Name- Sukrut Swapnil Mayekar
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - March 2022
Certificate Code- TCRIB2R28
Date of submission- 04-04-2022

Technical Coding Research Innovation, Navi Mumbai,
Maharashtra, India-410206

# (HR EMPLOYEE ATTRITION DATA ANALYSIS)

A Case-Study Submitted for the requirement of
**Technical Coding Research Innovation**

For the Internship Project work done during
**DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM**

by
Sukrut Swapnil Mayekar (TCRIB2R28)

Rutuja Doiphode
CO-FOUNDER &CEO
TCR innovation.

**Abstract -These instructions give you the basic guidelines for preparing papers for IEEE conference proceedings.**

**Index Terms - List key index terms here. No mare than 5.**

Name- Sukrut Swapnil Mayekar
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - March 2022
Certificate Code- TCRIB2R28
Date of submission- 04-04-2022

# I. Introduction

In today's world we are surrounded by raw, unstructured and structured data which can be extracted from physical world with the help of the sensors and from the virtual world with the help of cloud, server and other storage devices. The process of data science is that we have to first access the data and make insightful predictions from the data using analysis. This can be done by various python libraries like Matplotlib, Seaborn or Bokeh. In our project we used Plotly as the visualizing library. This requires analytical skills as well as business and domain understanding. The next step of the data scientist is to clean the data and preprocess the data. In cleaning the data, we have to check if the dataset contains missing, nan or duplicate values. If yes then we have to fill these values with the mean, median or mode of the concerned features. After this step, we have to check the distributions of data, how much numerical and categorical features are present in the dataset. Do they make a normal distribution or not. Are some values in the dataset are too large and some of them too small. In this situation we have to use min max scaler to take all of the values between the range of zero and one. If some of the features contains discrete values then we have to use one hot encoding. This preprocessing step is very important for applying our machine learning models on the data to get higher accuracy. As we have to predict if the employee leaves the company or not which comes in binary classification, we used Logistic Regression and Random Forests and Decision Tree Classifiers. We got the highest accuracy of 94 % using the random forest classifier.

The projects is created by following this six steps

1. Importing all libraries from packages
2. Fetching datasets from csv file
3. Checking missing data,unique values and also total records.
4. Analysis the datasets through various graph study
5. Data pre-processing and Converting the data to another form
6. Applying the ML algorithm to predict the data
7. Checking the accuracy score.

DATASETS

The data is given to us in the csv file format. In this data there are total 2940 rows and 35 columns, in this there are some numerical data and some categorical data. The types of data are numerical, categorical, date time values etc. First procedure whenever we fetch the data is to identify the data types and to check the shape of the data which tells that how much columns and rows are there in datasets. We are also checking any missing values are there in the datasets and after classify the missing values we have to replace it by simply mean, median or mode of that feature.

Fig 1.1 Datasets

| | EmployeeNumber | Attrition | Age | BusinessTravel | DailyRate | Department |
|---|---|---|---|---|---|---|
| 0 | 1 | Yes | 41 | Travel_Rarely | 1102 | Sales |
| 1 | 2 | No | 49 | Travel_Frequently | 279 | Research & Development |
| 2 | 3 | Yes | 37 | Travel_Rarely | 1373 | Research & Development |
| 3 | 4 | No | 33 | Travel_Frequently | 1392 | Research & Development |
| 4 | 5 | No | 27 | Travel_Rarely | 591 | Research & Development |

# II. METHODS

## ❖ DATA PRE-PROCESSING STEPS

### Step 1: Exploratory Data Analysis

We are going to check the amount of missing values present in all the features of the dataset by using command df.isnull().sum().

```
1 df.isnull().sum()
```

```
EmployeeNumber              0
Attrition                   0
Age                         0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
```

