gradient boosting:
A machine learning method called gradient boosting is often used, among other things, for classification and also the regression tasks. It provides a forecasting model that is made up of the ensemble of decision trees-like weak prediction methods. It is among the most effective methods for creating predictive models.

Random Classifier:
A random forest also called a 'meta-estimator' that employs averaging to increase predicted accuracy and reduce overfitting after fitting numerous decision tree classifiers to distinct database subsamples.

Model Evaluation:
For evaluating our model for its performance or accuracy, we will employ different methods. For instance, F-1 Score Precision, and Recall. Based on the results, we will then see if any hyperparameter tuning or any changes to the model is required.

Features:

- Age: This is to indicate the employee's age
- BussinessTravel: The overall travel requirement of the employee
- Department: The particular department that the employee is working for or under
- EducationField: The educational background of a particular employee
- Gender: To indicate if it is a male employee or female.
- HourlyRate: This is the pay or salary of an employee on the basis of hours
- JobRole: It is to indicate the role (job) of a particular employee.
- MonthlyIncome: It will indicate the monthly earnings or salary of an employee
- OverTime: To indicate the extra work or overtime an employee (beyond office hours) has worked
- PercentSalaryHike: The hike that the employee received along with his/ her salary
- TotalWorkingYears: It indicates the entire working years of a particular employee in his lifetime
- YearsAtCompany: It indicates the entire working years of a particular employee in this company
- YearsInCurrentRole: The number of years that an employee has worked in the present role.

Background

One of the significant issues faced by recruiting teams or departments is employee retention. It has been observed that the employees tend to expect or even demand certain perks from their company like growth aspects, work-life balance, their reputation, role (if it has demand in the market or not)[5], and job security being one of the important things among others.

In order to tackle this issue, since the past a lot of work is being done either individually or in different teams to come up with various solutions. Employee attrition or employees leaving the company is due to various reasons, for instance,

long-term career growth, pay scale, work culture, work-life balance basically work environment.

After a lot of research, researchers were able to understand the underlying relationship between employees leaving the company with their company's values or commitment including work environment and their job satisfaction was playing a key role in their attrition.

Model Architecture

In this experiment we have chosen a Random Forest model in order to predict if any given employee quits the organization or not. The Random Forest Model is considered a Supervised Machine Learning Algorithm[3] which is built from decision trees. It is trained by a bagging method which intends a combination of learning models to maximize the accuracy. This Supervised Machine Learning Algorithm has a special quality, it handles both regression and classification types of datasets continuous and categorical variables. In case of classification, it uses an ensemble method which combines multiple learning algorithms and provides solutions to complex scenarios. In case of regression, it takes the average of mean of trees and predicts the solution. The below diagram explains the working of a Random Forest Model.

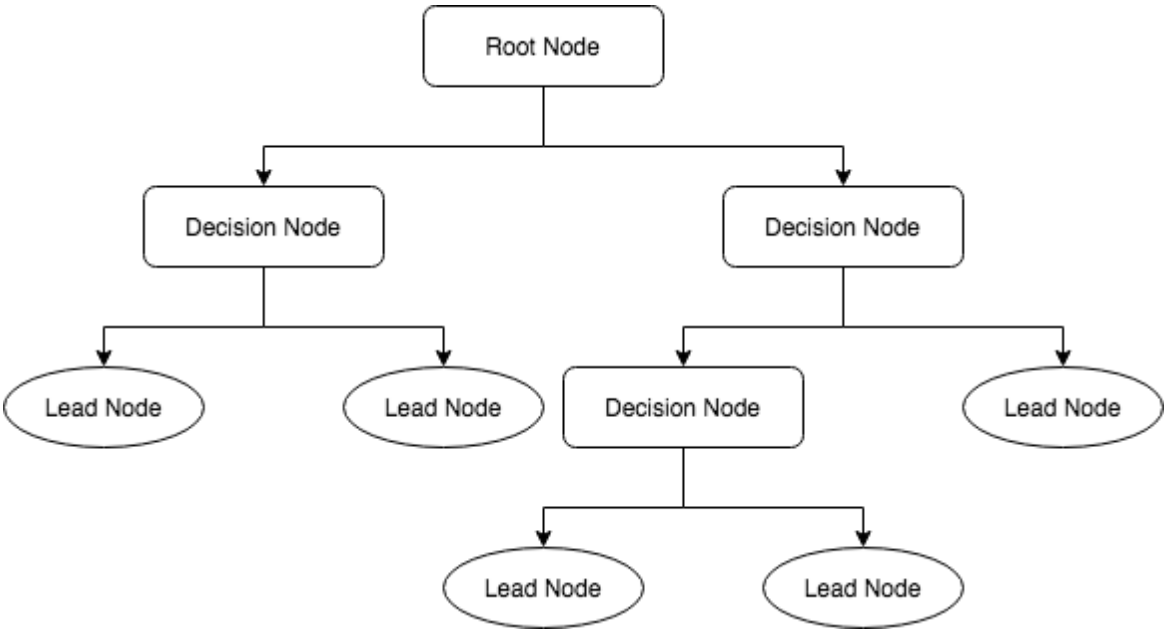

Fig 1: Random Forest decision tree architecture

For any Random Forest Algorithm, Decision[2] Trees are the building blocks. The Decision Tree is a tree like structure and a decision support technique. As shown in the above figure, it consists of three major components. They are decision nodes, leaf nodes and a root node. These nodes represent the attributes used in the model to predict the output. Root node and decision nodes represent features of the tree. The tree ends at leaf node, any decision tree wouldn't continue after leaf node. The other two important factors of a decision tree are entropy and information gain. Entropy is used to calculate uncertainty. Information gain is the measure. It checks how uncertainty is reduced in the target variable. It uses independent variables which are features to gain information on the target variable. It is used to train the decision tree. These two measures are very much important in splitting the trees.

