**Building a Model in Python to Predict Attrition in a Company**

Alisha Fal Desai

# Contents

# Problem Definition

# Data Analysis

# EDA Concluding Remarks

# Pre-processing Pipeline

# Building Machine Learning Model

# Concluding Remarks

# Problem Definition

# Employee attrition refers to a loss of employees voluntarily or involuntarily through a natural process, such as retirement, elimination of a position, personal health, or other similar reasons. It leads to a reduction in the strength of the workforce.

# There are two sides to employee attrition: positive and negative. When low performing employees leave the company or are fired, it is positive attrition. However, when top-performing employees responsible for driving sales and increasing revenue, due to various circumstances, start looking for the exit, it is known as negative attrition. Negative attrition generally implies a larger, more serious problem within the company.

# One study conducted by FurstPerson illustrates the financial toll employee attrition can take on an organization. Across several major industries, staff turnover can cost anywhere from $1,500 to 416,650 per agent.

# Attrition can bring down the motivation and morale of the company by creating a leadership gap in the organization if the senior employees are the ones leaving. It can affect the diversity of the workplace if several employees of a minority group are the ones leaving. The position that the employee vacates may not always be possible to be filled immediately.

Understanding why and when employees are most likely to leave can lead to actions to reduce employee attrition as well as possibly planning to hire new employees to fill vacancies in advance. We will be using a method that could be used for a variety of Machine Learning problems, following a step-by-step systematic approach. This project would fall under what is commonly known as **HR Analytics.**

# In this project, we will attempt to solve the following problems:

* What is the likelihood of an active employee leaving the company?
* What are the key indicators of an employee leaving the company?

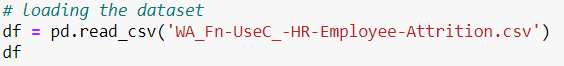
Given that we have data on former employees, this is a**standard supervised classification problem** where the label is 'Yes'(former employee) and 'No'(active employee). Here the target variable is 'Attrition', whether a particular employee is going to leave the company or not.

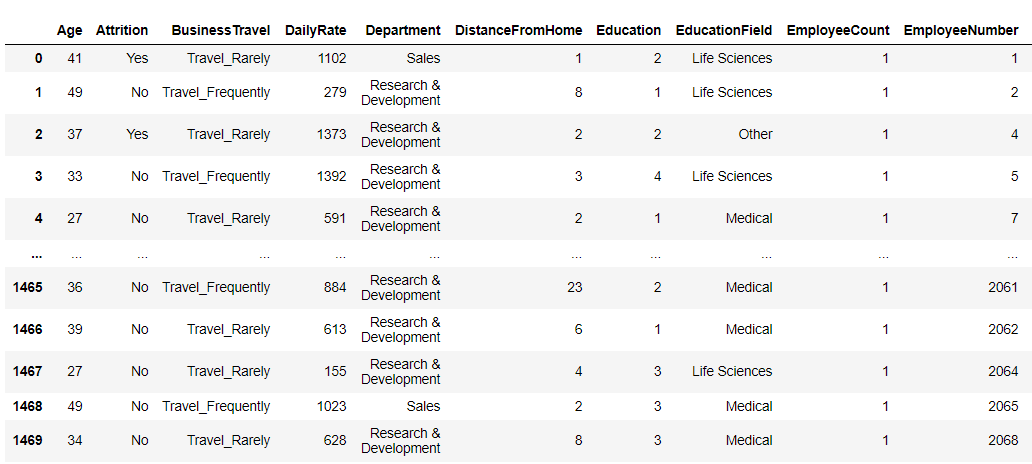
# Data Analysis

For this project, we have used ‘IBM HR Analytics Employee Attrition Performance’ Dataset, which was downloaded from GitHub, which contains the data for 1470 employees. The features in the dataset include ‘Age’, ‘Attrition’, ‘Department’, ‘Education’, ‘EmployeeCount’, ‘Gender’, ‘JobRole’, ‘JobSatisfcation’, ‘MonthlyIncome’, ‘PercentSalaryHike’, ‘PerformanceRating’ and ‘YearsAtCompany’ among others. We will study the features of this dataset to predict whether a particular employee is going to leave the company or not.

**2.1) Data Description**

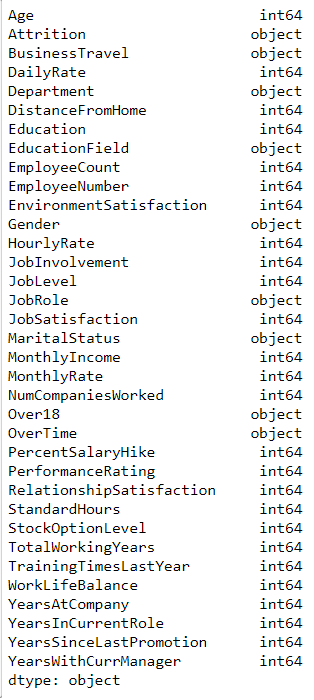
First we import the csv file. The dataset has 1470 rows and 35 columns.





The dataset contains several numerical and categorical features that give the details of the employee. Fortunately, none of the columns of the dataset contain any missing values.

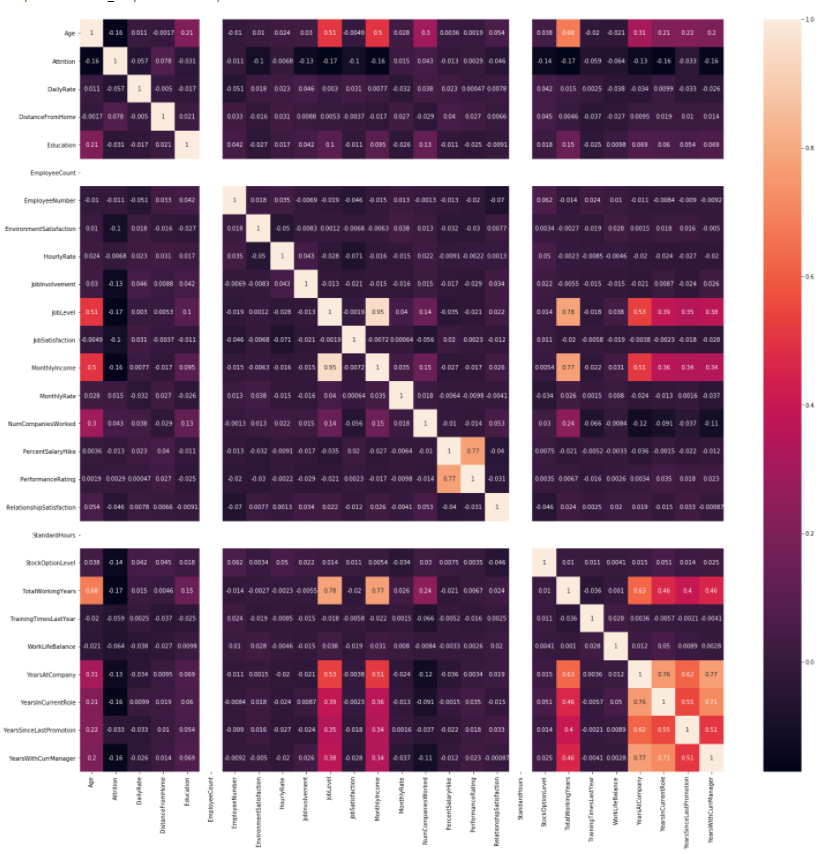




**2.2) Correlation**

In this project, we are hoping to predict the employee attrition, and in order to do that we need to know how the features are correlated to each other. We are plotting a seaborn heatmap of the correlation of the numeric features of that dataset to study the correlation between them.