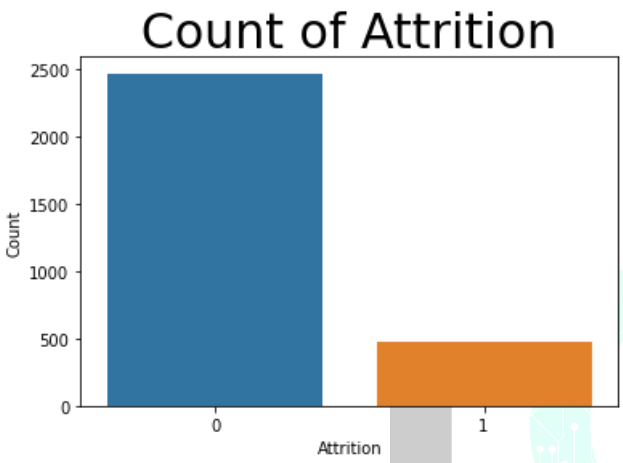
Name- Sukrut Swapnil Mayekar
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - March 2022
Certificate Code- TCRIB2R28
Date of submission- 04-04-2022

**Step 2 : Data Visualization with Plotly and matplotlib library**

We are going to form various inferences by data visualization with Plotly and Matplotlib.

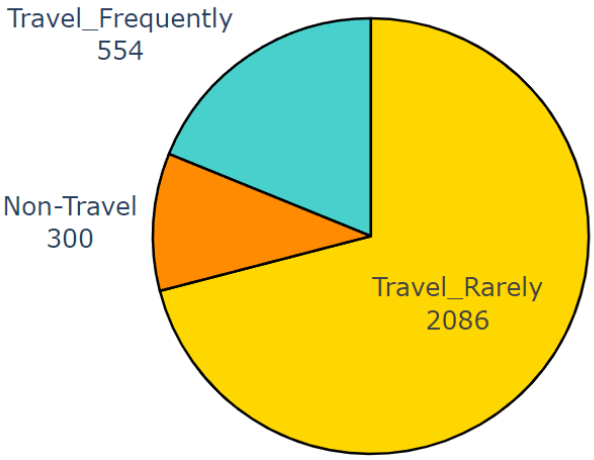
We first have to check how many employees left and remained in the organization

.



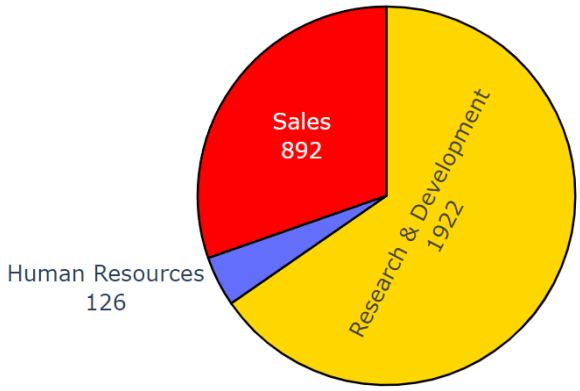

Total 500 employees left the company and about 2500 employees were retained.

Next we want to check the relation between the employees leaving the company and travelling.

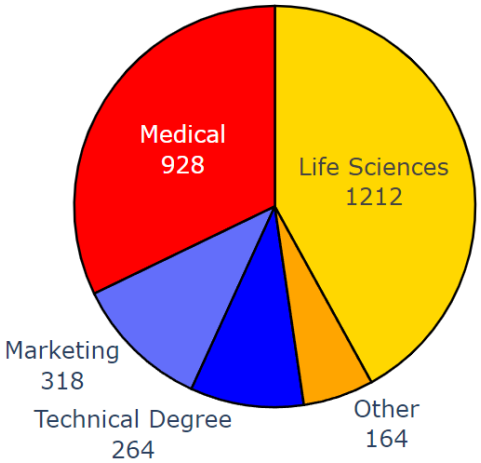



It was amusing because the employees who travelled rarely left the company in huge number rather than employees who travelled frequently.

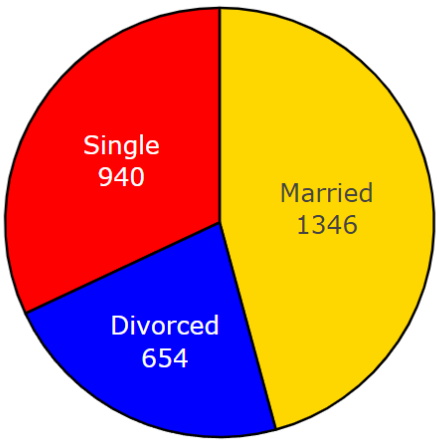
We are going to check that employees from which department chose to leave the organization.