Dataset

In this project we have used a Kaggle Dataset which consists of data related to Employee attrition rate. Multiple factors of an employee have been considered in this dataset. The first such factor of an employee has been considered is Age. There are various age groups in this dataset. The minimum age is 18 and the maximum is 60. Out of all age groups, there are more employees around the age 30. This age group is an integer data type. The second most important factor which has been considered is Attrition. It is a Boolean datatype. It says either yes or no based on the employee attrition status. To find the Attrition rate, the Business Travel of an employee has also been considered. It has been divided into two different classes which are employees who travel Rarely and employees who travel frequently. It is a string datatype. In any organization there are two different departments. They are Sales and Research & Development. These departments are also considered in the dataset. It is also a String Datatype.

Distance from is another important factor which employee attrition rate is heavily dependent on. It is an integer data type which indicates the number of miles the organization/company is situated from an employee's home. Education field and Job Role are other two string datatype factors on which employee attrition rate is dependent on. Education field describes the field of education of an employee pursued and job role describes what type of the employee is working on. Hourly rate, Job satisfaction and Job involvement are the three job related factors which were included in this data set. All three factors are integer data types and indicate an employee's salary rate, level of his/her satisfaction with the job and their involvement in the job.

It is believed that marital status plays an important role in an employee's work life. Hence it is also included in the dataset in order to predict the attrition rate. After marital status it is really important to understand an employee's work life balance. Hence it is also included in the dataset as an integer data type. Years at the company and Years in the same role are other factors which are related inside of an organization. Both these factors are considered as integer data types.

Design of Features

In any machine learning project, selecting features would be the key aspect in building the model. Similarly we have chosen a few specified features in order to predict employee attrition rate[1]. The first such feature we have selected is Attrition. It describes if the employee has left the organization or not. It is an integer data type which says either Yes or No.

The second feature we have selected is Business Travel. It is divided into three different types. They are Rarely Traveled, Frequently Traveled and Not Traveled. It is an integer datatype. One other feature we have selected is the Department. It describes the department the employee belongs to. It is divided into three different sub categories. They are Sales, Human Resources and Research & Development. Education field, Gender and Job role are few other features we have used in this project to determine the employee attrition rate.

As the mental health of an employee makes a difference on his ability to work we have considered one of the mental emotional factors Marital Status as a feature. It is an integer datatype. Based on these features we would build the model in order to determine employee attrition rate.

