HR Analytics Project – Understanding the Attrition in HR

Submitted by:

Nimit Shah

**ACKNOWLEDGMENT**

During completion of this project, I refer various sources like GitHub, Data Trained institute’s reference materials.

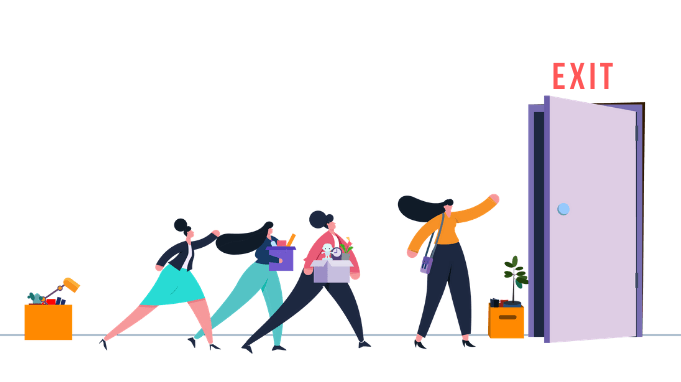
**INTRODUCTION**

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes



**Attrition**  
Employee attrition is when an employee leaves the company through any method, including voluntary resignations, layoffs, failure to return from a leave of absence, or even illness or death.



**Goal of the project**

The goal is to help the HR Analytics department of a company, by understanding the type of employees who have undergone attrition.Thereby making a model to understand the employees that may undergo attrition in the future.

**Problems caused to company due to Attrition**

1. The company has invested time and energy in building up human resource and when the employee leaves, the efforts value would diminish. Regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time.
2. if the business is customer-facing, as customers often prefer to interact with familiar people
3. Errors and issues are more likely if you constantly have new workers.
4. A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them.

**Buisness aim of the project**

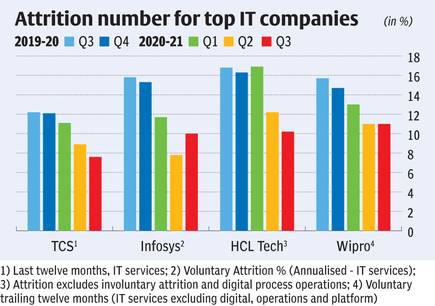
By understanding if a person will undergo attrition:

1.The HR can recruit those who will in future not undergo attrition, thereby avoiding the loss

2.The HR can focus on the employees that might show attrition by providing them benefits and also encouraging them through various means.

**Statistics of Attrition In India**

Attrition for different Indian companies varies between 7 per cent and 17 per cent. “Voluntary attrition for IT services calculated on an annualised basis increased to 15.2 per cent, as demand for talent increased,” Pravin Rao, COO, Infosys, told analysts while discussing the March quarter results 2021



**Models aim**

The aim of the model built will predict the Employee will undergo attrition or not.The model built to predict the Attrition will be aiming to obtain the highest Accuracy score. An accuracy score is fraction of predictions our model got right.

The model will also aim towards obtaining the highest ROC AUC score, precision score and recall.

**Methodology for execution of the modelling**

The data was cleaned and exploratory analysis was done

The data from the data set was split into the train and test data. The train data was then be checked for the best sample state. Using the sample state, the data was split again into the test and train data.

The data was tried to be balanced and it was balanced using under sampling.

This train data was modelled through various algorithms.The algorithms used here are Logistic regression, DecisionTreeClassifier, RandomForestClassifier and SVC. The model with least difference between accuracy score and Cross validation score was used for hyperparameter tuning and then the model was saved. The area under curve of the ROC (ROC AUC) will also be taken into consideration in model selection as a secondary criterion as it is important to distinguish between fraud and legit claims.

**Challenges in the project**

1. The dataset dependent variable is imbalanced
2. The sample size is small. Statistical models are more stable when data sets are larger. It also generalizes better as it takes a bigger proportion of the actual population.
3. the data only capture 3 department

**Criteria for success:**

The **model should be able to classify if attrition would occur or not on a data set**. The independent variables would be about the employee, through which the model would accurately predict if attrition is likely to occur or not. The model is successful if it can predict if employee/candidate will undergo attrition accurately.

**Analytical Problem Framing**

Mathematical/ Analytical Modelling of the Problem

During project building, we run statistical analysis of all available attributes, analyse existing data structure.

Tasks that have been performed from data point of view –

1. Analysis of available data types
2. Visual data analysis
3. Correlation analysis
4. Outlier detection
5. Analysis and definition of the ‘’target’’ variable.

Based on the results and insight obtained regarding these steps, we have a better understanding of what variables we will be able to generate at the data preparation stage and what the system architecture will look like.

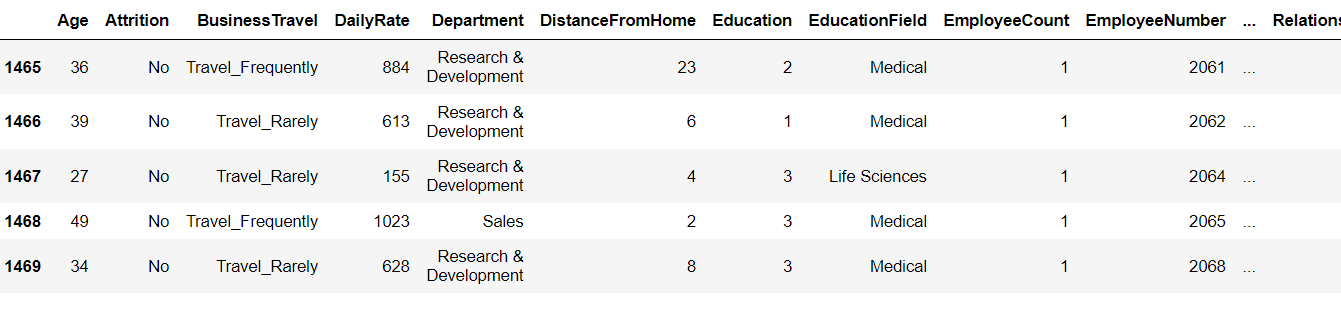
Data Analysis –

The dataset that we are going to used can be found on below mentioned link –

<https://github.com.drscientist/IBM_HR_Attrition_Rate_Analystics>.

The purpose of this project is to identify factors that lead to attrition and then predict employee attrition using multiple machine learning algorithms.

The dataset has thirty-five independent variables that ranges from Age to YearsWithCurrManager.



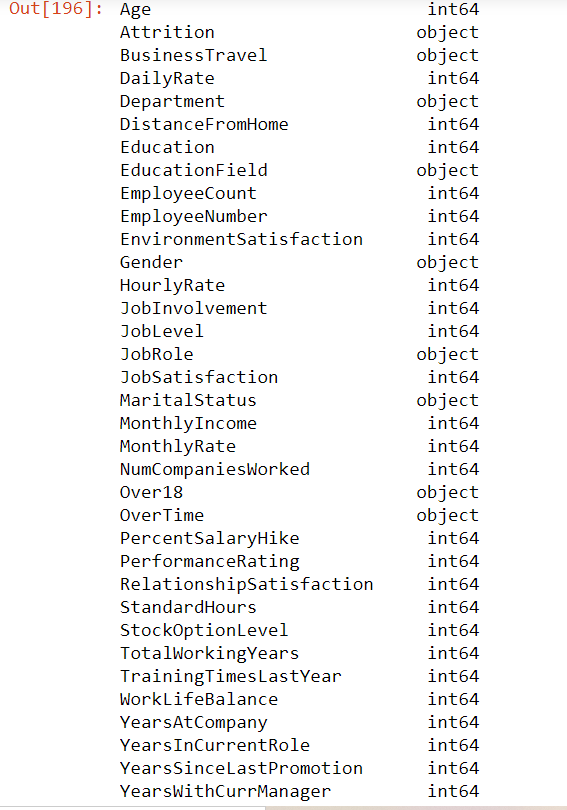
There is total 1470 rows and 35 columns.



BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, Over18, OverTime has object dataset and set columns has integer dataset.

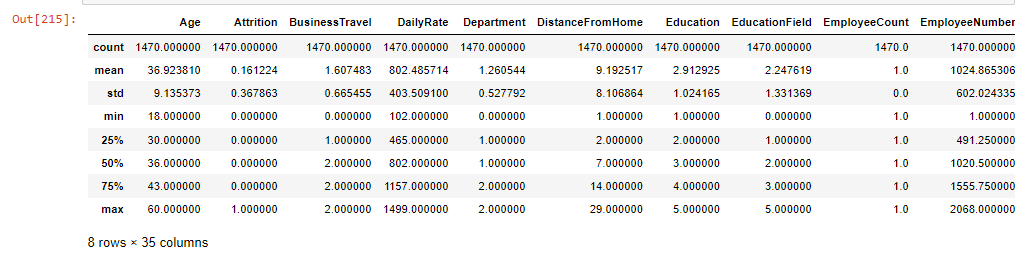
df.dtype helps to know about datatype.