Employees from which career field left the organization.




How many employees who left the company were singles and married and divorced people.



Name- Sukrut Swapnil Mayekar
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - March 2022
Certificate Code- TCRIB2R28
Date of submission- 04-04-2022

**Step 3: Representing the data in Graphical format**

Data visualization is a very important step in data science project. With the help of data visualization, we are able to detect the underlying patterns in the data, which will help us in building our inferences and arrive at certain conclusions.

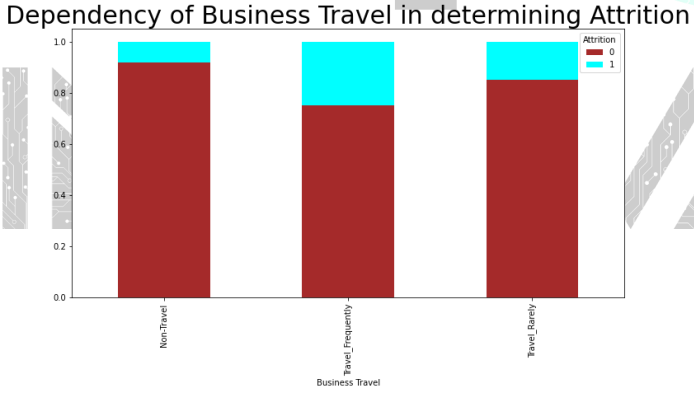
By comparing different input features with attrition target feature, we will be able to understand the many reasons which prompt the employees to leave the organization.

Finding the relationship between 'Attrition' that is our target column which decide whether the given employee will leave or not leave organization to other categorical dataset by using packages of matplotlib, seaborn for clearly classifying the relation between dataset.

By visualising the relation between employee attrition which is our target column and between our input features we will understand the relations between them which will help us in building our model.

We are going to relate the attrition of employees with each feature in the dataset to arrive at inferences and conclusions.

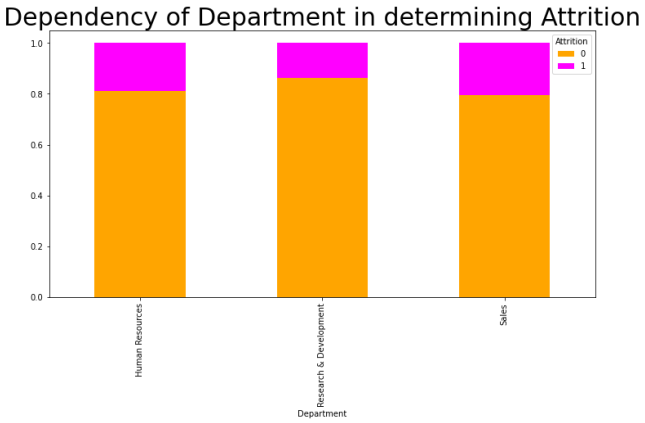
**Relation between Business Travel and Attrition**




By visualising the relation, we find that the employees who travel frequently have lesser chances of attrition.

It seems logical, because those who travel frequently would be tired of travelling as they are investing so much amount of their time, money and energy and thinking of switching to jobs which require less travelling.

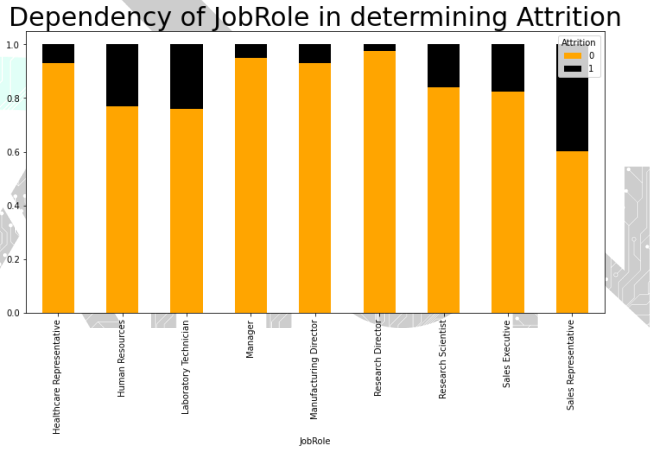
**Relation between Department and Attrition**




Employees working in Human Resources sector and in the sales department have greater chances of Attrition.