Analysis of data

Data Pre-processing

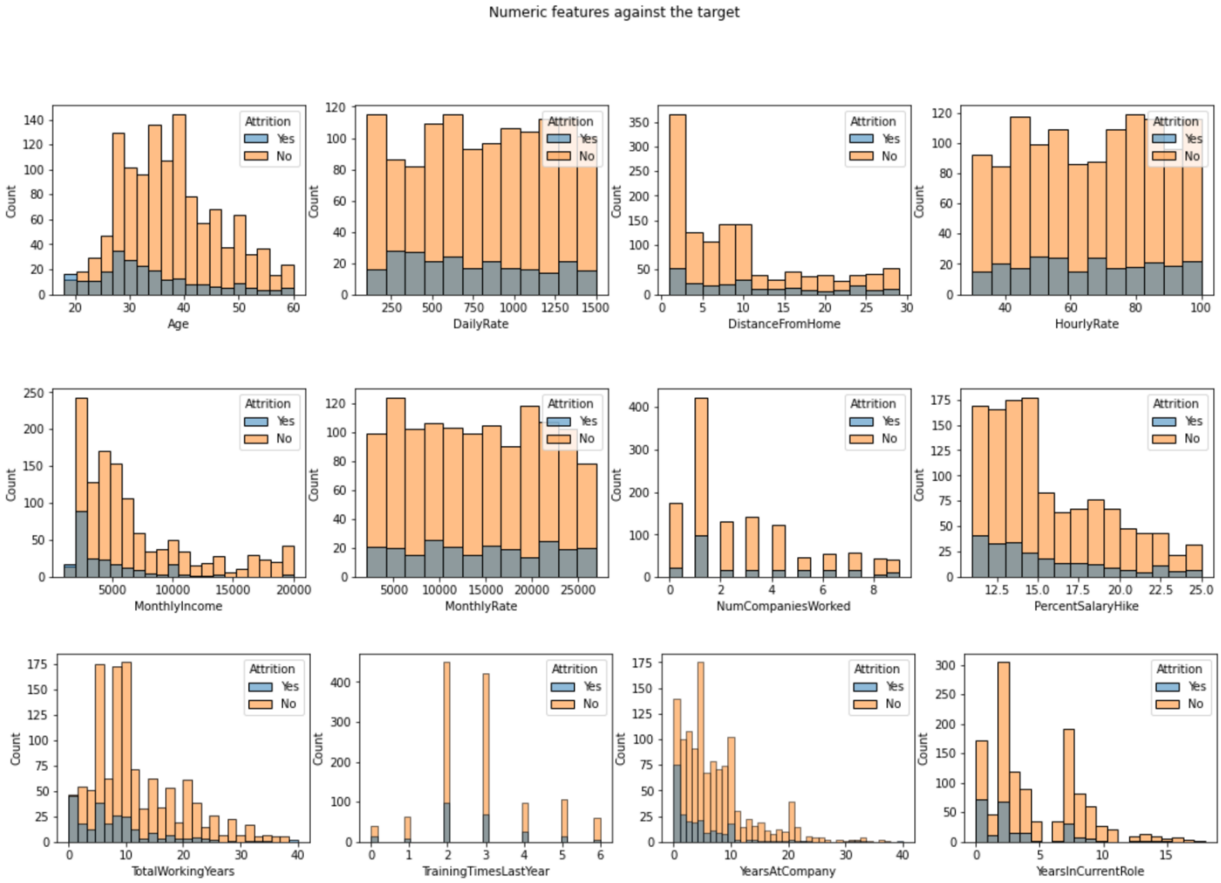

Fig 2: Correlation between different features

The Fig.: Which shows various features variations based on each employee. We could now understand the data based on the visualizations, based on the above data we could say the age group of people participated in the survey is of about 20 to 60 years of range, most employees are of about age 30 to 40. The graph which shows the data is in a form of gaussian curve.

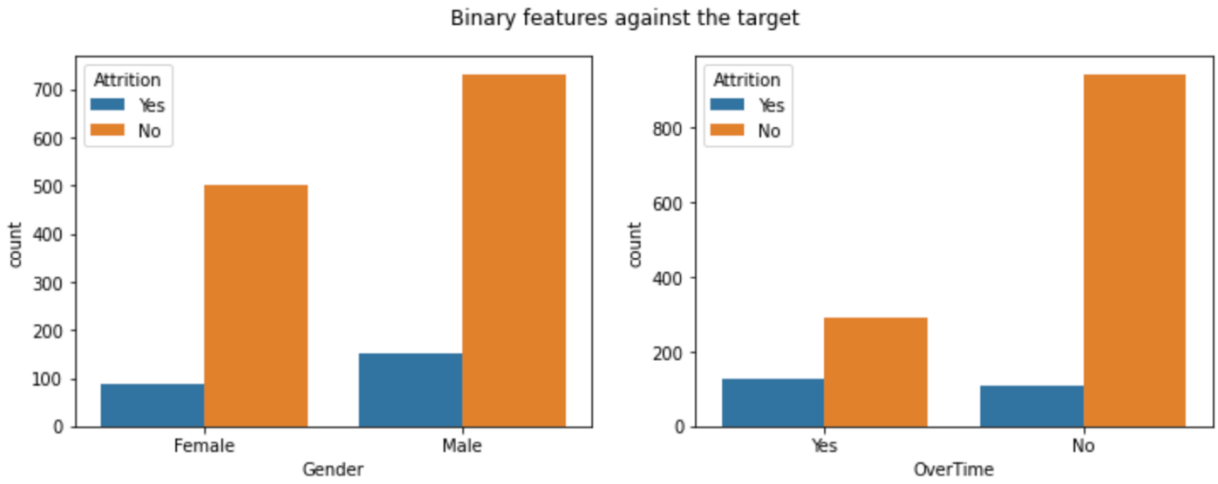

Fig 3 - Features based on Binary Classification

The binary classification is based on 2 types such as gender and overtime basis. The gender shows much clearer separation of data with male and female, we could observe the more employees towards the attrition is for male employees. The overtime worked employees which states the more employees who had worked more had less attrition.