From the above heatmap, we can see that JobLevel, TotalWorkingYears, YearsWithCurrManager, YearsInCurrentRole are negatively correlated to Attrition. And DistanceFromHome, MonthlyRate, NumCompaniesWorked are positively correlated to Attrition. None of the features are strongly correlated to the target variable.

**2.3) Categorical Features**





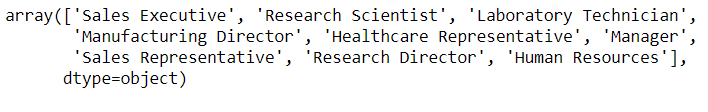






















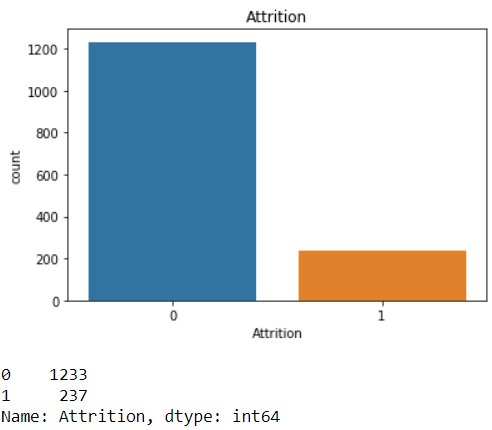


From the above blocks of code, we can see the unique values of the categorical features. We can see that all the employees are above 18 years of age. The ‘EmployeeNumber’ column refers to the unique number given to every employee. We can drop these two columns as they do not affect the target variable. The dataset also tell us whether an employee has opted for overtime or not, along with the department they currently work or used to work in before they left.

Next, we perform Exploratory Data Analysis (EDA) on the numerical and categorical features of the dataset.

**2.4) EDA**

**2.4.1) Target Variable: Attrition**

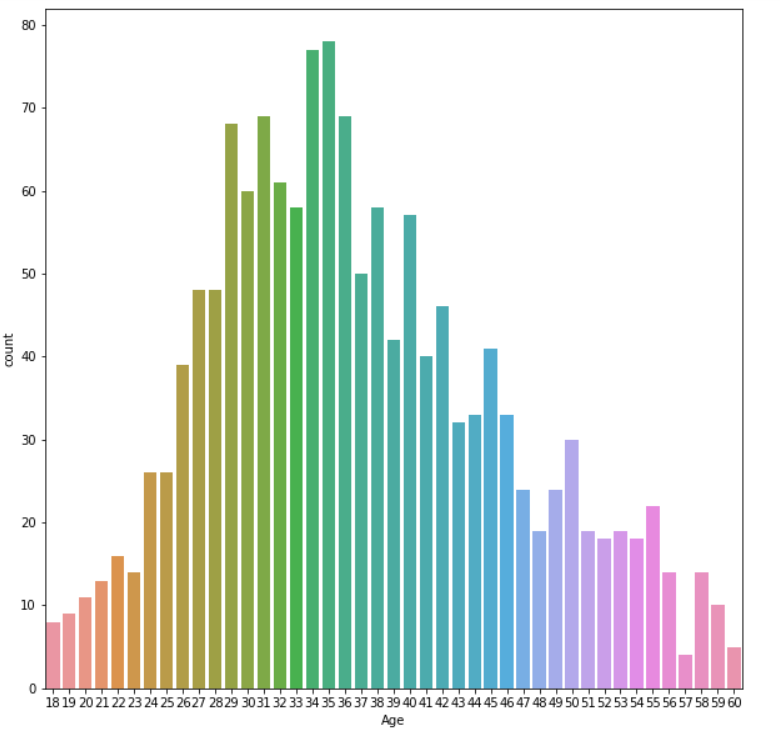


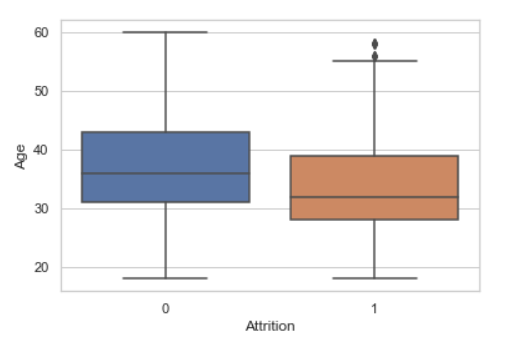
In this project, we are building a machine learning model to predict the values of ‘Attrition’, i.e. whether an employee will leave the company or not.

In this dataset, 1233 of the 1470 employees are still working in the company and 237 employees have left. This plot tells us that the dataset is imbalanced. Hence, we will have to make sure we split the dataset properly.

**2.4.2) Age**

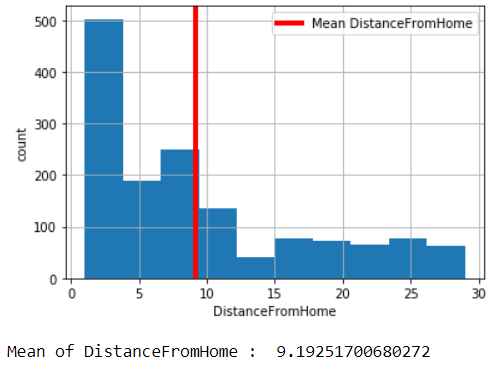
The Age column has a normal distribution. Most of the employees in the company are in their 30’s. The mean age of the employees that leave the company is 33, and the mean age of the employees that stay with the company is 37.