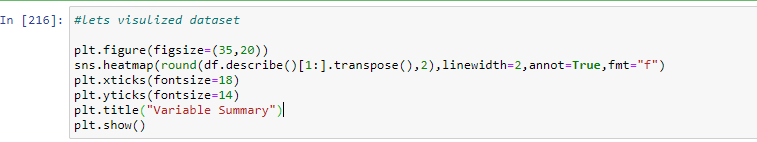
df.describe method gives stastical details like count, mean, std, min, max, 25%, 50%, 75%.

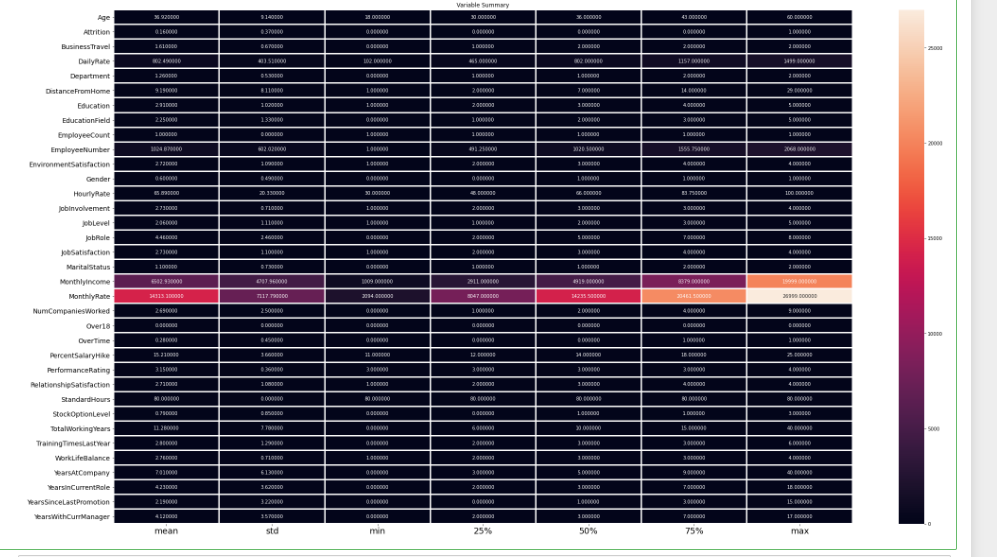




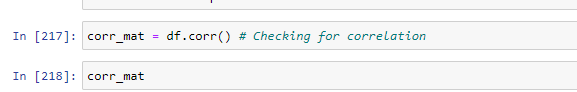
Above outcome shows, there are outlier present in dataset by comparing mean and 50% value. There is major difference between them like in Age, Attrition, BusinessTravel, DistanceFromHome, EmployeeNumber, StockOptionLevel, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole etc.

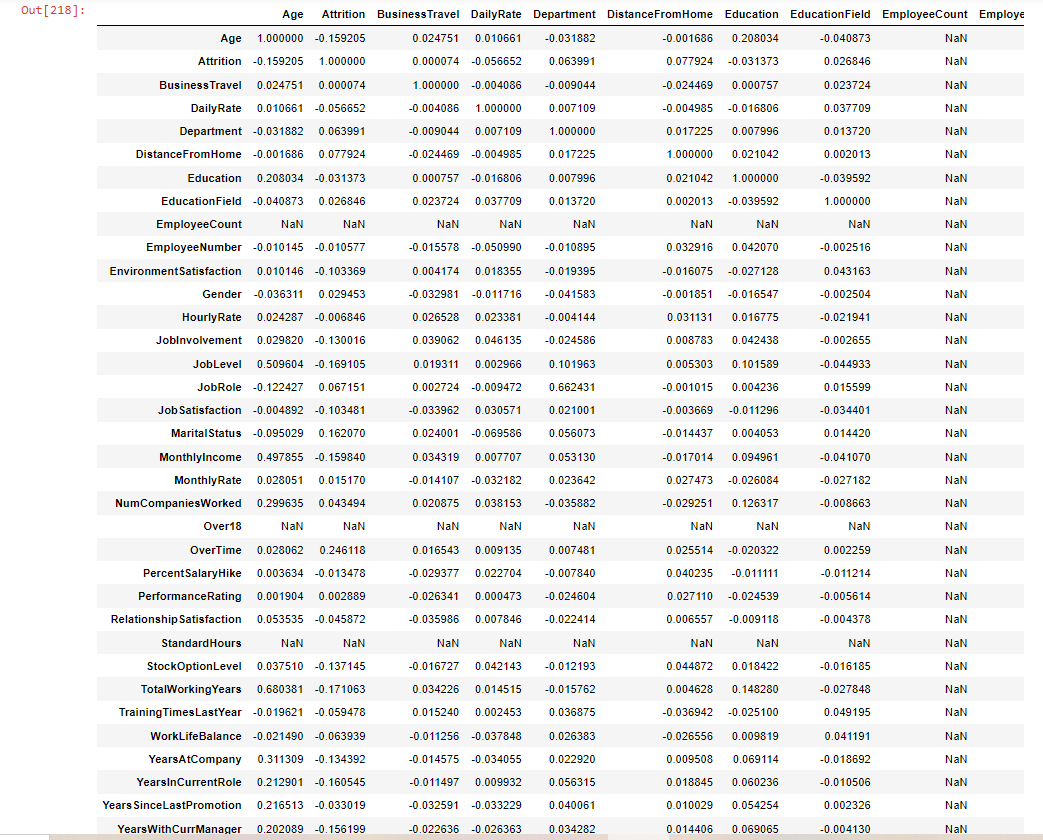
Let’s visualized above variable summary –



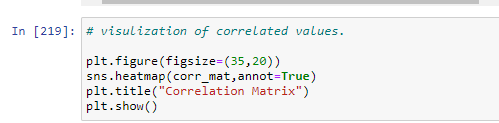


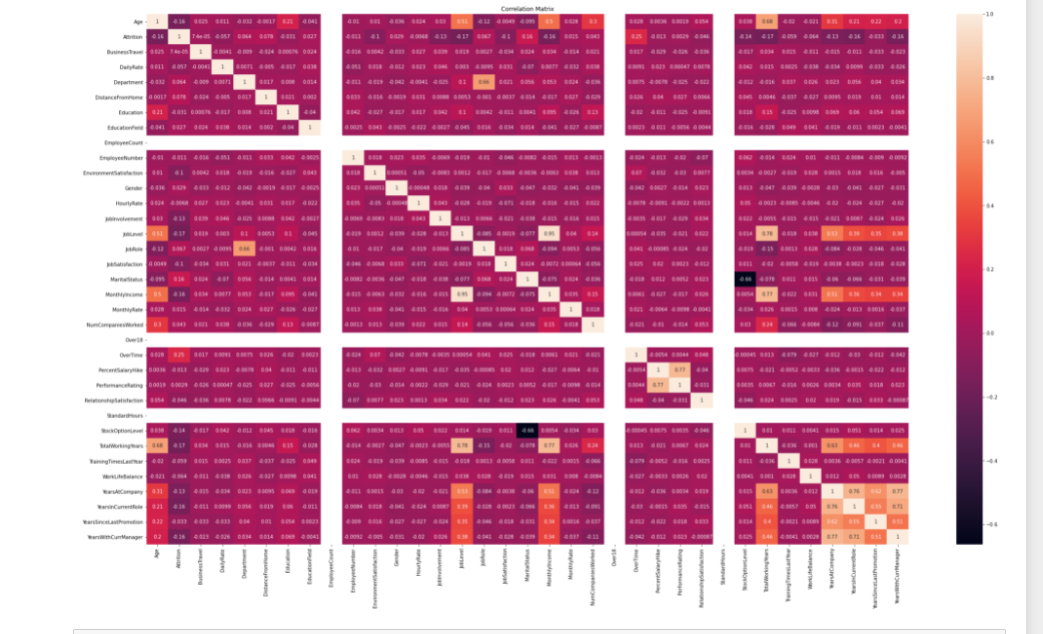
To check for correlation of features with that of label value, we used corr\_mat. Below outcome, we received it–





Let’s visualized it –





All columns of database are positively correlated.

Age has 20%, Attrition has -16%, BusinessTravel has -2.3, DailyRate has -2.6, Department has 3.4%, DistanceFromHome has 1.4%, Education has 5.3%, EducationField has -0.41%, EmployeeNumber has -0.92%, EnvironmentSatisfaction has -0.5%, Gender has-3.1%, HourlyRate has -2%, JobInvolvement has 2.6%, JobLevel has 38%, JobRole has -4.1%, JobSatisfaction has -2.8%, MaritalStatus has -3.9%, MonthlyIncome has 34%, MonthlyRate has -3.7%, NumCompaniesWorked has -11%, OverTime has -4.2%,PercetSalaryHike has -1.2%, PerformanceRating has 2.3%, RelationshipSatisfaction has-0.0087%, StockOptionLevel has 2.5%, TotalWorkingYears has 46%, TrainingTimesLastYear has -0.041%, WorkLifeBalance has 0.28%, YearsAtCompany has 77%, YearsAtCurrentRole has 71%, YearsSinceLastPromotion has 51% coorelation with target value.