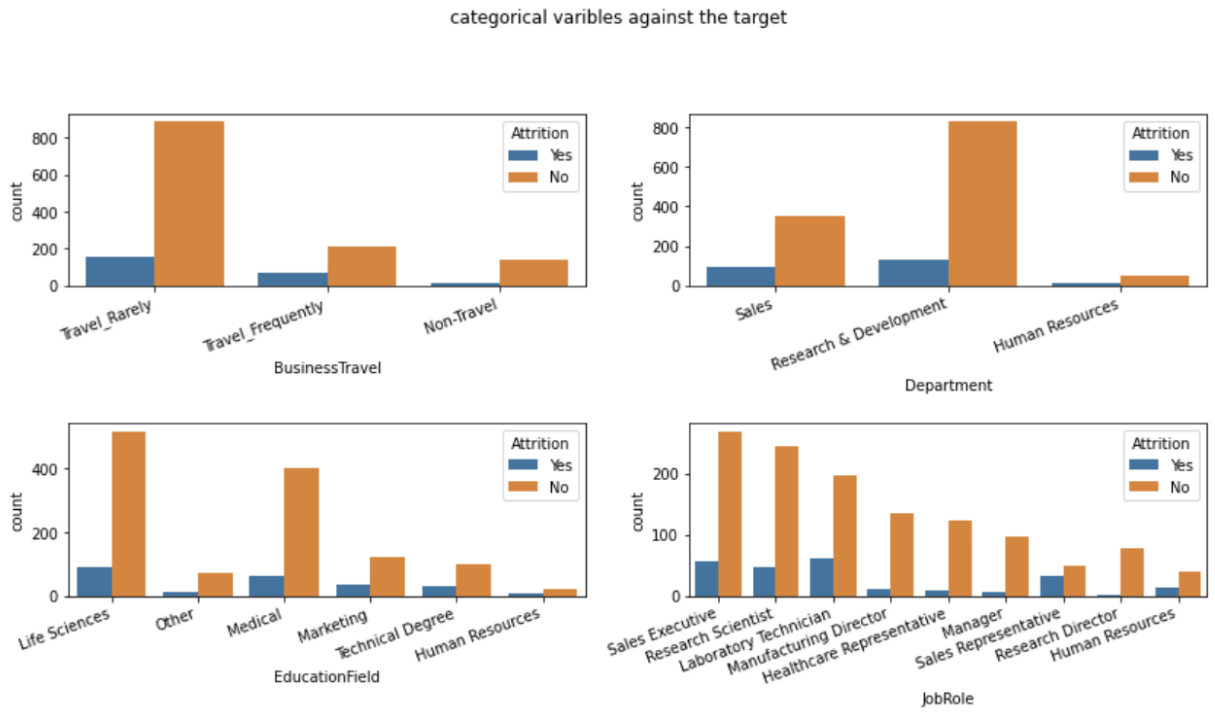

Fig 4 - Features based on Categorical data

The above Fig.: Which explains the data based on the categorical information of data. This shows the attrition rate based on the overall picture, then it further to department wise and then its further to education field and then the job roles. Based on the observations we could see the most affected job role is Lab technician, the min is done in manufacture director role. The lowest affected field in Human resources, as the no. of employees are required to handle more information about the employees , hence the attrition rate is less when compared to the other information.