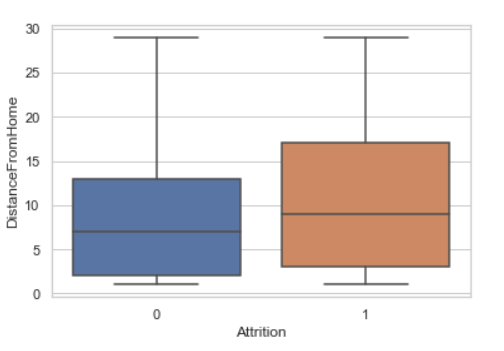




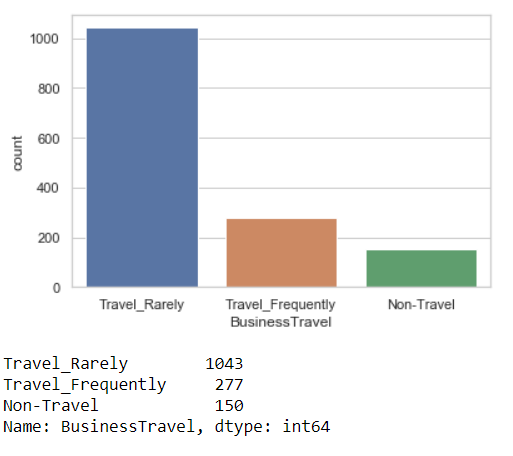
**2.4.3) DistanceFromHome**

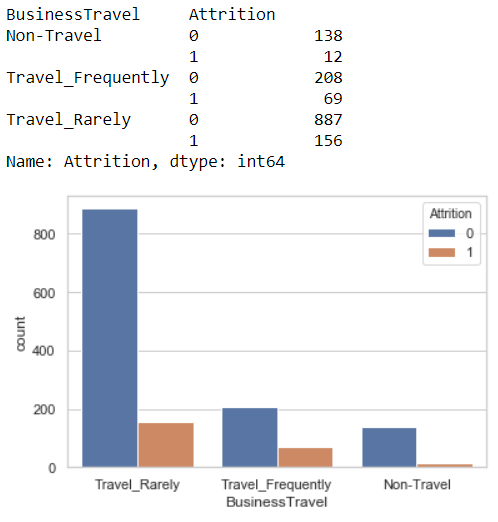
DistanceFromHome is heavily right skewed. The mean of distance from home to the company for the employees is 9.2. The average distance from home to the company for employees who are currently working in the company is 7 and for the employees who have left the company is 9. We will have to treat the column for skewness before building the model.





**2.4.4)** **BusinessTravel**

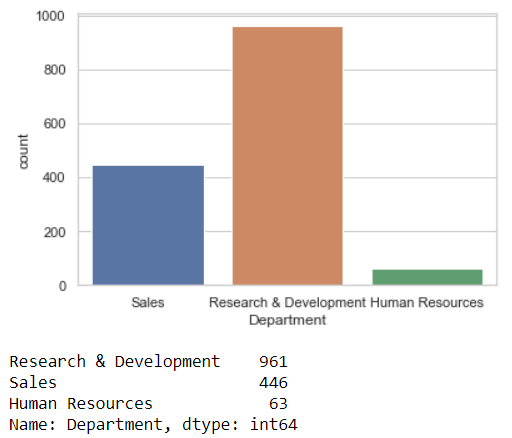


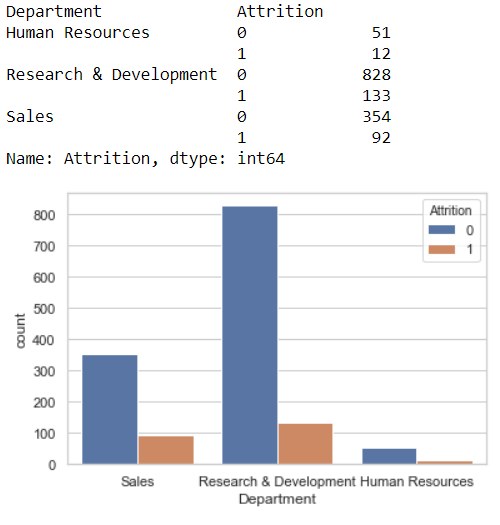


From 1043 employees that rarely travel for business, 887 employees are still working in the company and 156 have left the company. Out of 227 that travel frequently, 69 have left and 208 are still with the company. And among the employees who do not travel at all 138 are still in the company and 12 have left.

**2.4.5)** **Department**

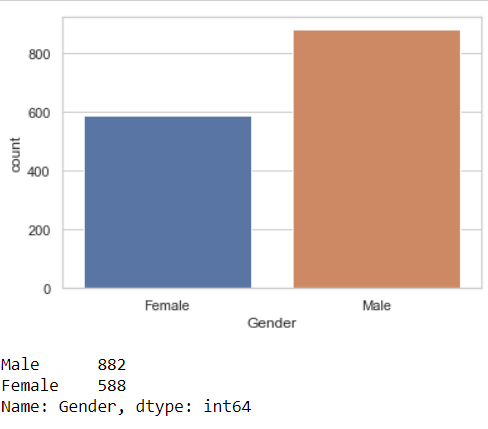
13.84% of the employees working in the Research & Development department have left the company while 20. 62% of the employees from Sales department have also left. From Human Resources, 19.05% of the employees have left the company.

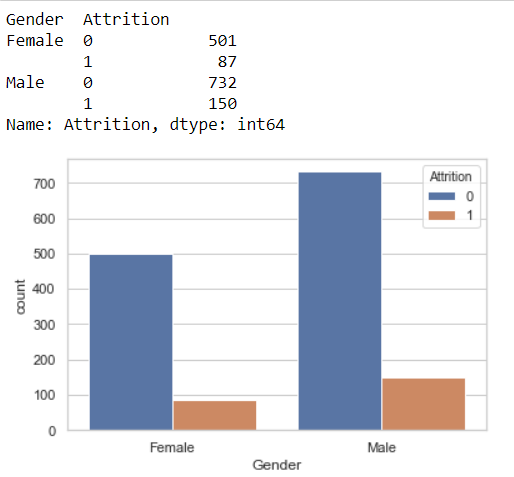




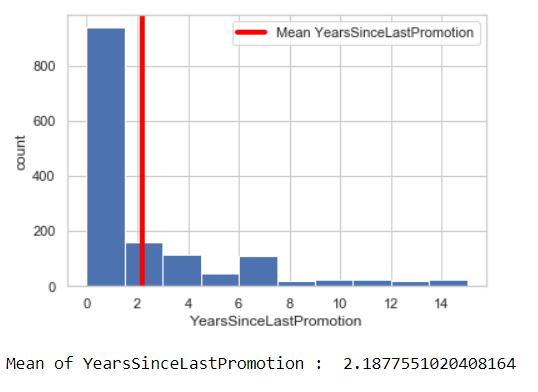
**2.4.6) Gender**

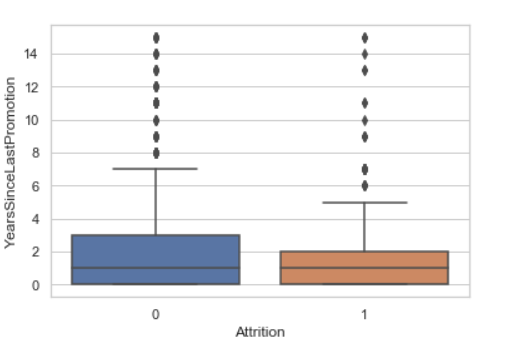
We can see that the company has a higher number of male employees than female. We can also see that the number of male employees leaving the company is higher than the female employees leaving the company.





