## **1. Problem Definition:**

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

#### **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 in HR:**

#### Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

#### **Attrition affecting Companies:**

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. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

**2. Data Analysis:**

The dataset consists of basic information about the employees working in the company like Age, Education, Working years, etc. The dataset consists of 1470 employees' data & 34 columns of information about those employees. The features Description are mentioned below.

| **Name** | **Description** |
| --- | --- |
| AGE | Numerical Value |
| ATTRITION | Employee leaving the company (0=no, 1=yes) |
| BUSINESS TRAVEL | (1=No Travel, 2=Travel Frequently, 3=Tavel Rarely) |
| DAILY RATE | Numerical Value - Salary Level |
| DEPARTMENT | (1=HR, 2=R&D, 3=Sales) |
| DISTANCE FROM HOME | Numerical Value - THE DISTANCE FROM WORK TO HOME |
| EDUCATION | Numerical Value |
| EDUCATION FIELD | (1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TECHNICAL) |
| EMPLOYEE COUNT | Numerical Value |
| EMPLOYEE NUMBER | Numerical Value - EMPLOYEE ID |
| ENVIRONMENT SATISFACTION | Numerical Value - SATISFACTION WITH THE ENVIRONMENT |
| GENDER | (1=FEMALE, 2=MALE) |
| HOURLY RATE | Numerical Value - HOURLY SALARY |
| JOB INVOLVEMENT | Numerical Value - JOB INVOLVEMENT |
| JOB LEVEL | Numerical Value - LEVEL OF JOB |
| JOB ROLE | (1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= RESEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIVE, 9= SALES REPRESENTATIVE) |
| JOB SATISFACTION | Numerical Value - SATISFACTION WITH THE JOB |
| MARITAL STATUS | (1=DIVORCED, 2=MARRIED, 3=SINGLE) |
| MONTHLY INCOME | Numerical Value - MONTHLY SALARY |
| MONTHLY RATE | Numerical Value - MONTHLY RATE |
| NUMCOMPANIES WORKED | Numerical Value - NO. OF COMPANIES WORKED AT |
| OVER 18 | (1=YES, 2=NO) |
| OVERTIME | (1=NO, 2=YES) |
| PERCENT SALARY HIKE | Numerical Value - PERCENTAGE INCREASE IN SALARY |
| PERFORMANCE RATING | Numerical Value - PERFORMANCE RATING |
| RELATIONS SATISFACTION | Numerical Value - RELATIONS SATISFACTION |
| STANDARD HOURS | Numerical Value - STANDARD HOURS |
| STOCK OPTIONS LEVEL | Numerical Value - STOCK OPTIONS |
| TOTAL WORKING YEARS | Numerical Value - TOTAL YEARS WORKED |
| TRAINING TIMES LAST YEAR | Numerical Value - HOURS SPENT TRAINING |
| WORK-LIFE BALANCE | Numerical Value - TIME SPENT BETWEEN WORK AND OUTSIDE |
| YEARS AT COMPANY | Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPANY |
| YEARS IN CURRENT ROLE | Numerical Value -YEARS IN CURRENT ROLE |
| YEARS SINCE LAST PROMOTION | Numerical Value - LAST PROMOTION |
| YEARS WITH CURRENT MANAGER | Numerical Value - YEARS SPENT WITH CURRENT MANAGER |

**Data type of each feature:**

The data type of all the column need to be checked and try to convert those object data type to integer or float. So, our model will able to predict the target variable.

* Age : int64
* Attrition : object
* BusinessTravel : object
* DailyRate : int64
* Department : object
* DistanceFromHome : int64
* Education : int64
* EducationField : object
* EmployeeCount : int64
* EmployeeNumber : int64
* EnvironmentSatisfaction : int64
* Gender : object
* HourlyRate : int64
* JobInvolvement : int64
* JobLevel : int64
* JobRole : object
* JobSatisfaction : int64
* MaritalStatus : object
* MonthlyIncome : int64
* MonthlyRate : int64
* NumCompaniesWorked : int64
* Over18 : object
* OverTime : object
* PercentSalaryHike : int64
* PerformanceRating : int64
* RelationshipSatisfaction : int64
* StandardHours : int64
* StockOptionLevel : int64
* TotalWorkingYears : int64
* TrainingTimesLastYear : int64
* WorkLifeBalance : int64
* YearsAtCompany : int64
* YearsInCurrentRole : int64
* YearsSinceLastPromotion : int64

Some columns need to be encoded for building the model.

**Missing values:**

Checking if there's any missing values in the dataset if there are any missing values. It can be rectified through the imputation technique.