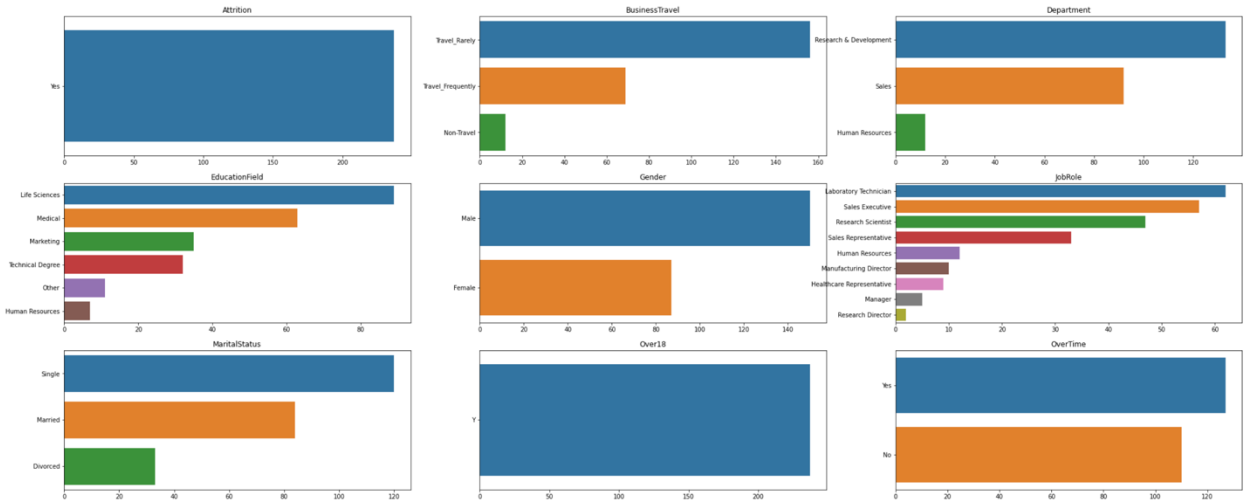

Fig 5 - Features based on Categorical data

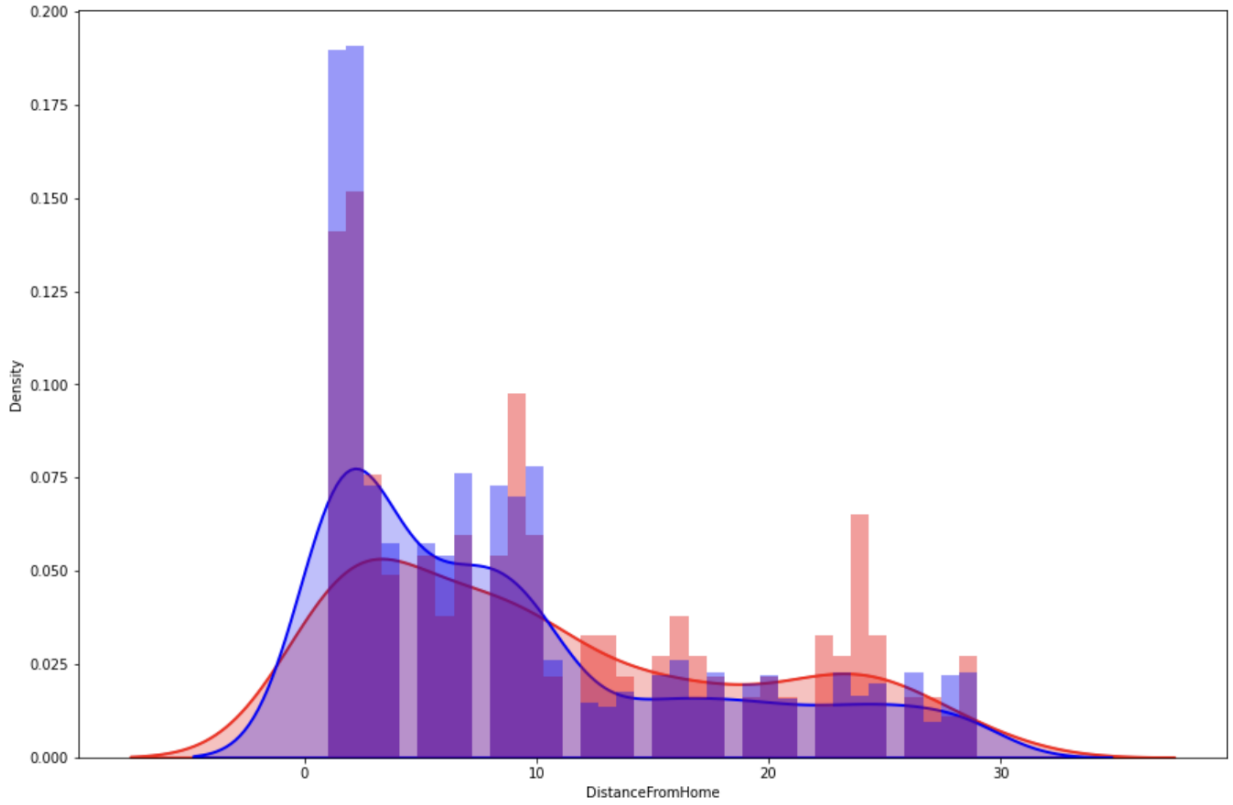

Fig 6 - CDF, PDF for a Feature (Distance from home, density)

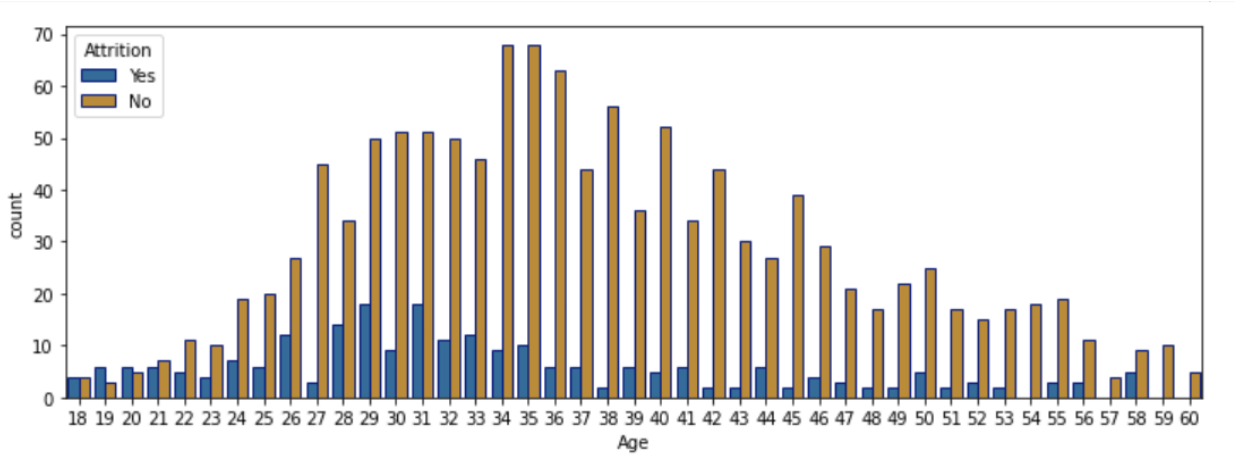

Fig 7 - Age vs count Features

The data which displays the information in which maximum attrition rate affected in based on age 31. The attrition rate based on the age is based on the other factors of employees. The no. of employees of age 35 shows maximum not of people are found, the data shows the employees count with less of age 25 and age greater than 40 the count of employees are less.

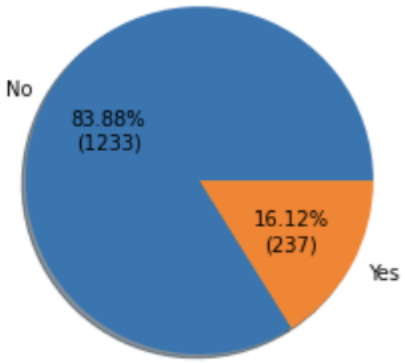

Fig 8 - Data Balanced information

The data which has information which is biased of saying the attrition rate is accepted of about 16.12 % and the no attrition of about 83.88%. We need to normalize the data before we split the data for test and train information.