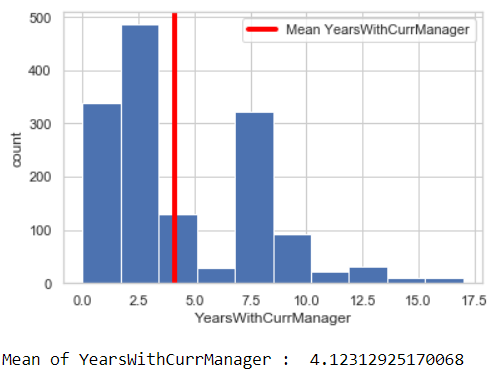
**2.4.7) YearsSinceLastPromotion**

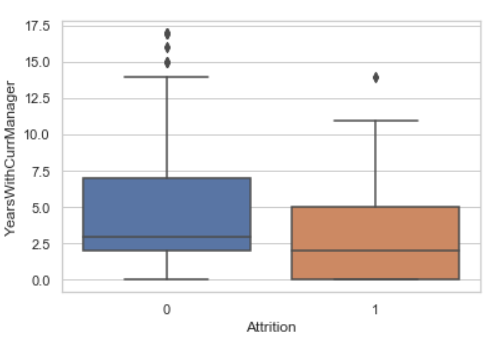




YearsSinceLastPromotion column is highly right skewed and it has a lot of outliers. The average years before an employee gets a promotion is roughly 2. Both the employees who are still working in the company and those who have left have an average of 1 year since their last promotion. We will have to treat the data for skewness and outliers before training the model.

**2.4.8) YearsWithCurrManager**

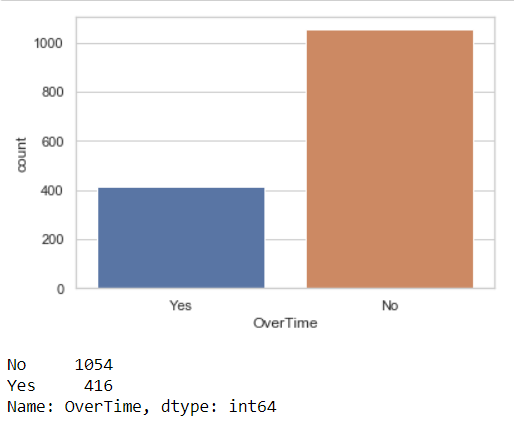


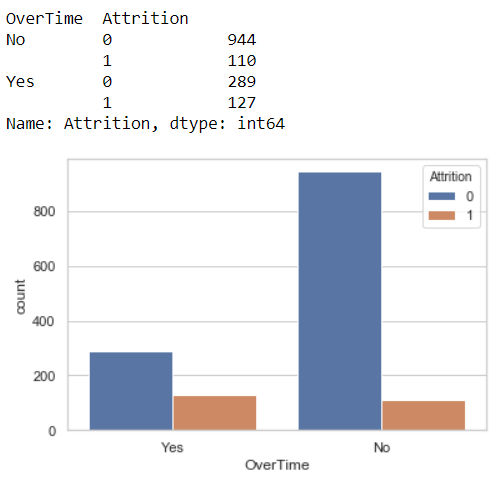


YearsWithCurrManager column is right skewed and it has a few outliers. The average years an employee of the company works under a particular manager is roughly 4. Employees that leave the company spend an average of almost 2 years with the same manager. On the other hand employees that stay with the company spend an average of roughly 3 years with the same manager.

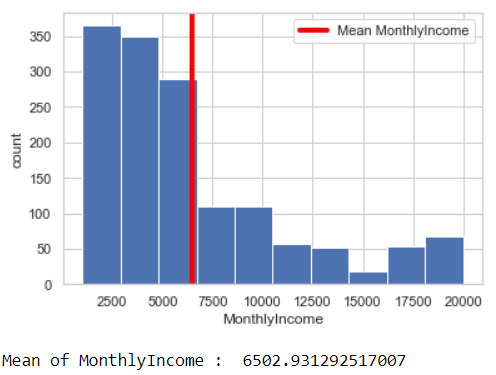
**2.4.9) OverTime**

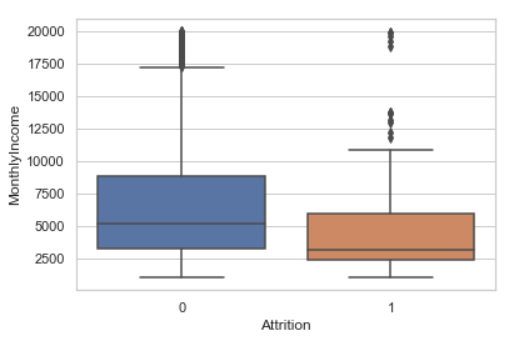
**As we can see, only about 28% of the employees opt for overtime, and more than 30% of those employees leave the company. On the other hand, around 10% of the employees who did not opt for any overtime left the company. Hence, we can say that a significantly larger number of employees who have opted for overtime are likely to leave the company as compared to those that have not.**





**2.4.10) MonthlyIncome**

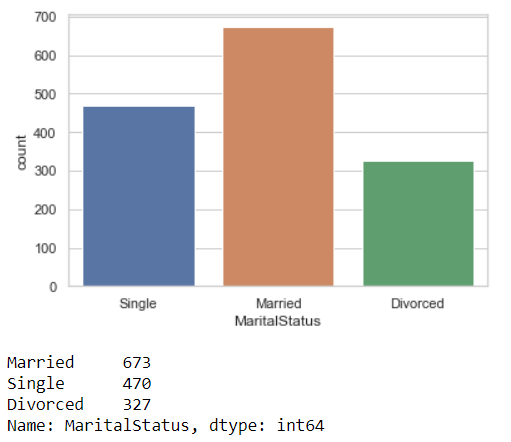


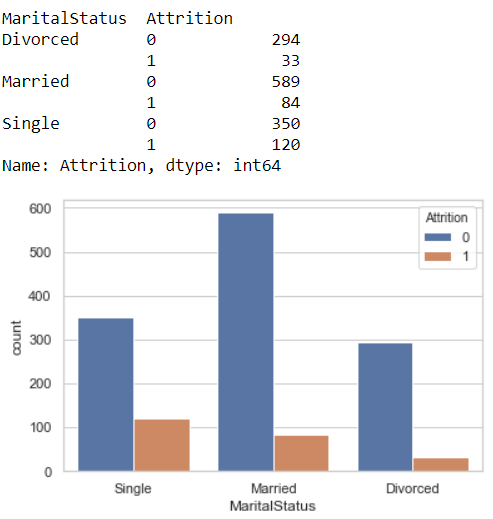


The MonthlyIncome column is highly right skewed and it has outliers. On an average the company pays its employees a monthly income of 6502.93. From the boxplot we can see that employees that leave the company have an average monthly income of around 3000 and employees who stay with the company have an average monthly income more than 5000. Hence, we can say that employees that are getting paid less are more likely to leave the company.