Max Correlation: YearsAtCompany

Min Correlation: NumCompaniesWorked

Data Pre-processing

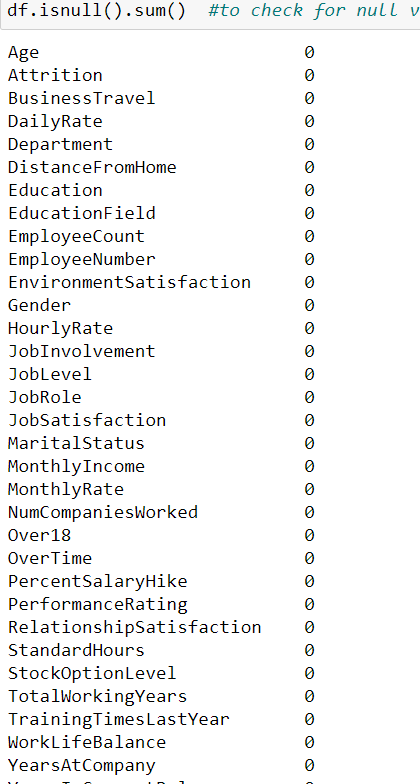
At this stage, the main task is to prepare data for machine learning modelling. It is important to properly aggregate data, create all available variables.

It is also very important to define the target variables.

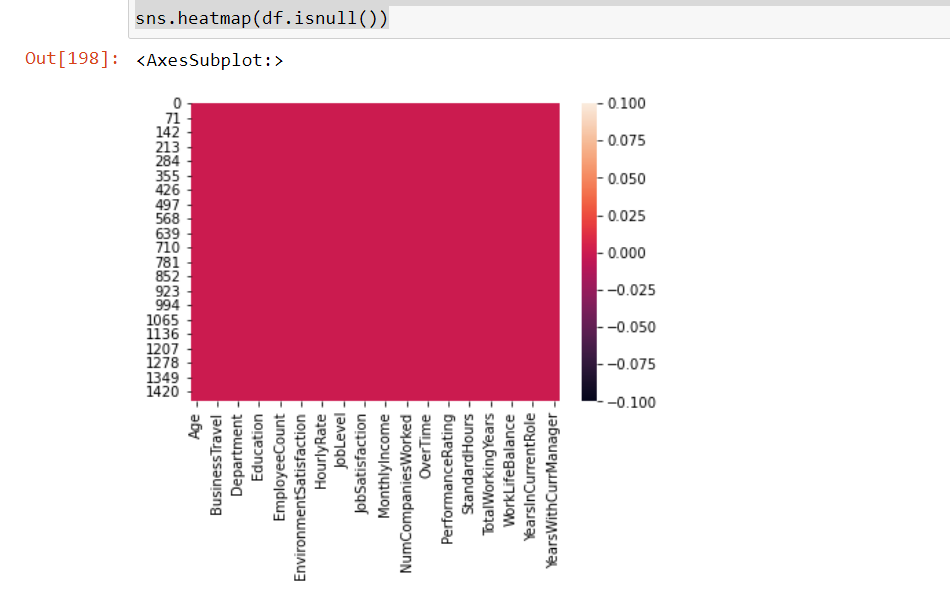
In data processing stage, we checked for dimension of data (df.shape), Type of data (df.info()), Null values (df.isnull().sum()) present in dataset. If null values present in dataset, then fill it with data with help of mean/median or mode methods.

There are no null values present in dataset. Sns.heatmap(df.isnull()) helps to visualized it better.

There are no null values present in this dataset. We can check it with the help of df.isnull().sum().

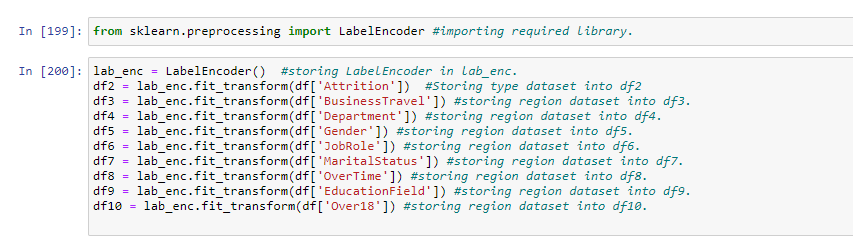


Let's visualized null values with the help of sns.heatmap(df.isnull()). It is clear that dataset does not have any null values because red color is uniformly distributed.

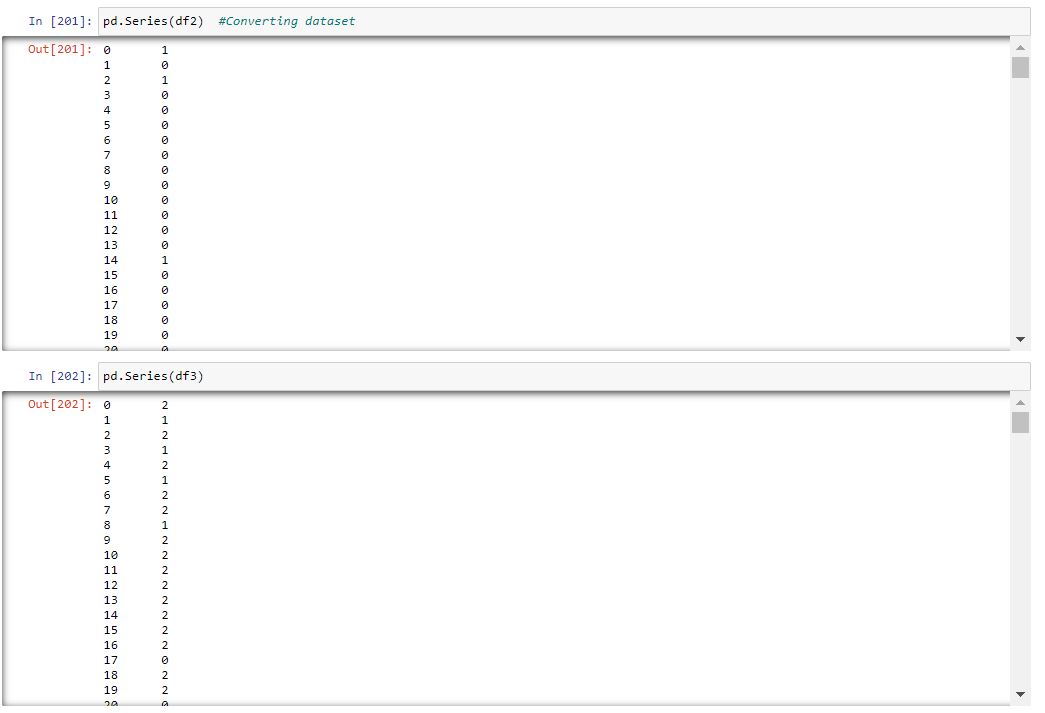


For string/object type of data, it is important to convert it into integer datatype. So for these purpose LabelEncoder() used.

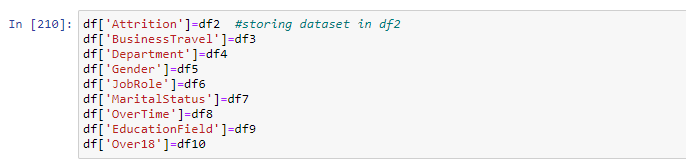
Attrition, BusinessTravel, Department, Gender, JobRole, MaritalStatus, Overtime, EducationaField and Over18 dataset converted from string to integer dataset.



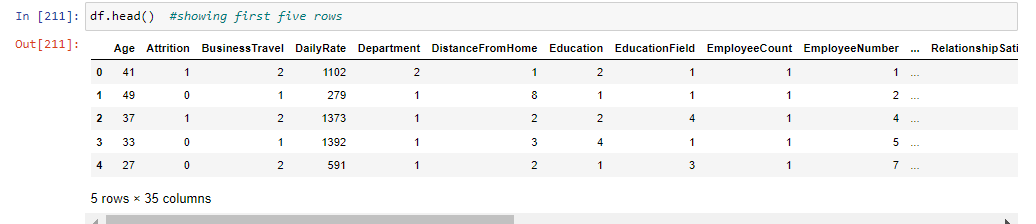
Converting dataset -



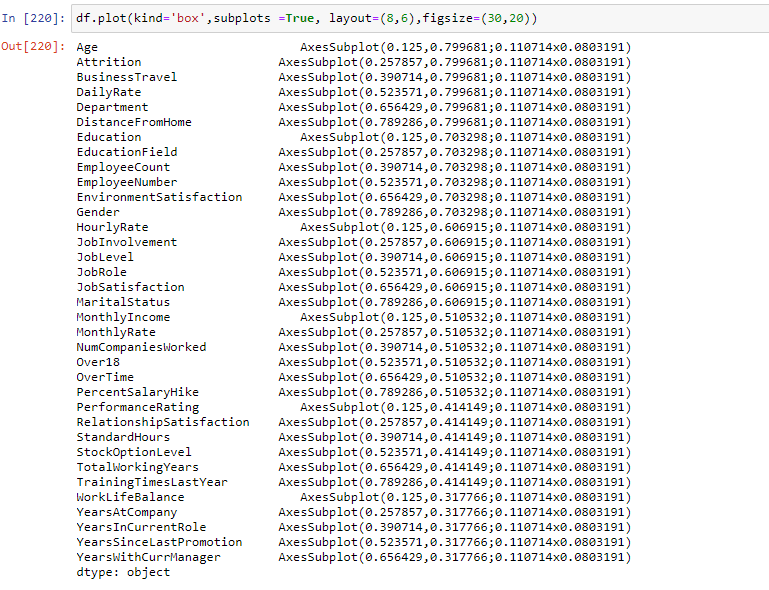
Storing dataset into particular variable –