This could be their weak relations between the employees or with the upper management.

**Relation between Job Role and Attrition**

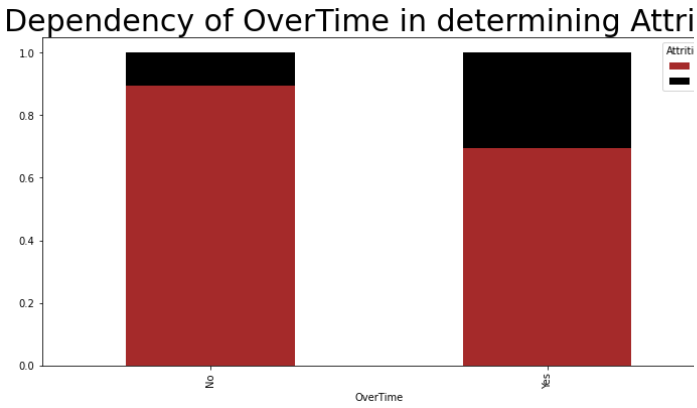



The employee who is working as the sales representative has greater chance of leaving the company followed by Laboratory Technician and HR manager.

This could be because of the work load, working relations or less salary which the employee cannot tolerate and he starts looking for different roles in different organization.

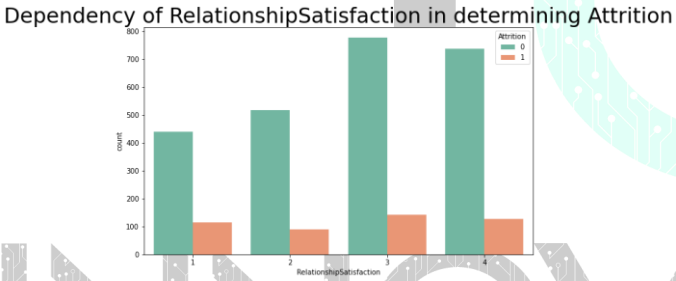
Name- Sukrut Swapnil Mayekar
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - March 2022
Certificate Code- TCRIB2R28
Date of submission- 04-04-2022

**Relation between Overtime and Attrition**



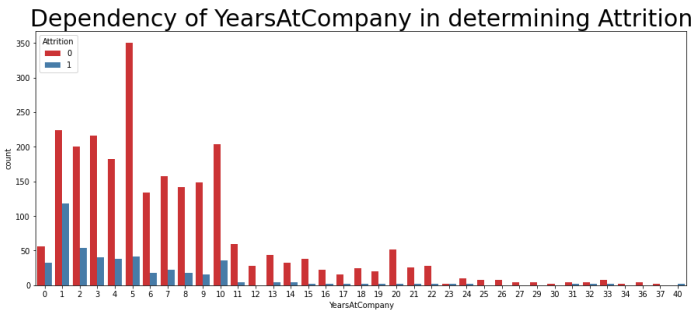

The employee who is working overtime currently is more eligible to leave the company.

**Relation between Relationship Satisfaction and Attrition**



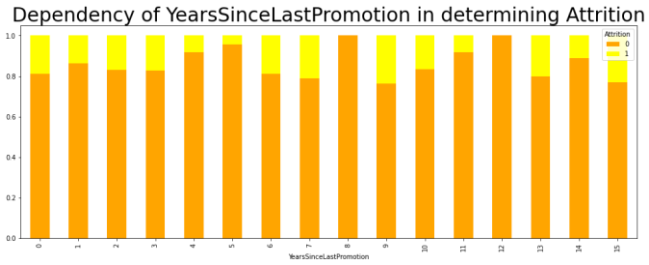

Employees's relationship satisfaction has very less impact in them leaving their jobs. Because employees in approximately equal numbers have left their jobs and also been retained.