**2.4.11) MaritalStatus**

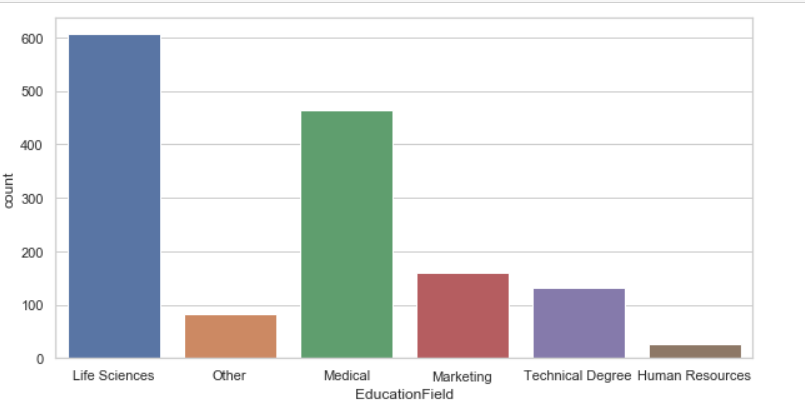
**More than 45% of the employees are married, of those only 12.5% employees leave the company. On the other hand, more than 25% of the employees that are not in any committed relationship and over 10% of the ones who are divorced leave the company. From this we can conclude that employees who are single are more likely to leave the company.**

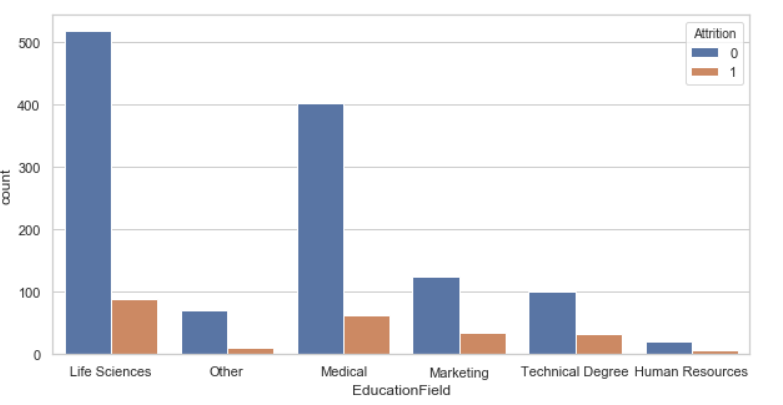




**2.4.12) EducationField**

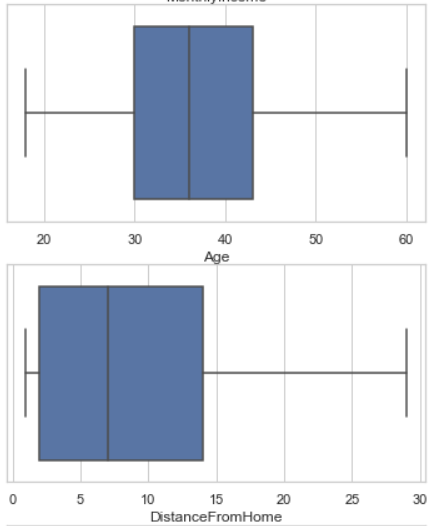
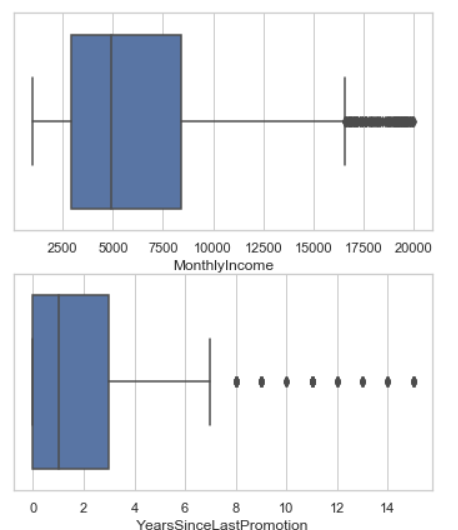
The employees listed in the dataset have various education backgrounds. Majority of the employees have studied Life Sciences. From the above graphs we can see that employees of Life Science education background are more likely to stay with the company and employees with a technical degree are more likely to leave the company.

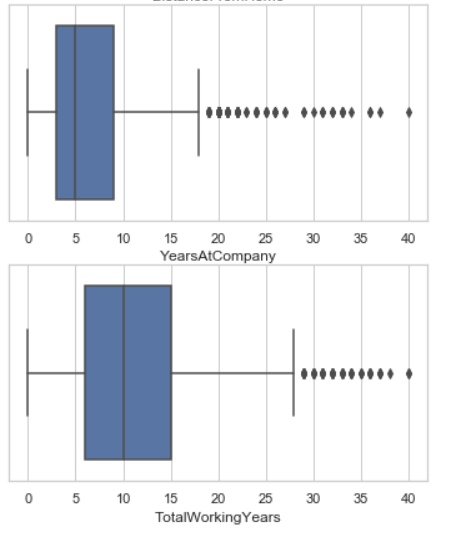
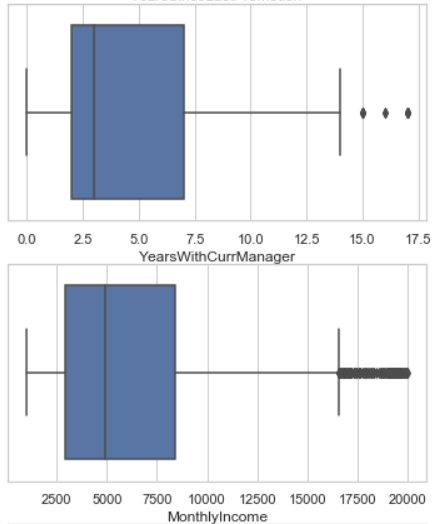




**2.5) Outliers**

We use boxplots to plot the outliers in the numerical columns of the dataset. From the below plots we can see that ‘MonthlyIncome’, ‘YearsSinceLastPromotion’, ‘YearsAtCompany’, ‘TotalWorkingYears’, and ‘YearsWithCurrManager’ columns have outliers. We will have to treat these outliers before building the predictive model.