Implementation

GitHub Link: https://github.com/adityapujari98/Employee_Attrition_Analysis

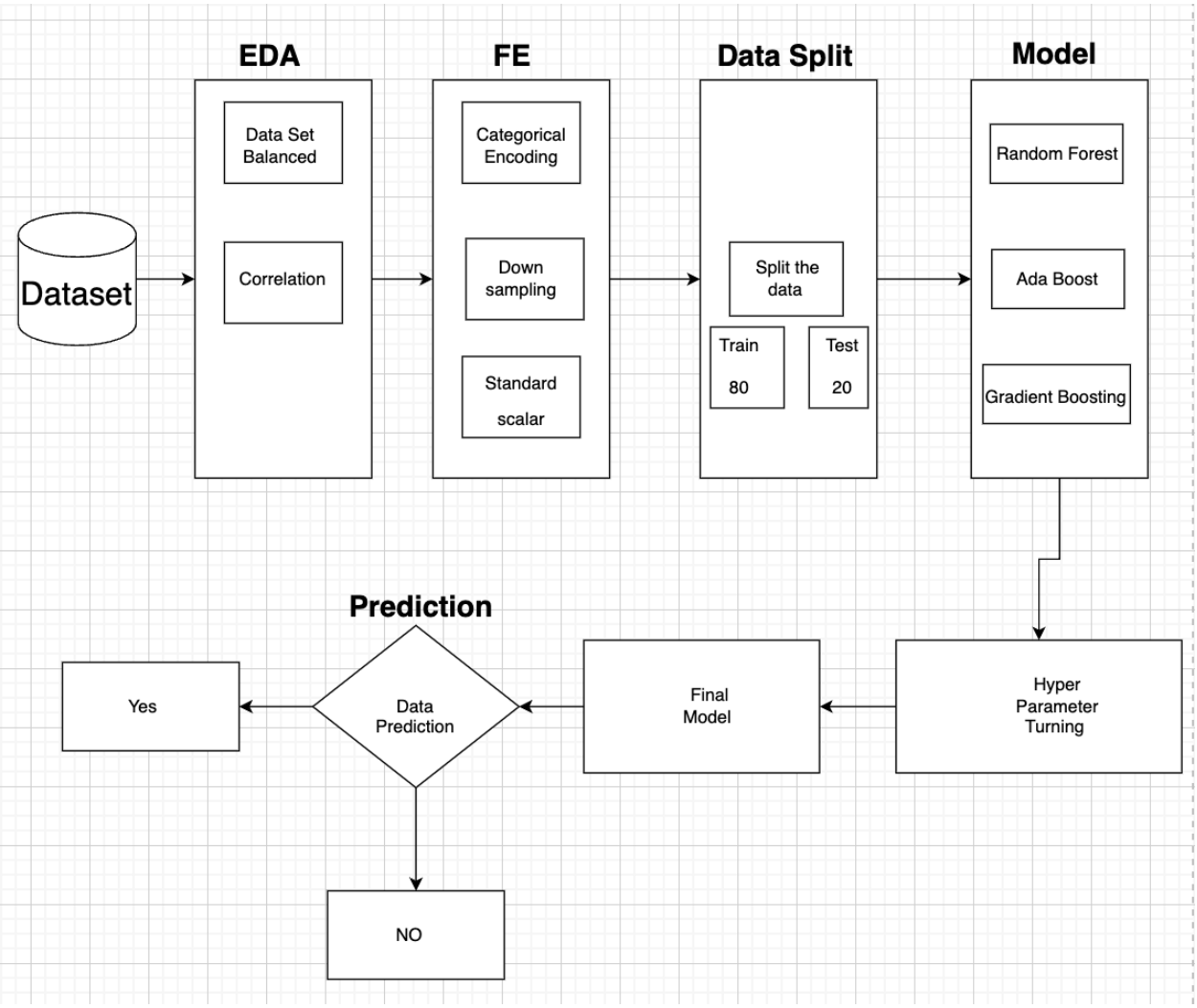

Fig 9 - Architecture Flow

Dataset:

The data set which contains information about 1470 set of employees based on 35 types of features for a particular employee.

Exploratory Data Analysis:

Dataset balance:

We are checking the data set is balanced or not, we could observe that No set of value is of about 83.33% and the rest is 16.66% are of about yes value, which shows that we have a biased dataset, and we need to normalize it before we use or else, we would get the result in a biased format.

Correlation:

The Feature importance can be calculated in the section, this can be further classified as much variant features such as “Age, Daily Rate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager”. The 14 features are compared based on the relationship with each other the dependency of the data based on each feature is calculated and graph is displayed to identify the relationship. Based on the feature importance we are taking it to classify as the important value and prior data is selected as an important feature to classify the data.

Feature Engineering:

Categorical Encoding:

The data is split with the categorical information such as ‘Business Travel', 'Department', 'Education Field', 'Job Role', 'Marital Status' based on the categorical the data is processed and visualized to know about the categorical information, this is future done to get more information based on gender and split the information in much useful manner.

Down Sampling:

The data is then down sampled based on the majority and minority of data in the ratio format and then the samples data is then further sent to the next process to identify the features.