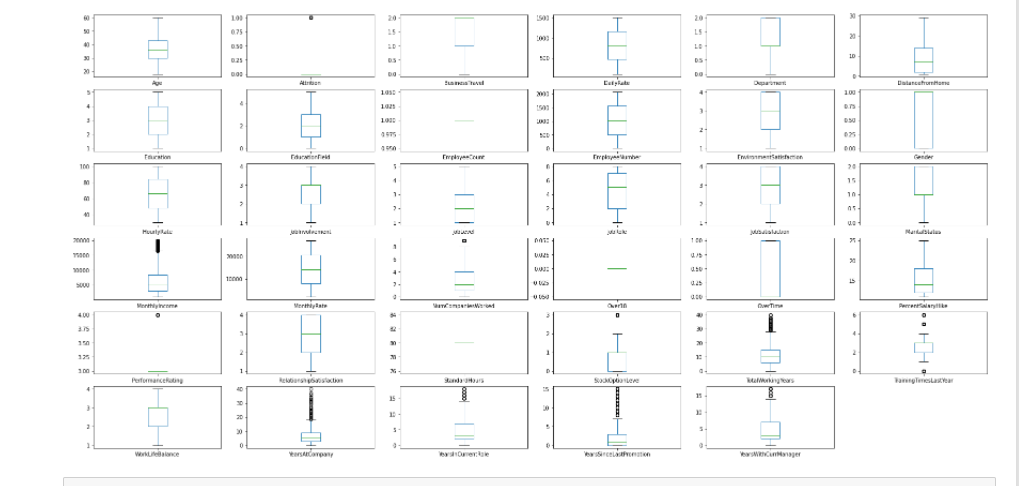


Dataset after converting string dataset into integer one –

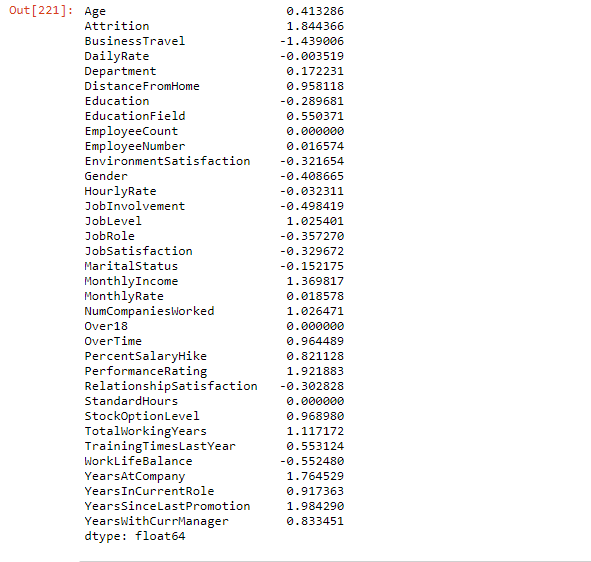


Now visualized outlier present in dataset with the help of box plot –





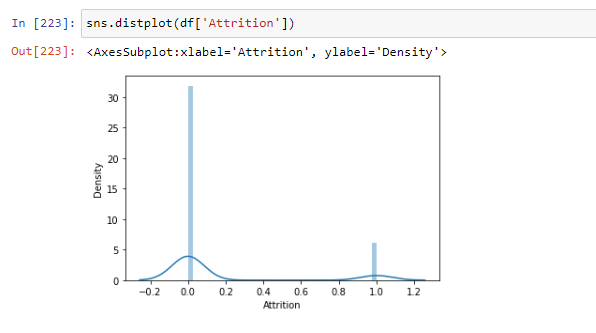
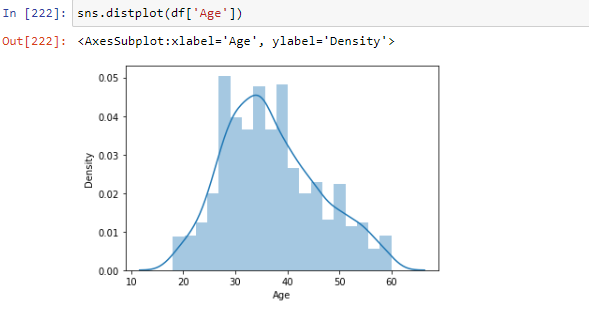
df.skew() helps to check skewness present in dataset –

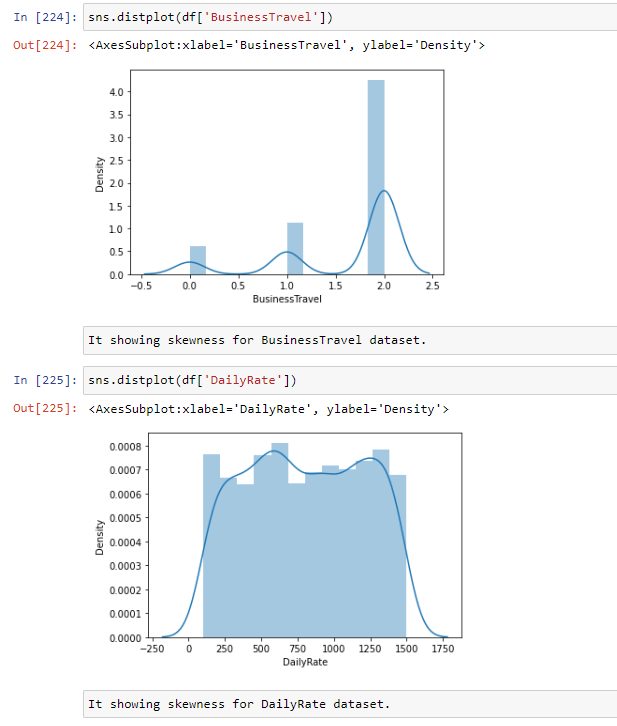
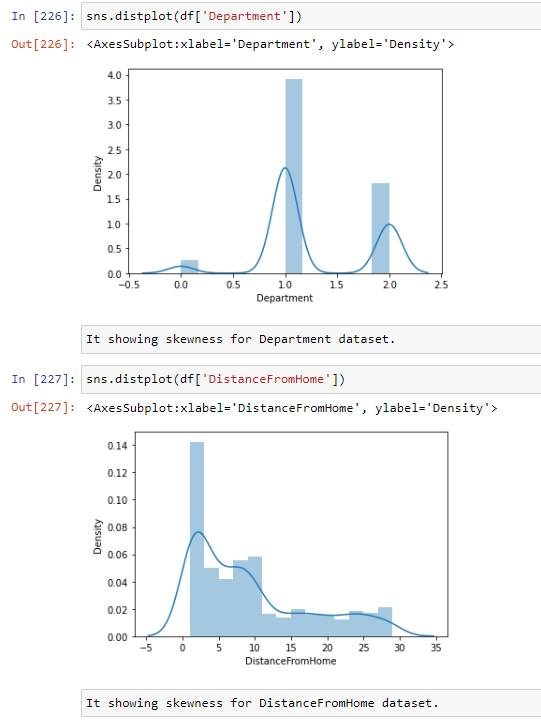


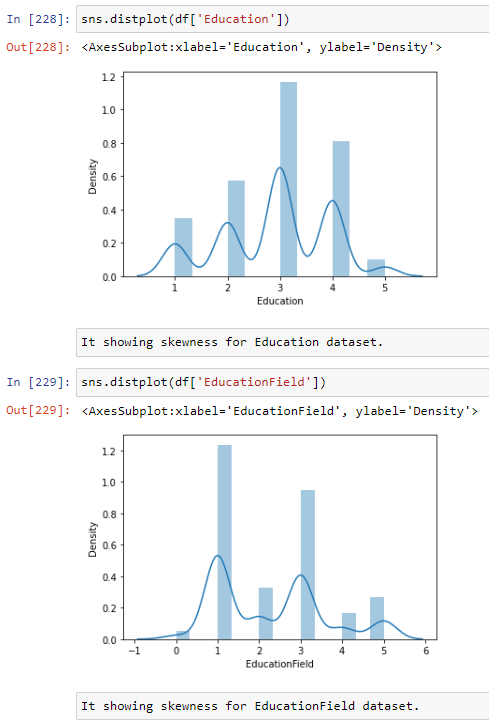
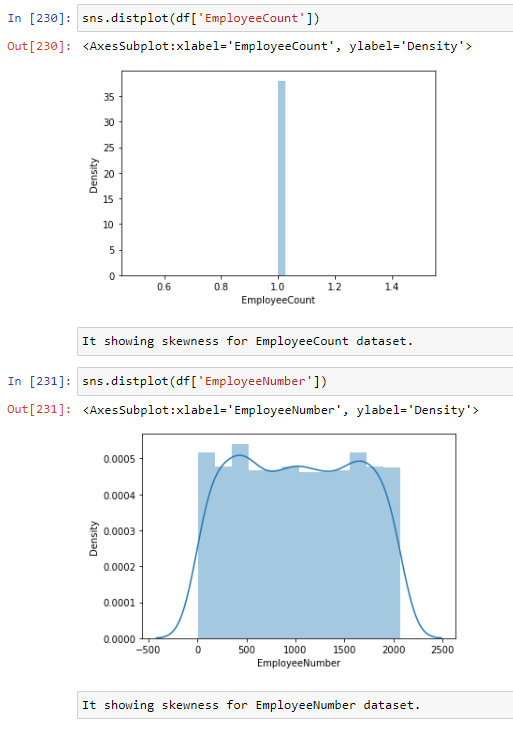
Normalized data range has skewnees ranges between +0.5 to -0.5.