**Relationship of YearsAtCompany and Employee Attrition**



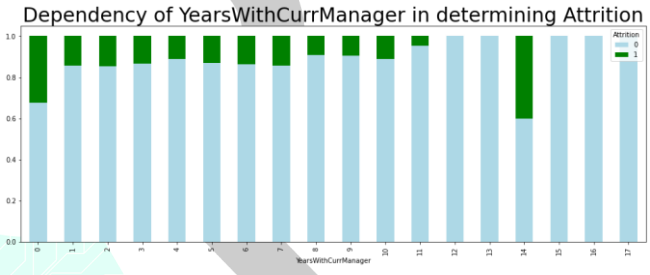

Employees who were previously freshers or who have only one or two years of experience had left their jobs.

**Relationship of YearSinceLastPromotion and attrition**




It shows that employees who received their promotion 15 years ago are more likely to leave their jobs. It also could mean that the employees leaving the company were retiring from their jobs.

**Relationship of YearswithCurrManager and Attrition**




Employees who collaborated with manager for 15 years mostly left their jobs. It means that they were retiring. Along with that it was observed that the employees who collaborated with manager for zero years two left their jobs in huge numbers. It means they were not on common ground with their manager or they grabbed the other career opportunities.

Name- Sukrut Swapnil Mayekar
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - March 2022
Certificate Code- TCRIB2R28
Date of submission- 04-04-2022

- **Data Pre-Processing Steps**

**Step 4: Feature Engineering**

With the help of python, we have separated the dataset features as numerical features and categorical features.

Numerical Features

```
numerical_features = [feature for feature in df.columns if df[feature].dtype != "O"]
numerical_features
```

```
['Attrition',
 'Age',
 'DailyRate',
 'DistanceFromHome',
 'Education',
 'EmployeeCount',
 'EnvironmentSatisfaction',
 'HourlyRate',
 'JobInvolvement',
 'JobLevel',
 'JobSatisfaction',
 'MonthlyIncome',
 'MonthlyRate',
 'NumCompaniesWorked',
 'PercentSalaryHike',
```

Categorical Features

```
categorical_features = [feature for feature in df.columns if df[feature].dtype == "O"]
categorical_features
```

```
['BusinessTravel',
 'Department',
 'EducationField',
 'Gender',
 'JobRole',
 'MaritalStatus',
 'Over18',
 'OverTime']
```

The numerical features can further be subdivided into int values and floating-point values. They can be continuous or they can be discrete.

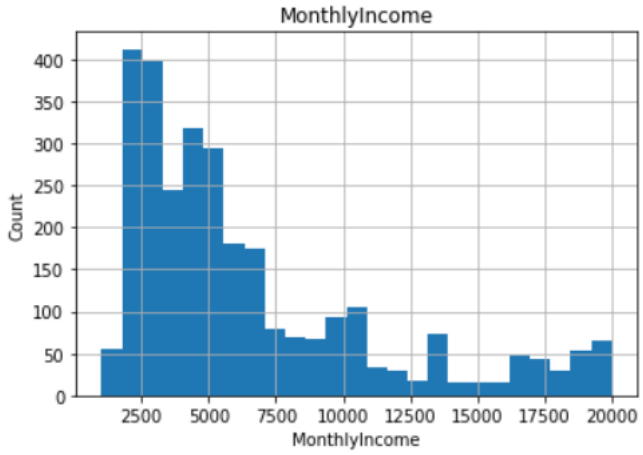
```
continuous_features = [feature for feature in df.columns if len(df[feature].unique()) > 10 and df[feature].dtype != "O"]
continuous_features
```

```
['Age',
 'DailyRate',
 'DistanceFromHome',
 'HourlyRate',
 'MonthlyIncome',
 'MonthlyRate',
 'PercentSalaryHike',
 'TotalWorkingYears',
 'YearsAtCompany',
 'YearsInCurrentRole',
 'YearsSinceLastPromotion',
 'YearsWithCurrManager']
```