Standard Scalar:

The data is future distributed based on the classification as 0 and 1. The resize of values are done, we are using Standard normal distribution to find the mean and SD based on the features selected

Data Split:

The data is split with 80:20 for the train and test respectively to train and test the model based on multiple models, The data is split to make sure the model which predicts the data is not biased and make sures the information passed has a corrected data.

Results

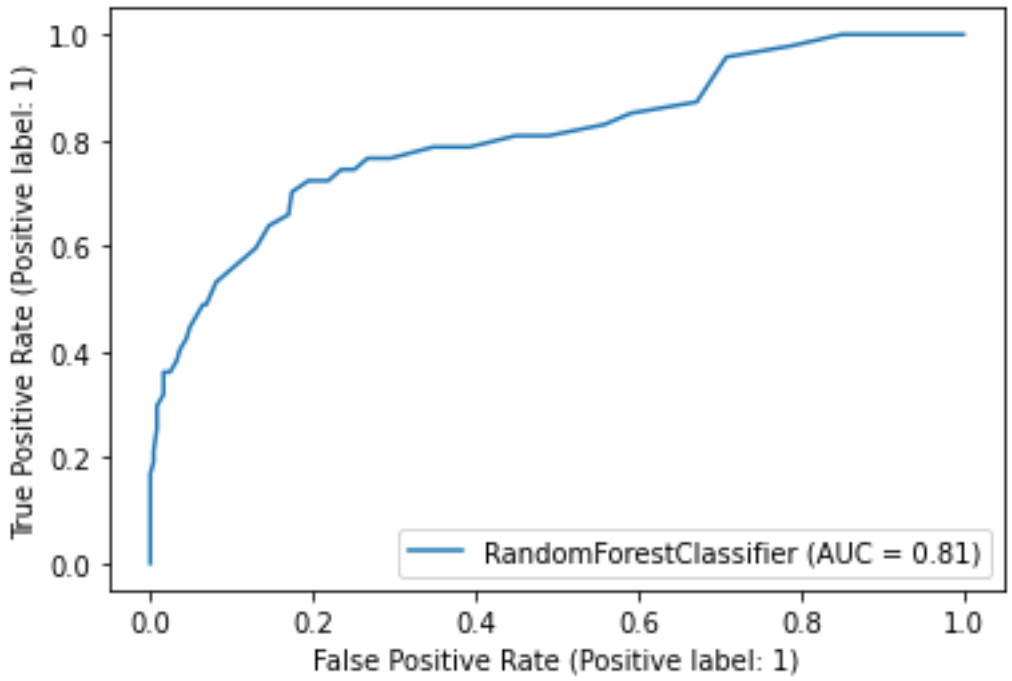

Fig 10 – AUC Curve for Random Forest Classifier

The above graph represents relationship between true positive and false positive. This is the result of the Random Forest Classifier's output. From the AUC (Area Under the Curve) plot, we can see that Class 0 is predicted by the model to be 0 and also the class 1 to be 1. The model performs efficiently or betters the higher the AUC.

```
_____Model Performance_____

              precision    recall  f1-score   support

           0       0.87      1.00      0.93       247
           1       0.91      0.21      0.34        47

    accuracy                           0.87       294
   macro avg       0.89      0.60      0.64       294
weighted avg       0.88      0.87      0.84       294


Base Model Accuracy: 87.07%.
```

Fig 11 – Evaluation metric for Random Forest Classifier

The above picture represents the Evaluation Metrics. As we can see we got overall precision of 0.87 for class 0 and 0.91 for class, correspondingly an f1-score of 0.93 for class 0 and 0.34 for class 1.

```
**------------------Top 5 Important Features-----------------**
               Varname       Imp
49        OverTime_No  0.150406
0                 Age  0.078165
9       MonthlyIncome  0.071732
15   StockOptionLevel  0.066066
16  TotalWorkingYears  0.061439
1           DailyRate  0.061074
```

Fig 12 – Top 5 Features for final model building

After training the baseline model, we got the above-displayed top five features.