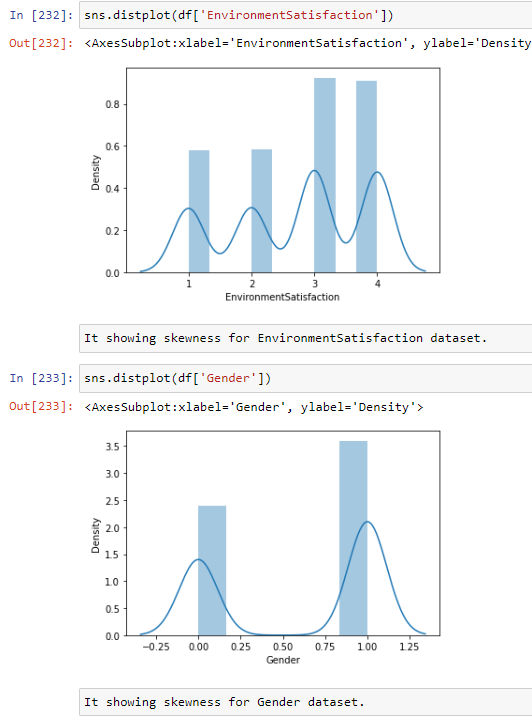
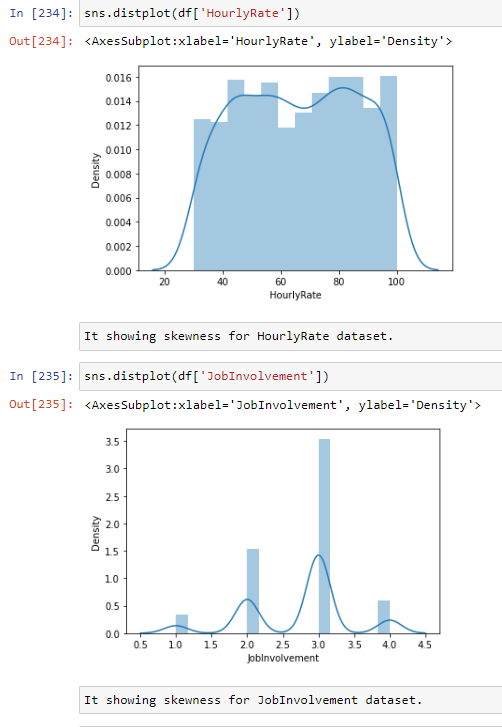
Columns has skewness - Attrition, BusinessTravel, DistanceFromHome,EducationFeild, JobLevel, MonthlyIncome, NumCompaniesWorked, OverTime, PercentSalaryHike, PerformanceRating, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrentManager.

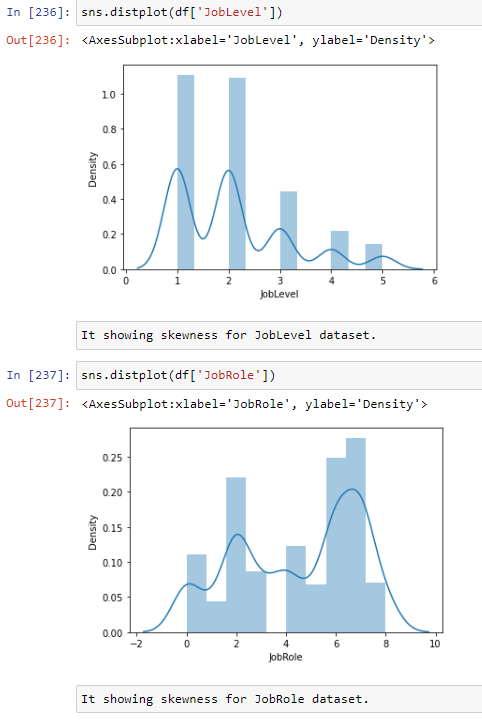
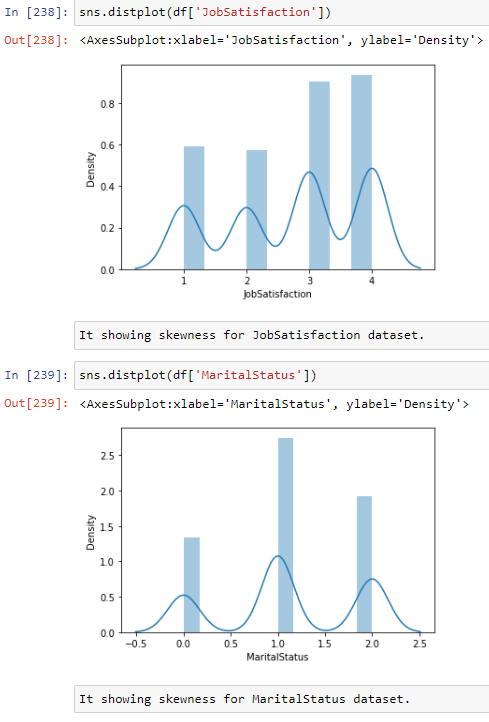
Let’s visualized skewness with the help of distplot.

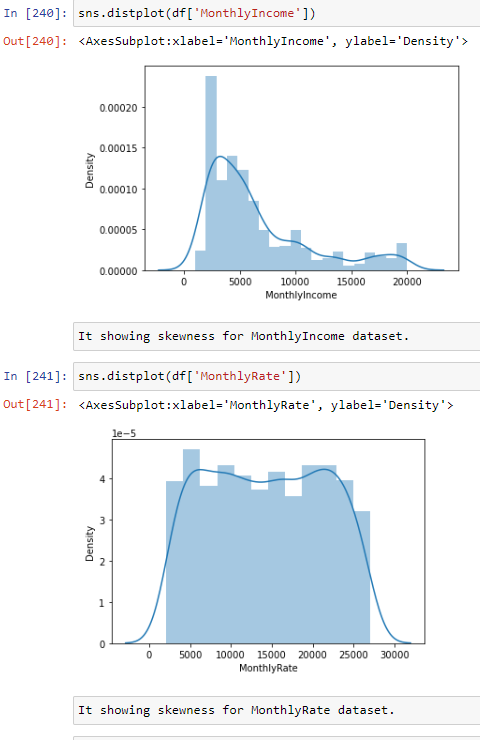
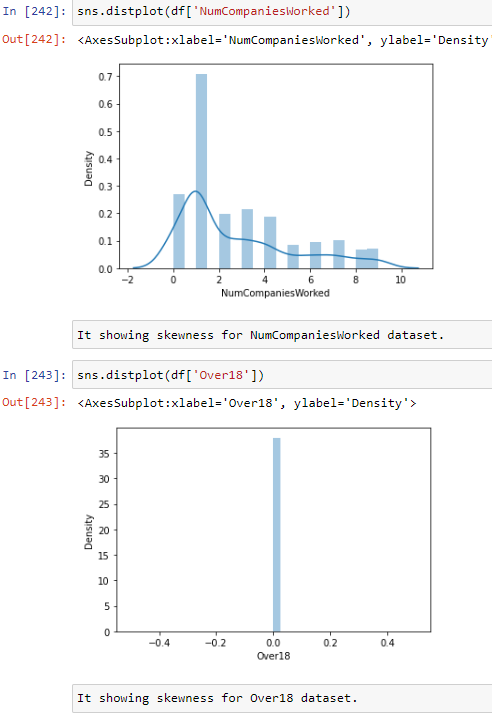