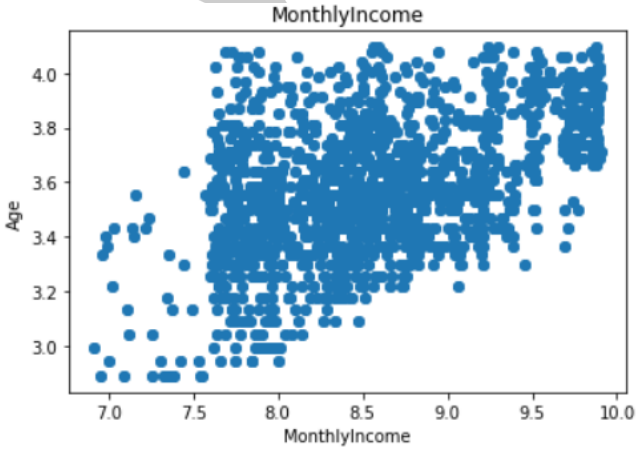
```
discrete_features = [feature for feature in df.columns if len(df[feature].unique()) < 10 or df[feature].dtype == "O"]
discrete_features
```

```
['Attrition',
 'BusinessTravel',
 'Department',
 'Education',
 'EducationField',
 'EmployeeCount',
 'EnvironmentSatisfaction',
 'Gender',
 'JobInvolvement',
 'JobLevel',
 'JobRole',
```

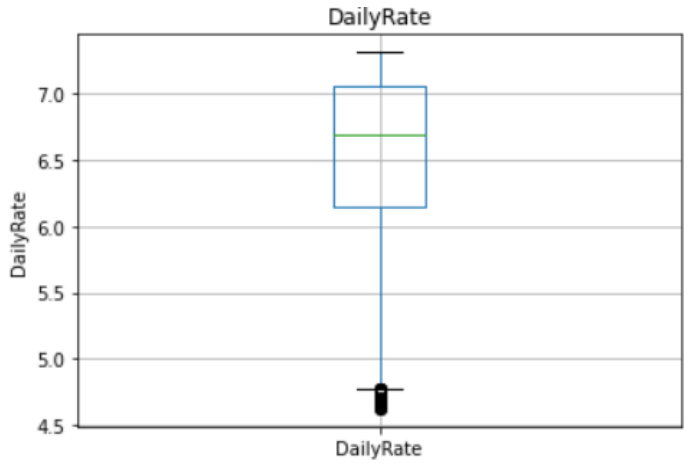
After the categorization of features, we then checked the distributions of this features, that is are the distributions gaussian or continuous in nature or are they exponential or skewed distributions. If the distributions are skewed in nature, then we have to apply log normal to the skewed distribution and convert it into normal distribution.




As we see that the above distribution is skewed in nature, we will apply log to it and convert the distribution to gaussian distribution.

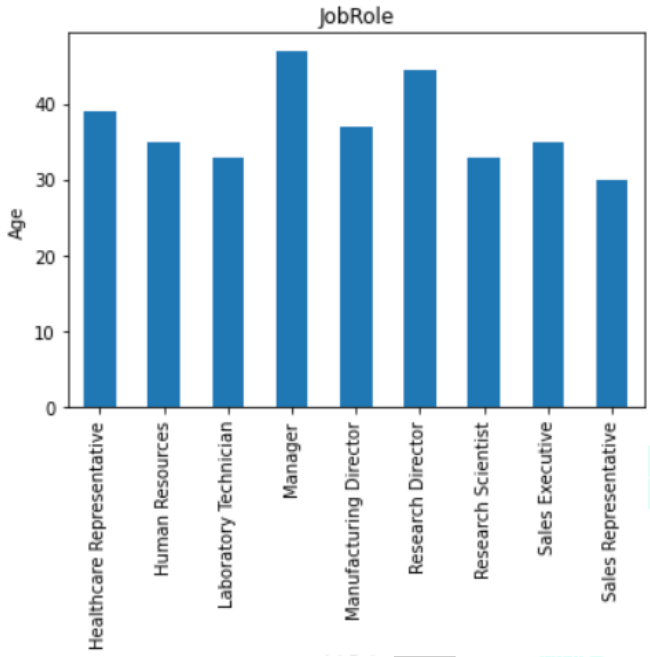



Next step is to visualize the total number of outliers present in our data. In order to detect these outliers, we will use boxplot.



Name- Sukrut Swapnil Mayekar
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - March 2022
Certificate Code- TCRIB2R28
Date of submission- 04-04-2022

Next, we will visualize the employees from which age category take part in which responsibilities in the organization.




The discrete features in our dataset contains discrete values which are not numeric. But when we create our machine learning models in order to predict the target feature ie. Attrition. Our model will not understand the categorical values, but it will understand the numeric values.