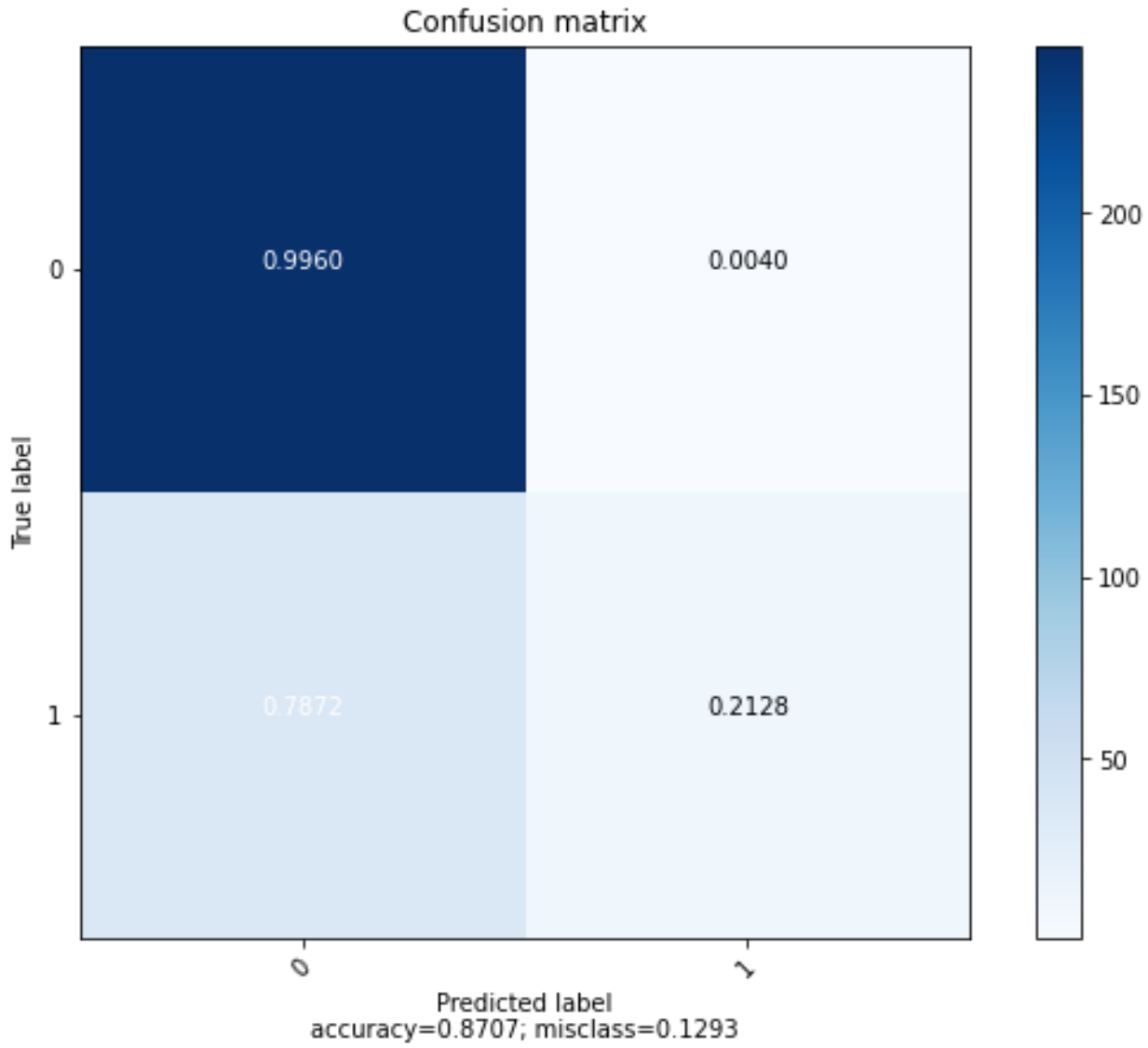

Fig 13 – Confusion Matrix with 87% Accuracy

A performance indicator for our machine learning categorization is the confusion matrix. The obtained accuracy is 87.07% and the misclassified percentage is 12.93%.

Project Management

Implementation status report
Work completed

Description:
Built model to find the classification of employee attrition, predicting whether the employee may leave the job or not. Performed hyperparameter for getting good accuracy of the model.

Responsibility (Task, Person):

Aditya Pujari:

- Implemented some visualizing for understanding the correlation with the high needed features.
- Did feature engineering to find new features and find patterns among them

Brinda Potluri:

- Found different ways to balance the unbalanced data
- Performed data scaling and down sampling of data and proper validation of data to check loss of data
- Validated the data input parameters before splitting the dataset for checking data leakage

Nitin DM:

- Performed hyperparameter tuning to get the best model among the three models
- Trained model with different samples of data as part of cross validation to avoid overfitting

Sai Tarun Gunda:

- Performed statistical analysis of data for understanding the different aspects of data
- Using baseline model figured out the important features for further training the final model

Contributions (members/percentage)

Aditya Pujari	25%
Brinda Potluri	25%
Nitin DM	25%
Sai Tarun Gunda	25%

Issues/Concerns

- As the data which we got is not balanced the model built is some what biased with the majority class that is NO.
- As the number of samples, we got are less as compared to normal dataset, so we had to go with machine learning model only which was our best option to go

References

1. PRIIT ULMAS1, Tallinn University of Technology, Akadeemia Tee. "Segmentation of Satellite Imagery using U-Net Models for Land Cover Classification"
   - **https://arxiv.org/pdf/2003.02899.pdf**
2. Roger Xu Jiang, "Neural network for satellite image segmentation"
   - **https://towardsdatascience.com/dstl-satellite-imagery-contest-on-kaggle-2f3ef7b8ac40**
3. J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders, University of Trento, Italy, University of Amsterdam, the Netherlands. "Selective Search for Object Recognition"
   - **http://www.huppelen.nl/publications/selectiveSearchDraft.pdf**
4. Voodan, "Satellite Image Segmentation: a Workflow with U-Net"
   - **https://medium.com/vooban-ai/satellite-image-segmentation-a-workflow-with-u-net-7ff992b2a56e**
5. Hannah Peterson, "A Beginner's Guide to Segmentation in Satellite Images"
   - **https://medium.com/gsi-technology/a-beginners-guide-to-segmentation-in-satellite-images-9c00d2028d52**