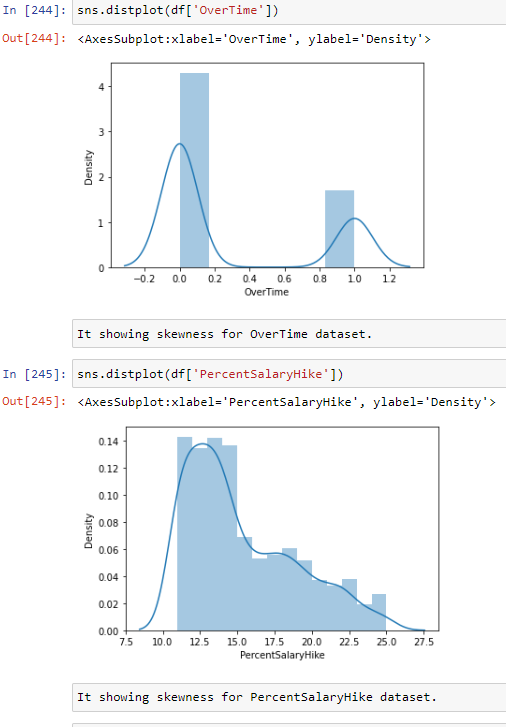
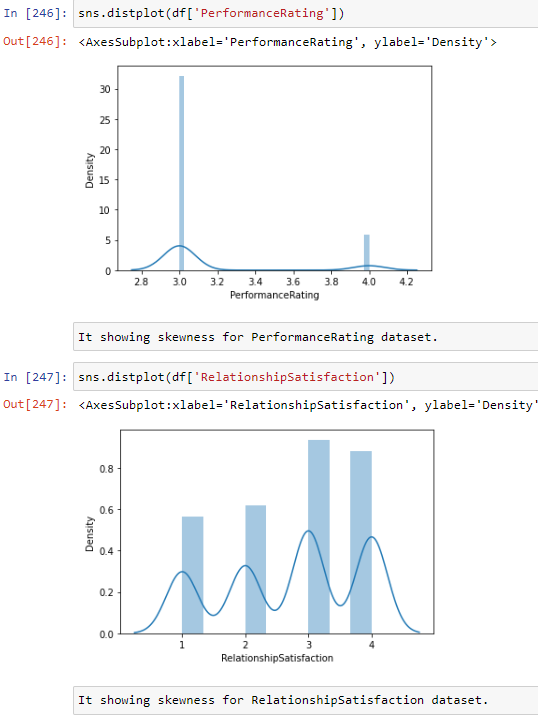


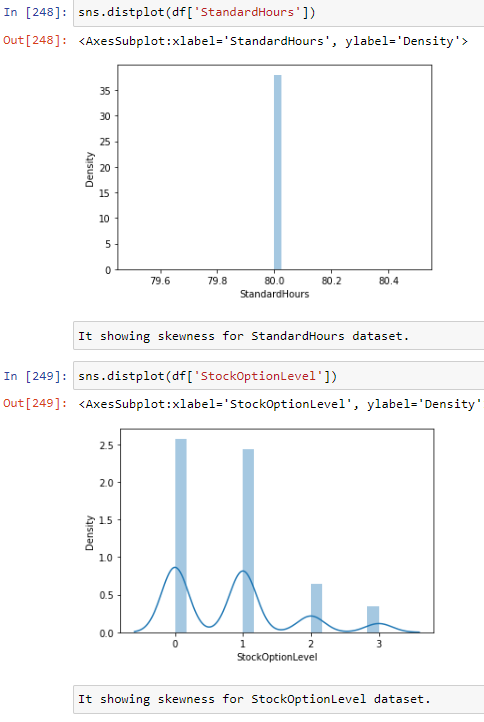
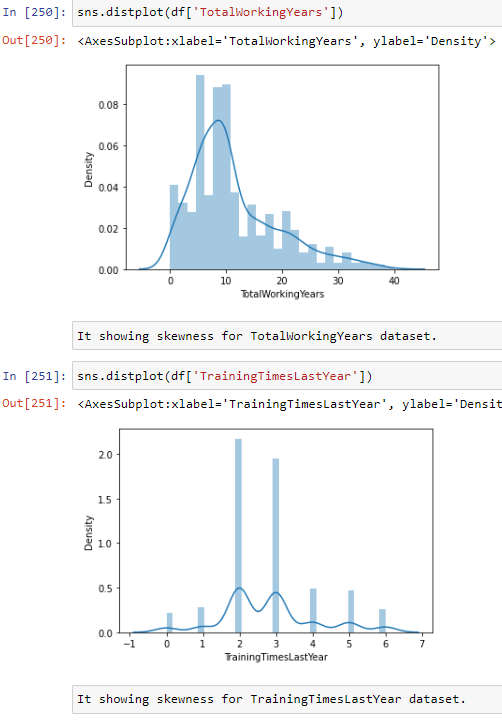


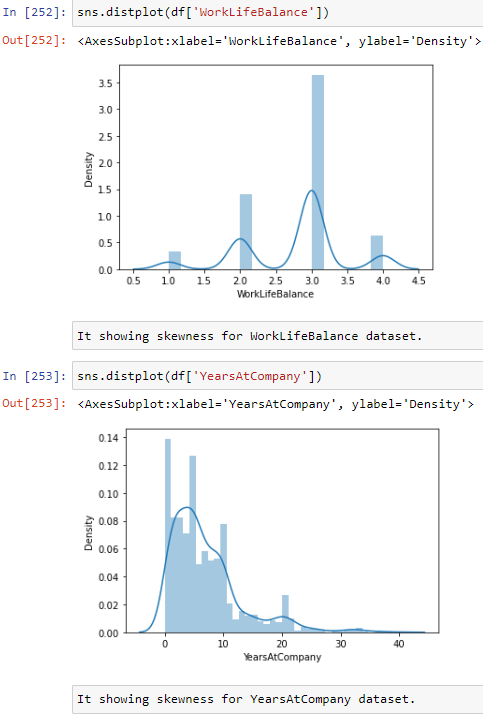
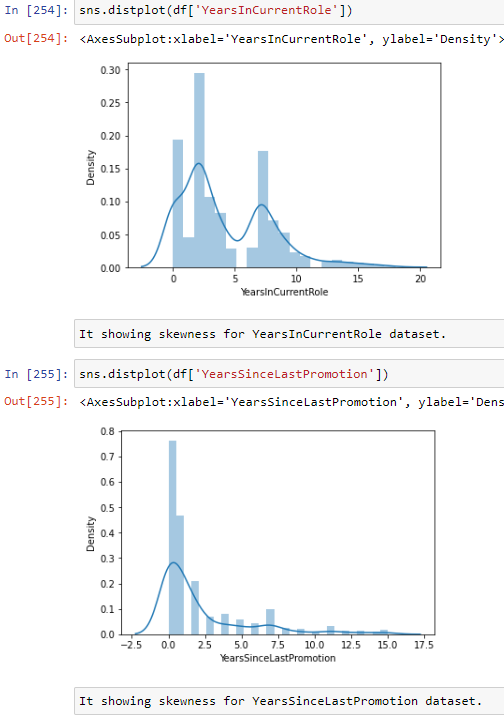