So we will use label encoder to convert the categorical values into numeric values.

```
df["Gender"] = label_encoder.fit_transform(df["Gender"])
df["Gender"].head()
```

```
0    0
1    1
2    1
3    0
4    1
Name: Gender, dtype: int64
```

In order to convert multiple discrete values into numeric values, we will use get dummies function in pandas. The reason that we did not used the label encoder is because the label encoder will convert the discrete values into numeric values and will put them in the same feature. Our machine learning model will take the values to be priorities but in reality, they are randomly ordered. This will in turn affect the accuracy of our model.

```
df = pd.get_dummies(df, columns = ['BusinessTravel', 'Department', "EducationField", "JobRole"])
df.head()
```

| EducationField_Human Resources | EducationField_Life Sciences | EducationField_Marketing | EducationField_Medical | EducationField_Other |
|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |

After checking the values in the dataset, we understood that some of the values are two- or three-digit values, while others are single digit values. There is a huge difference between single digit and multi digit values which in turn will affect the accuracy of our model. So, we used MinMaxScaler to put all the values in the range between zero and one. This is turn increased the accuracy of our model.

| | Attrition | Age | DailyRate | DistanceFromHome | Education | EmployeeCount | EnvironmentSatisfaction | Gender | HourlyRate |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0.547619 | 0.715820 | 0.000000 | 0.25 | 0.0 | 0.333333 | 0.0 | 0.914286 |
| 1 | 0 | 0.738095 | 0.126700 | 0.250000 | 0.00 | 0.0 | 0.666667 | 1.0 | 0.442857 |
| 2 | 1 | 0.452381 | 0.909807 | 0.035714 | 0.25 | 0.0 | 1.000000 | 1.0 | 0.885714 |
| 3 | 0 | 0.357143 | 0.923407 | 0.071429 | 0.75 | 0.0 | 1.000000 | 0.0 | 0.371429 |
| 4 | 0 | 0.214286 | 0.350036 | 0.035714 | 0.00 | 0.0 | 0.000000 | 1.0 | 0.142857 |

**Step 5: Feature Selection**

We separated the input features and the target feature "Attrition" before doing so.

```
x = df.drop(columns = "Attrition", axis = 1)
y = df["Attrition"]
```

Before splitting our dataset into training and testing set, we selected the most important features from our dataset. This was done with the help of lasso regression. Instead of selecting all of the features which are not even required and which will hamper the accuracy of our model, we selected the features which are important and which will help us in giving accurate results.

```
selected_feat
```

```
Index(['Age', 'DistanceFromHome', 'EnvironmentSatisfaction', 'Gender',
       'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MaritalStatus',
       'NumCompaniesWorked', 'OverTime', 'RelationshipSatisfaction',
       'StockOptionLevel', 'WorkLifeBalance', 'YearsInCurrentRole',
       'YearsWithCurrManager', 'BusinessTravel_Non-Travel',
       'BusinessTravel_Travel_Frequently', 'Department_Research & Development',
       'EducationField_Medical', 'EducationField_Technical Degree',
       'JobRole_Laboratory Technician', 'JobRole_Sales Representative'],
      dtype='object')
```

We then splitted our dataset with selected features into training set and test set which we will feed into our machine learning models.

Name- Sukrut Swapnil Mayekar
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - March 2022
Certificate Code- TCRIB2R28
Date of submission- 04-04-2022