# EDA Concluding Remarks

* The dataset has features of int64 and object datatype.
* The dataset does not have any missing values.
* ‘DistanceFromHome’ has the strongest positive correlation with the target variable ‘Attrition’.
* ‘JobLevel’ and ‘TotalWorkingYears’ have the strongest negative correlation with the target variable ‘At
* trition’.
* The dataset is heavily imbalanced, with majority of the employees still working with the company and very few who have left.
* Majority of the employees are in their 30’s.
* Employees who live closer to the company are more likely to continue working with the company.
* Employees working in the sales or the human resource department are more likely to leave the company as compared to employees working in the research and development department.
* Majority of the employees working in the company are male. However, female employees are less likely to leave the company, while male employees are more likely to leave.
* Employees that spend more time with the same manager are less likely to leave the company.
* Very few employees opt for overtime, however of those who do more than 30% are likely to leave the company. Whereas, employees who do not opt for overtime are more likely to continue working with the company.
* Employees who get paid less are more likely to leave the company.
* Employees who are divorced are most likely to stay with the company, whereas employees who are single are most likely to leave.
* Employees with a technical background are most likely to leave the company.

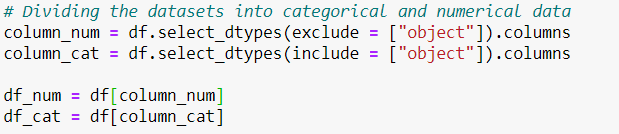
# Pre-processing Pipeline

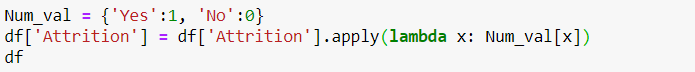
We have to pre-process the data before feeding it to a machine learning algorithm to build a model that will predict attrition with maximum accuracy.

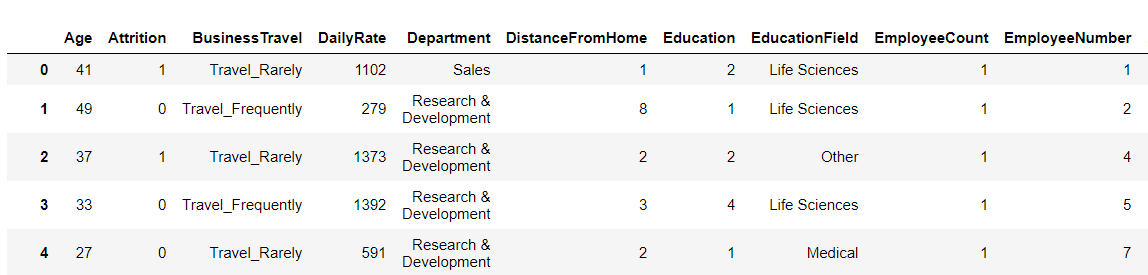
**4.1) Encoding**

Machine learning algorithms can only take numerical values as input. Therefore, we have to encode the categorical values to convert them into numerical values before feeding them to the algorithm. We use LabelEncoder to encode columns which have only two values, we represent these values as 0 and 1. To encode columns with more than 2 values we use get\_dummies() method of pandas library which creates a separate column for every value in the column and the values in the new columns are either 1 or 0.

First, we split the dataset into numerical and categorical data. Then, we replace ‘yes’ and ‘no’ in ‘Attrition’ column with 1 and 0 respectively.

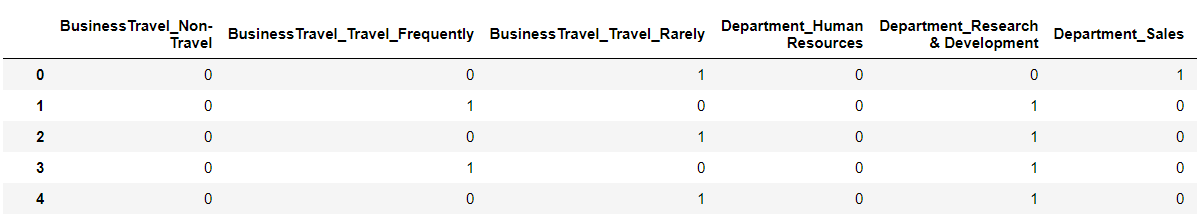






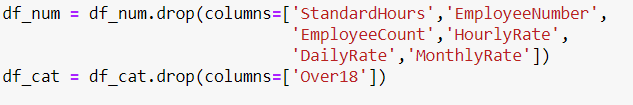
Next, we use get\_dummies() method to encoded the remaining categorical column of the dataset.





**4.2) Dropping Columns**

To make the model more accurate we drop columns that do not contribute to the output or the ones which have the same value over the entire dataset.

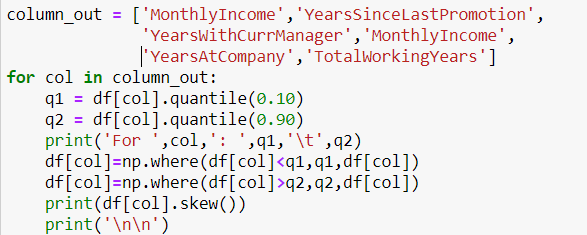


After dropping the columns, we concatenate the numerical and encoded categorical data.



**4.3) Treating Outliers**

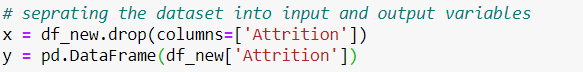
Since the dataset has lots of outliers, instead of dropping the rows with outliers, we replace the outliers with the quantile values.





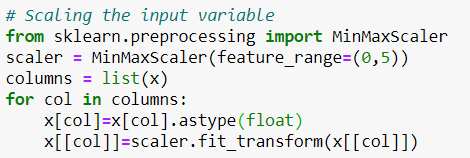
**4.4) Splitting the Dataset into Input and Output Variables**

Before scaling the dataset we have to split the dataset into input and output variables. We only scale the input variables.



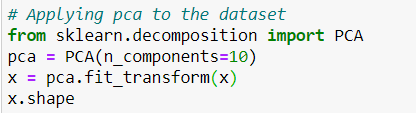
**4.5) Feature Scaling**

Before applying any machine learning algorithm on the dataset, we need to scale the data. Here, we use MinMaxScalar to scale the input parameters. MinMaxScalar transforms the features by scaling each feature to a given range. The estimator scales and translates each feature individually such that it is in the given range. For this dataset, we scale the features in the range 0 to 5.



**4.6) PCA**

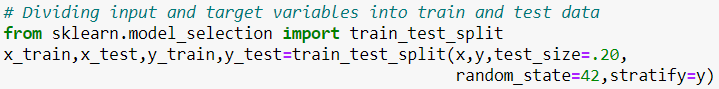
Since the number of columns in the dataset is high, we reduce the number by applying Principal Component Analysis (PCA). PCA reduces linear dimensionality using Singular Value Decomposition (SVD) on the data and projects it to a lower dimensional space. The input data is not scaled, but centered for each feature before applying SVD. Here, we reduce the number of columns from 48 to 10.