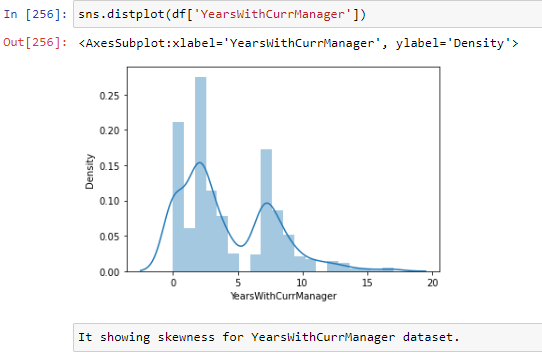




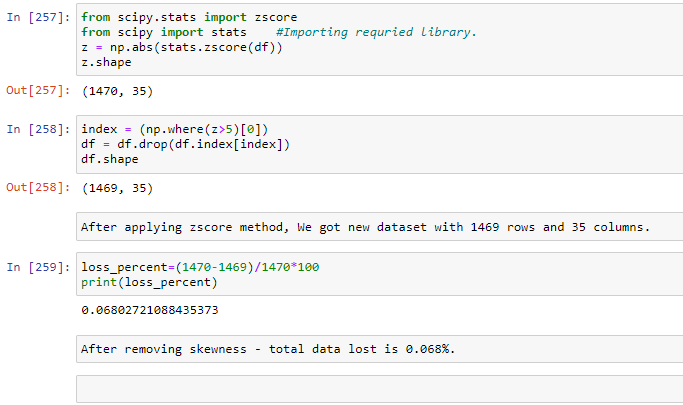




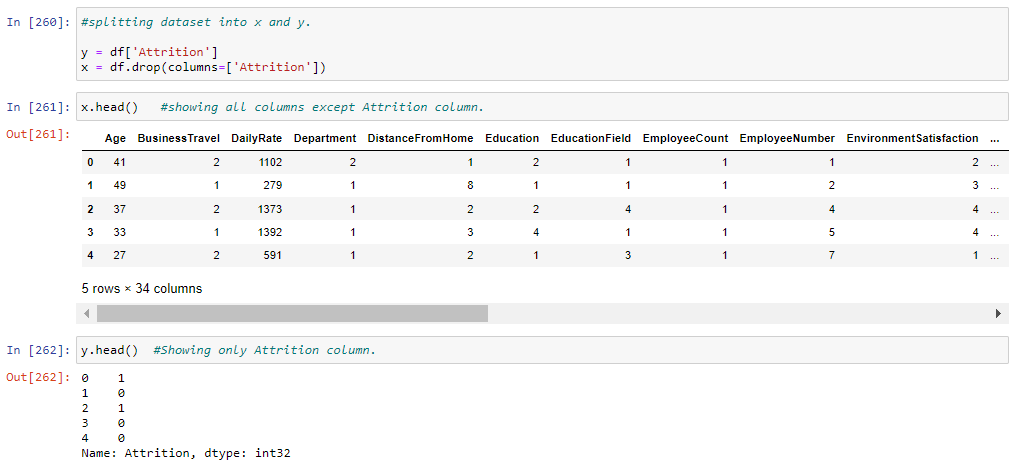




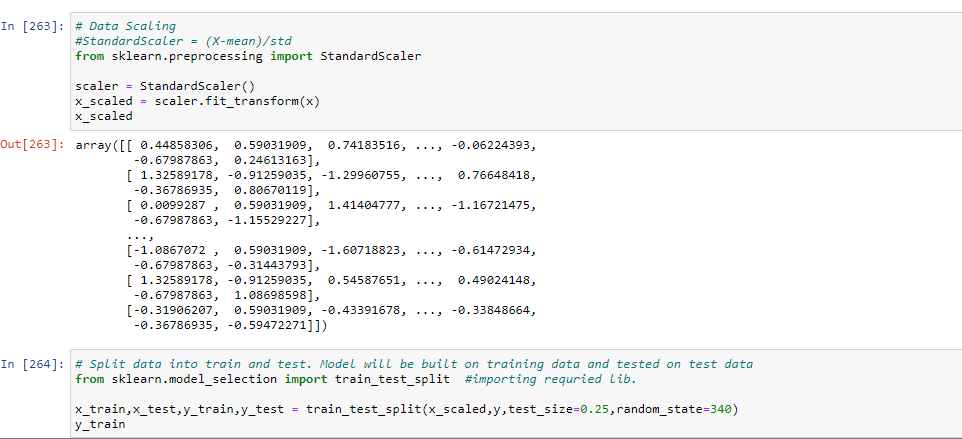
Remove skewness with the help of zscore. After applying zscore 0.068% data is lost.



Dividing dataset into label and features –



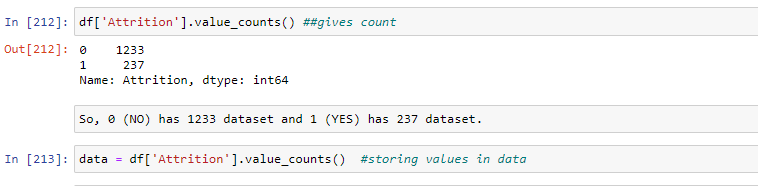
Data standardization and splitting dataset into train and test dataset-

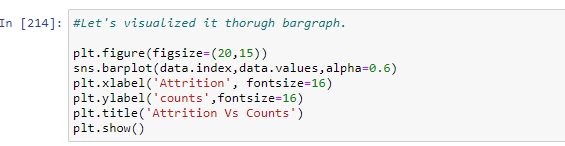


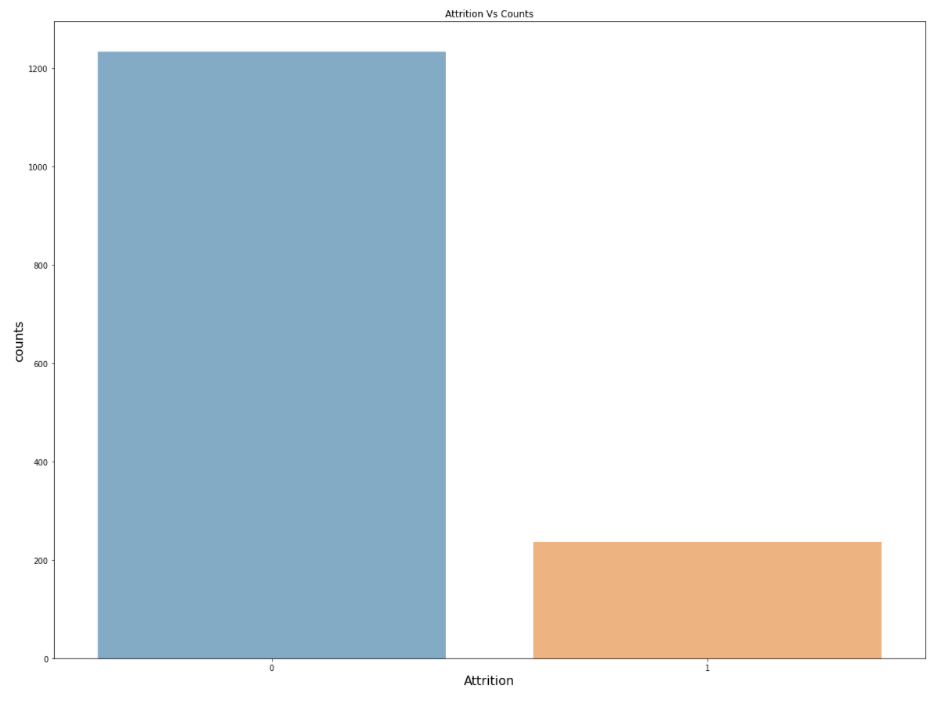
Data Inputs- Logic- Output Relationship

Input parameters (features) and Output (labels/target values) are two important parameters of any dataset. Based on features, target values changed. So, it is important to analysed features parameter to predict correct target values.

In this project, ‘Attrition is taken as output parameter. Output is based on input parameters like Age, BusinessTravel, DailyRate etc







Hardware and Software Requirements and Tools Used

Libraries used while building model are –

1. pandas and numpy – pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values etc. Pandas allows various data manipulation operations such as data cleaning, data wrangling, selecting etc.

numpy provides a multidimensional array object. It can be used for various math operations.

1. Matplotlib.pyplot and seaborn – These are visualization techniques. It helps to plot various graph based on datatypes like scatter plot, Bar graph, distplot etc which are used in this model.
2. Warnings – used to avoid any unnecessary popup while running model.
3. LabelEncoder – It helps to convert string/object dataset into integer dataset.
4. Zscore – Helps to remove skewness present in dataset.
5. Classification\_report, accuracy\_report,

confusion\_matrix – classification\_report is used to measure the quality of prediction from classification algorithm, accuracy\_report gives Number of correct predictions to Total number of predictions, confusion\_matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier.

1. LogisticRegression/KNeighborsClassifier/RandomForestClassifier/DecisionTreeClassifier – It helps for model instantiating and training.
2. Cross\_val\_score – It gives cross validation score.
3. GridSearchCV – It helps to give correct accuracy score for model after adjusting any overperformance of model.
4. Joblib – It helps to save the model.

**Model/s Development and Evaluation**

Identification of possible problem-solving approaches (methods)

At this stage, It is important to create a proper machine leaning model in accordance with best practices. It involves steps –

* Data pre-processing – Clean and transform data into an appropriate format
* Conduct features selection in order to choose the most relevant set of variables.
* Selecting appropriate metrics to measures the performance of the model.
* Train several models
* Validate stability of the model
* Analyse result of model

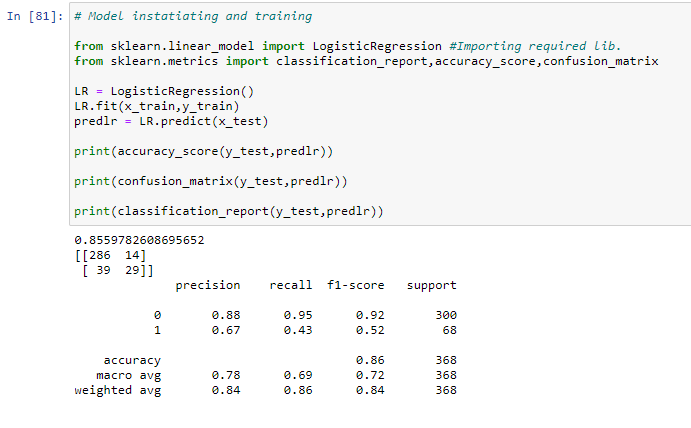
Testing of Identified Approaches (Algorithms)

List of algorithms used in models are –

1. LogisticRegression
2. KNeighborsClassifier
3. RandomForestClassifier
4. DecisionTreeClassifier

Run and Evaluate selected models

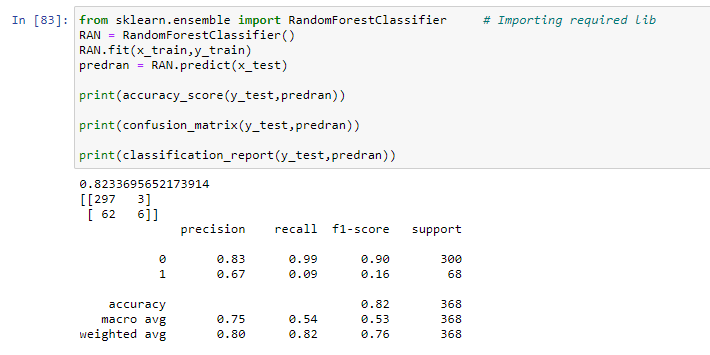
1.LogisticRegression – It is classification Model. It is used when output parameter is discreet.