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state = 42)
```

```
print(x_train.shape, x_test.shape, y_train.shape, y_test.shape)
```

```
(2205, 22) (735, 22) (2205,) (735,)
```

### Step 6: Applying the Machine Learning Models

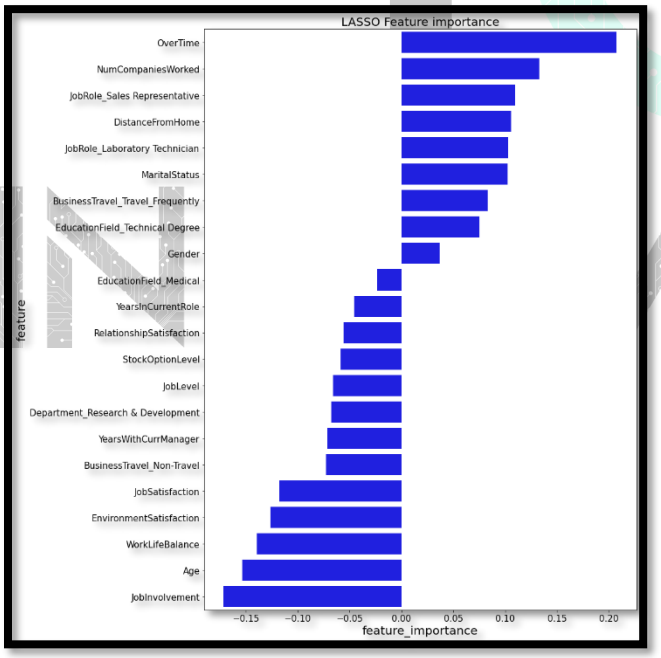
As we have to predict two results, that is if the employee will leave the organization (1) or not (0). This is a problem of binary classification. For classification, we have to use classifier algorithms. Some of them are Decision Trees Classifiers and Random Forest Classifiers. Along with that we have also used Logistic Regression, which is specifically used for Binary Classification purpose.

Using Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier(criterion='gini', max_depth=20, random_state=0)
dtc.fit(x_train, y_train)
```

```
DecisionTreeClassifier(max_depth=20, random_state=0)
```

Along with that we have used Lasso Regression in selecting the most important features from our dataset.




By applying our models on the training and test set, we recorded the accuracy score of our models. As we had performed feature engineering and feature selection, almost all of our algorithms have scored well.

### Step 7: Prediction Results

#### Logistic Regression Results

```
print(f"The test accuracy of the logistic regression model is {test_data_accuracy * 100}")
```

```
The test accuracy of the logistic regression model is 90.61224489795919
```

#### Random Forest Classifier Results

```
score = rf.score(x_test, y_test)
print('Randomforest Classifier', np.abs(score)*100)
```

```
Randomforest Classifier 94.96598639455782
```

#### Decision Tree Classifier Results

```
print('Decision Tree Classifier', np.abs(accuracy_score(y_test, y_predict))*100)
```

```
Decision Tree Classifier 95.91836734693877
```

With Logistic Regression we scored 90 % accuracy in our test dataset, with Random Forest Classifier we scored 94 % accuracy and with Decision Tree Classifier we scored 96 % accuracy.

## III. RESULTS

By applying the machine learning models we have scored pretty good accuracy on the training data set. We can successfully predict if an employee will leave the company or not with great accuracy.

```
predictions = model.predict(input_data_to_array_reshape)
print(predictions)
if (predictions[0] == "0"):
  print("The employee will not leave the company")
else:
  print("The employee will leave the company")
```

```
[1]
The employee will leave the company
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning:

X does not have valid feature names, but LogisticRegression was fitted with feature names
```

```
print(y[2])
```

```
1
```

## IV. Conclusion:

This Research paper tells about the use of Data science in datasets to find out prediction part of data.in data science we can do lots of thing for application part, like we do in our project. The machine learning can be a game changer in the field of data science which helps a lot in processing the data and also for prediction of data. Data analytic is a futuristic job which is growing rapidly from the other field of profession.

## V. References:

D. S. Sisodia, S. Vishwakarma and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," 2017 International Conference on Inventive Computing and Informatics (ICICI), 2017, pp. 1016-1020, doi: 10.1109/ICICI.2017.8365293.

V. Mehta and S. Modi, "Employee Attrition System Using Tree Based Ensemble Method," 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), 2021, pp. 1-4, doi: 10.1109/C2I454156.2021.9689398.

R. Chakraborty, K. Mridha, R. N. Shaw and A. Ghosh, "Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 2021, pp. 1-6, doi: 10.1109/GUCON50781.2021.9573759.

I. K. Nti, J. A. Quarcoo, J. Aning and G. K. Fosu, "A mini-review of machine learning in big data analytics: Applications, challenges, and prospects," in Big Data Mining and Analytics, vol. 5, no. 2, pp. 81-97, June 2022, doi: 10.26599/BDMA.2021.9020028.