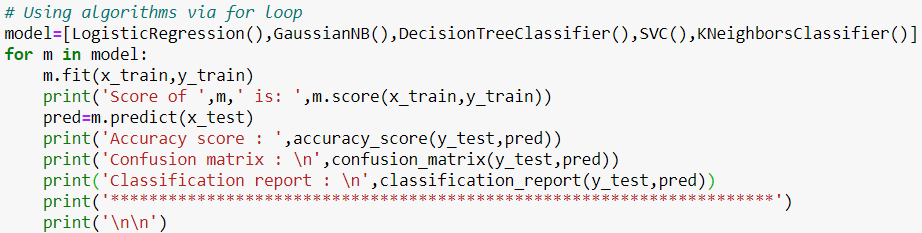
**4.7) Splitting the Dataset into Test and Train**

Before building the model we split the dataset into test and train data. We split it in the ratio 80:20. We use the train data to train the machine learning model which will then predict the results for the test data. The test data will only be used to make predictions and to evaluate the model performance in order to select the best model. We use stratify on y to make sure the y values are evenly split.



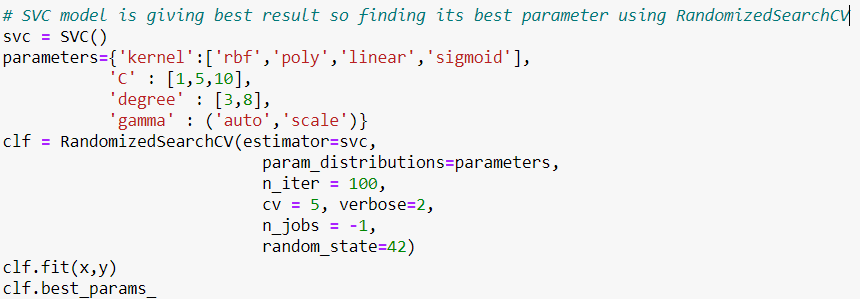
# Building Machine Learning Model

First, we run the baseline algorithms before moving on to hyper parameter tuning for algorithm that give the best accuracy scores. The algorithms we run for this problem are: LogisticRegression, GaussianNB, DecisionTreeClassifier, KNeighborsClassifier and SVC. We use a ‘for’ loop to run these algorithms. We print out the accuracy, confusion matrix and classification report for each model. LogisticRegression gives an accuracy of 0.86, GaussianNB gives and accuracy of 0.85, DecisionTreeClassifier gives an accuracy of 0.76, SVC gives an accuracy of 0.86 and KNeighborsClassifier gives an accuracy of 0.86.

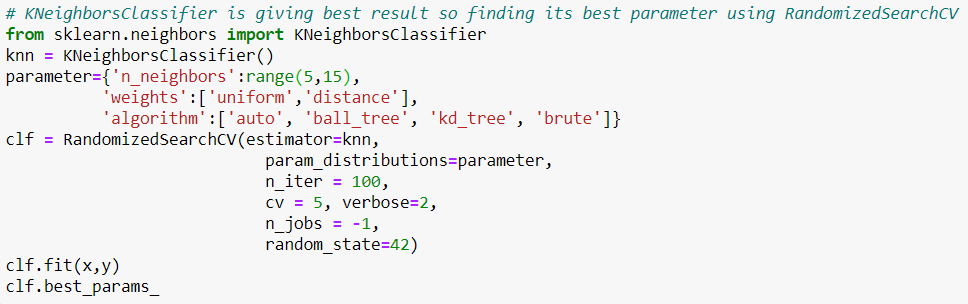


**5.1) Hyperparameter Tuning**

After getting the scores for baseline algorithms, we perform hyperparameter tuning on model which give the best scores to improve the accuracy and reduce overfitting or underfitting. We use RandomizedSearchCV for hyperparameter tuning for the models. RandomizedSearchCV is used as we consider more than one parameter for tuning and it gives results faster than GridSearchCV.

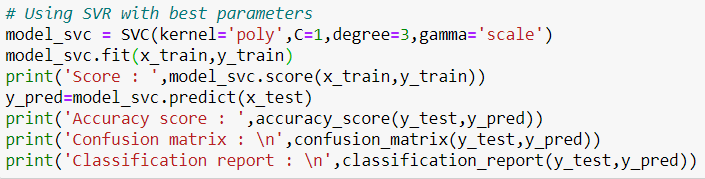


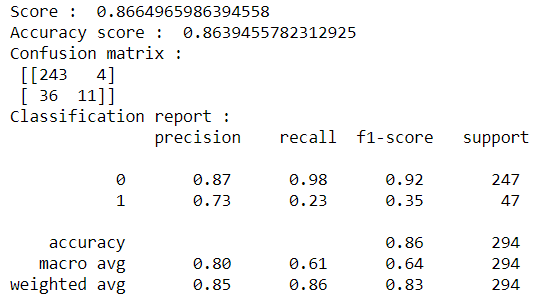


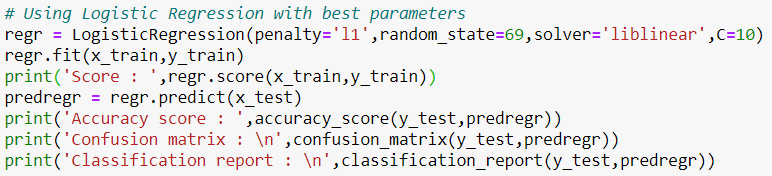


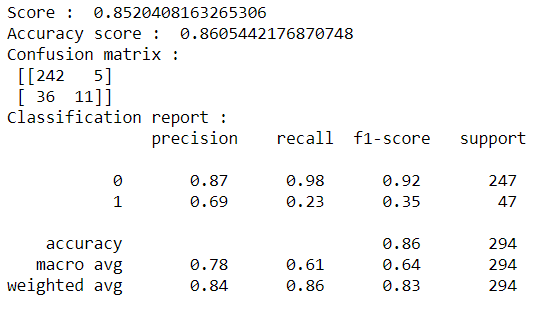
**5.2) Using the Models with the Best Parameters**

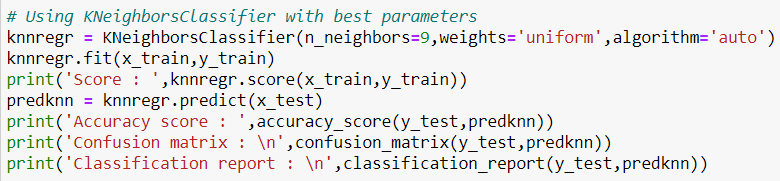
On applying RandomizedSearchCV we get the best parameters for the models. We then use the model with these parameters.