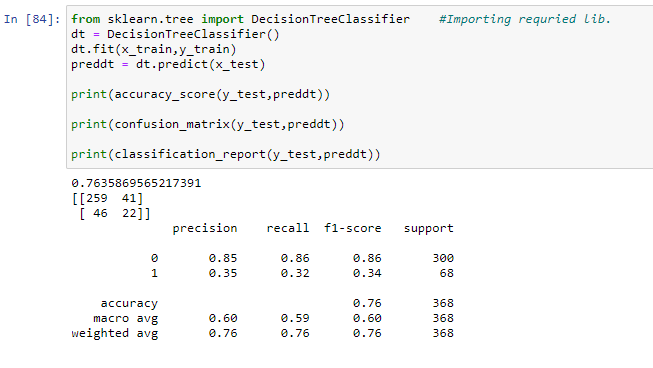
2.KNeighborsClassifier – It is also classification model. It looks for the 5 nearest neighbours.



3.RandomForestClassifier – It is ensemble learning method for classification constructing a multitude of decision trees at training.

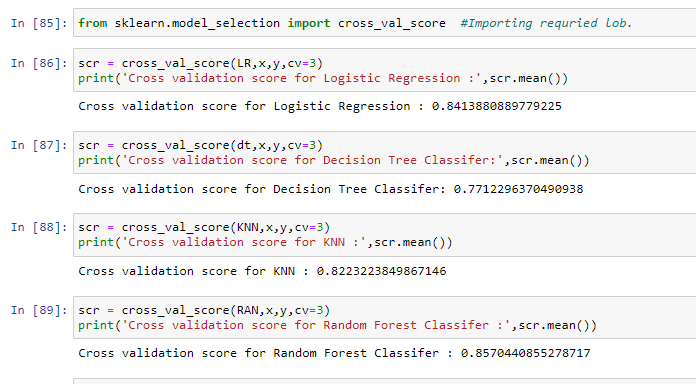


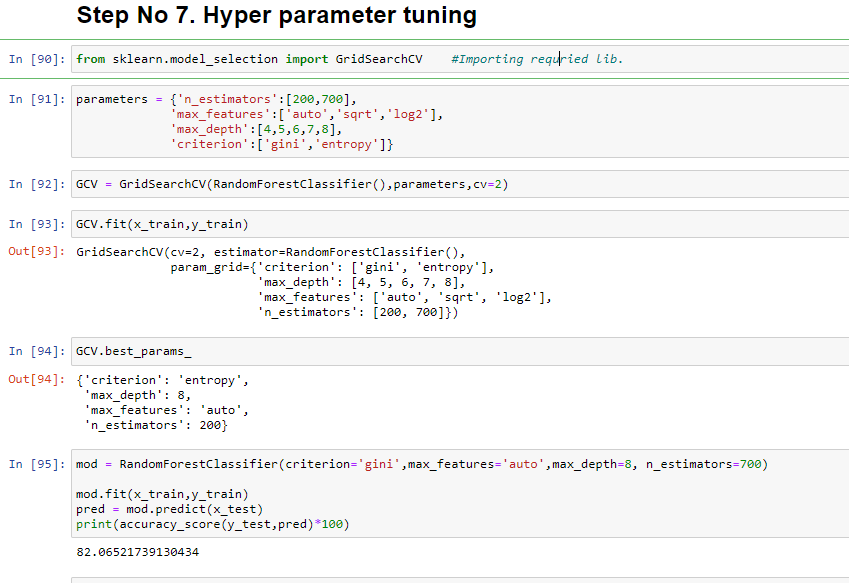
4.DecisionTreeClassifier – It is tree structured classifier, where internal nodes represent the features, branches represent the decision rules and each leaf node represents the outcome.



Key Metrics for success in solving problem under consideration

Cross validation score and Hyperparameter tuning used to avoid any overperformance of model**.** After taking cross validation score into consideration, RandomForestClassifier is best model. So, it is used in hyperparameter tuning. After running hyperparameter tunning, we got model score of 82.06%.





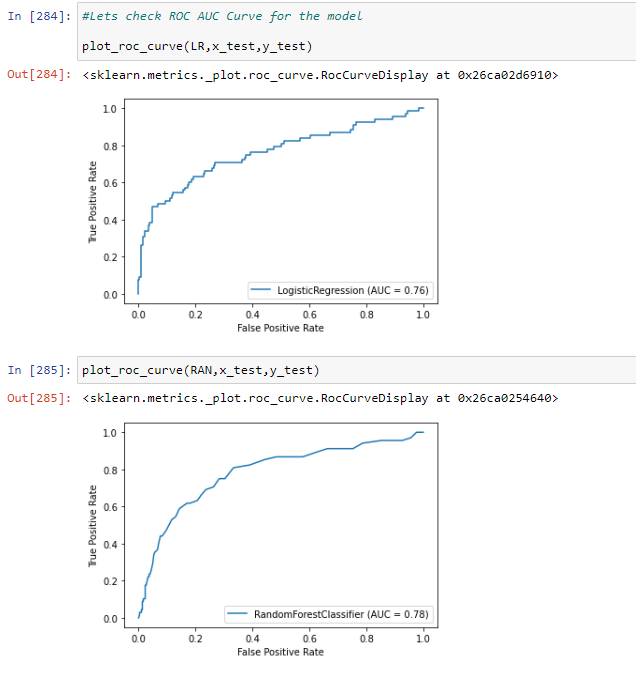
Visualizations

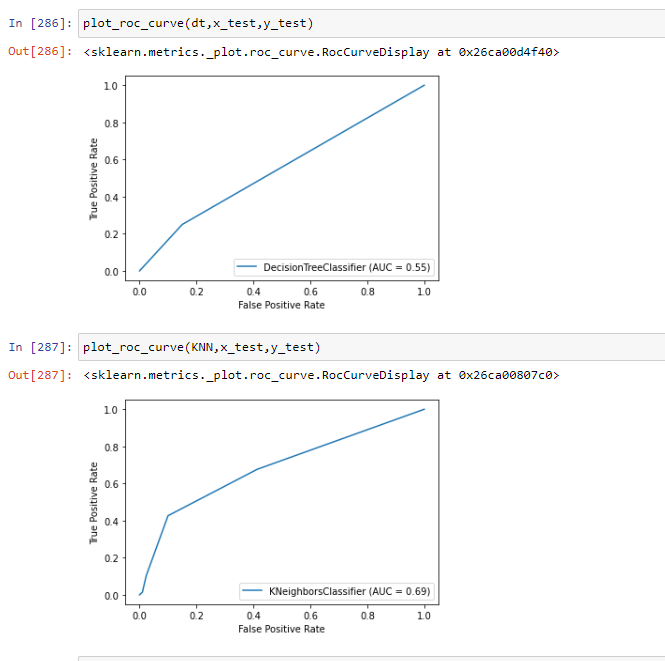
ROC curves typically feature true positive rate on the Y axis and false positive rate on the X axis. This means that the top left corner of the plot is the ideal point – a false positive rate of zero and a true positive rate is one.

The steepness of ROC curve is also important, since it is ideal to maximize the true positive rate while minimize the false positive rate.

ROC curves are typically used in binary classification to study the output of a classifier. In order to extend ROC curve and ROC area to multi-label classification, it is necessary to binarize the output. One ROC curve can be drawn per label, but one can also draw ROC curve by considering each element of the label indicator matrix as a binary prediction.

ROC curves for this dataset for respective models are –





**CONCLUSION**

Employees are the backbone of the organization. Organization's performance is heavily based on the quality of the employees. Challenges that an organization has to face due employee attrition are:

1. Expensive in terms of both money and time to train new employees.
2. Loss of experienced employees
3. Impact in productivity
4. Impact profit

This model will allow the company to calculate the probability of an employee to leave the company and to act on key-factors to avoid departures. The satisfaction of employees and the amount of workload they have to bear seem to be important causes of withdrawals. A particular attention on the work-life balance would be crucial to improve the turnover rate.

The workers with low JobLevel, MonthlyIncome, YearAtCompany, and TotalWorkingYears are more likely to quit their jobs.

BusinessTravel : The workers who travel alot are more likely to quit then other employees.

Department : The worker in Research & Development are more likely to stay then the workers on other departement.

EducationField : The workers with Human Resources and Technical Degree are more likely to quit then employees from other fields of educations.

Gender : The Male are more likely to quit.

JobRole : The workers in Laboratory Technician, Sales Representative, and Human Resources are more likely to quit the workers in other positions.

MaritalStatus : The workers who have Single marital status are more likely to quit the Married, and Divorced.

OverTime : The workers who work more hours are likely to quit then others.

**Github link:**

<https://github.com/Nimit96/Evaluation_Projects/tree/main/HR%20Attrition>