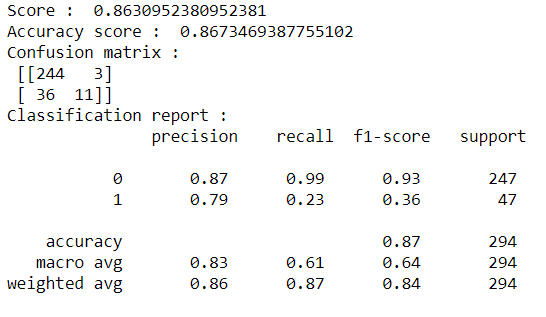






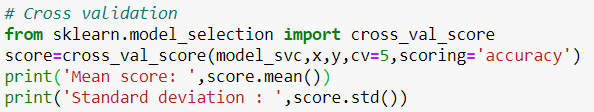




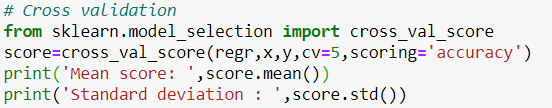


**5.3) Cross validation**

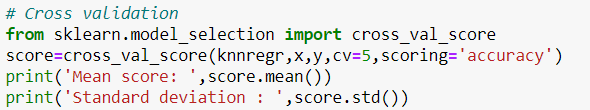
After building the model we cross validate the model to figure out whether the model is overfitting or underfitting. We use the cross\_val\_score() method from the model\_selection library to cross validate the model.









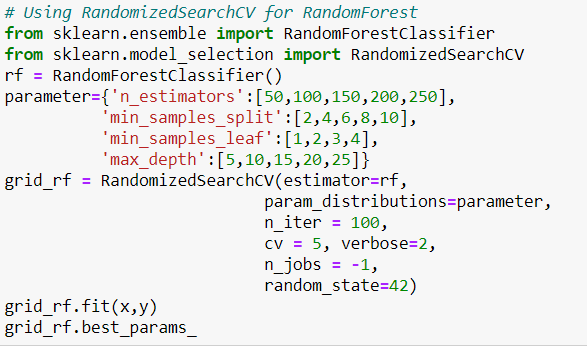




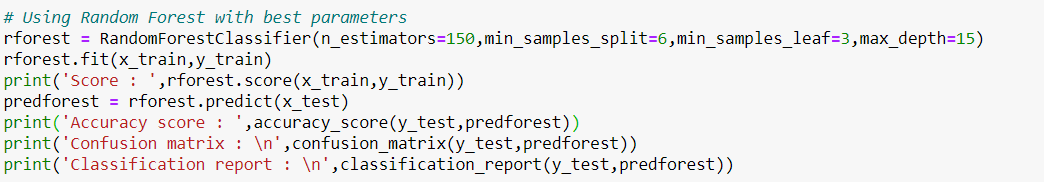
On cross validation we see that even though SVC and KNeighborsClassifier give better accuracy scores, LogisticRegression model gives the least difference between the accuracy score and the cross validation score. Therefore, among all the models LogisticRegression is the best fit for the dataset.

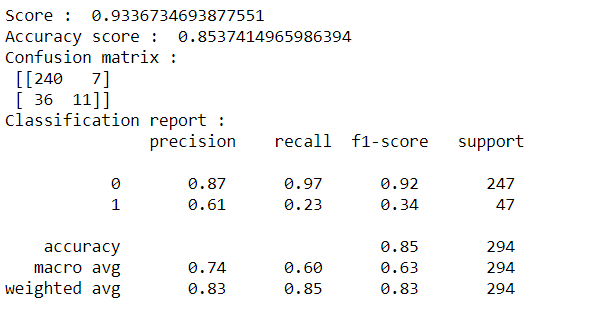
**5.4) RandomForestClassifier**

We use the RandomForestClassifier model on the dataset to try for better scores. We use hyperparameter tuning before running the model to get the best parameters.

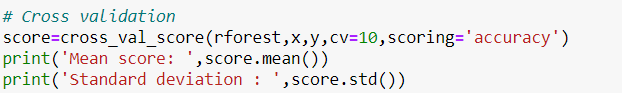


On getting the best parameters, we use the model with these parameters.





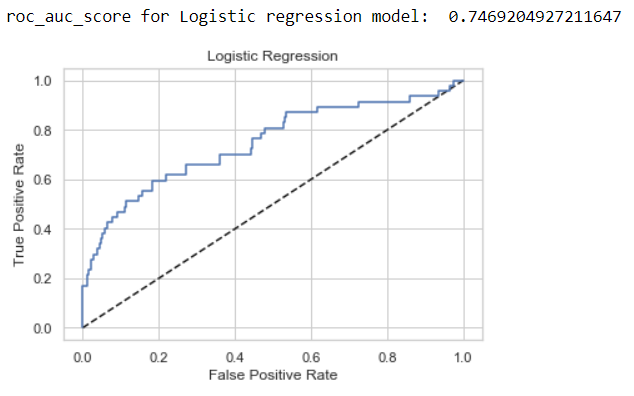
We then cross validate the model.





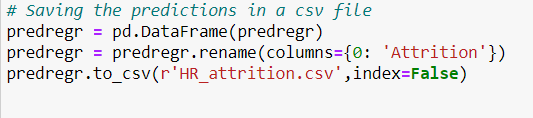
**5.5) ROC Curve**

ROC (Receiver Operating Characteristic) curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate and False Positive Rate. The area under this curve (AUC) is called ROC AUC Score. The higher the AUC score, the better is the performance of the model at distinguishing between positive and negative classes. We plot the ROC curve for LogisticRegression model and find the roc auc score.

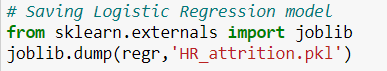


**5.6) Saving**

After testing various models on the dataset, we find that the LogisticRegression model gives the best fit for this dataset. Therefore, we select LogisticRegression as our final model and save the predictions made by that model is a csv file.



After saving the predictions we save the model as a pickle file using the joblib method of externals library.



# Concluding Remarks

# As new entries are added to the dataset and the dataset grows, we can retrain the model with the new data to get more accurate predictions. On getting better predictions we can move on to further classify the employees who are most likely to leave the company, those who may or may not leave the company and those who are least likely to leave the company. This data can also be used while hiring new employees to reduce the attrition rate.

# The most relevant features that indicate people leaving are:

# MonthlyIncome: Employees with lower wages are more likely to leave the company. Therefore, the company can provide regular bonuses and incentives for those employees.

# DistanceFromHome: Employees who live closer to the workplace are less likely to leave the company as compared to those who live farther away. So, the company should provide proper communal transport for the employees who live away from the workplace to make traveling easier and more convenient.

# Age: Employees below 30 years of age are more likely to leave the company. Screening the employees in their age will be considered as discrimination, so the company can work toward making the workplace friendlier towards that age group instead of turning them away.

# OverTime: Employees who work overtime are more likely to leave the company. Hence, the company can work toward having proper working hours for all the employees and sufficient workforce to reduce overtime as much as possible.

# YearsWithCurrManager: Employees that have worked under the current manager for less than 2 years are most likely to leave the company. To avoid that, the company should try not to change managers very frequently.

# In addition to the above listed suggestions, a through interview with the employee before hiring and knowledge of previous jobs will help in reducing the attrition rate.