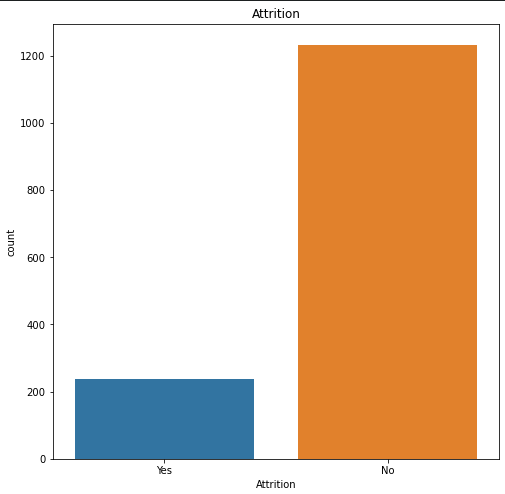
* Age 0
* Attrition 0
* BusinessTravel 0
* DailyRate 0
* Department 0
* DistanceFromHome 0
* Education 0
* EducationField 0
* EmployeeCount 0
* EmployeeNumber 0
* EnvironmentSatisfaction 0
* Gender 0
* HourlyRate 0
* JobInvolvement 0
* JobLevel 0
* JobRole 0
* JobSatisfaction 0
* MaritalStatus 0
* MonthlyIncome 0
* MonthlyRate 0
* NumCompaniesWorked 0
* Over18 0
* OverTime 0
* PercentSalaryHike 0
* PerformanceRating 0
* RelationshipSatisfaction 0
* StandardHours 0
* StockOptionLevel 0
* TotalWorkingYears 0
* TrainingTimesLastYear 0
* WorkLifeBalance 0
* YearsAtCompany 0
* YearsInCurrentRole 0
* YearsSinceLastPromotion 0
* YearsWithCurrManager 0

There are no missing values in the dataset with 1470 observations and 34 columns. Let's move on with our analysis process.

**Dependent variable (Attrition):**

Checking if there's an imbalance in the target variable.The target variable "Attrition" has,

* No - 1233
* Yes- 237



The dataset is imbalanced.So, if we build the model based on this observatins. our model will be biased. Thus we need to make the dataset balanced.

**3. Exploratory Data Analysis:**

**Descriptive Statistics:**

We are going to see the mean, standard deviation, and the 5 point summary of the variables in the dataset to understand the pattern.

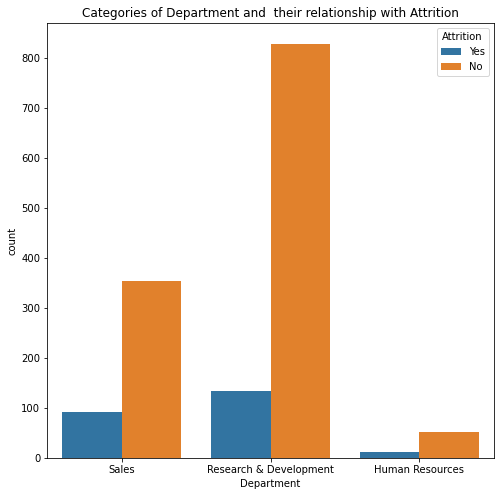
|  | **min** | **25 percent** | **median** | **mean** | **75 percent** | **max** | **std** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 18 | 28.00 | 32.0 | 34.11 | 40.00 | 58 | 9.98 |
| BusinessTravel | 1 | 2.00 | 3.0 | 2.61 | 3.00 | 3 | 0.59 |
| EducationField | 1 | 2.00 | 3.0 | 3.33 | 4.00 | 6 | 1.41 |
| EnvironmentSatisfaction | 1 | 1.00 | 3.0 | 2.45 | 3.75 | 4 | 1.17 |
| Gender | 1 | 1.00 | 2.0 | 1.61 | 2.00 | 2 | 0.49 |
| HourlyRate | 31 | 50.00 | 65.5 | 65.29 | 83.75 | 100 | 20.38 |
| JobInvolvement | 1 | 2.00 | 3.0 | 2.50 | 3.00 | 4 | 0.77 |
| JobLevel | 1 | 1.00 | 1.0 | 1.63 | 2.00 | 5 | 0.95 |
| JobRole | 1 | 3.00 | 7.0 | 5.71 | 8.00 | 9 | 2.63 |
| JobSatisfaction | 1 | 1.00 | 3.0 | 2.48 | 3.00 | 4 | 1.10 |
| MaritalStatus | 1 | 2.00 | 2.5 | 2.35 | 3.00 | 3 | 0.72 |
| MonthlyIncome | 1081 | 2363.00 | 3090.5 | 4805.40 | 6098.75 | 19859 | 3681.86 |
| MonthlyRate | 2326 | 8885.25 | 14465.0 | 14343.95 | 20700.75 | 26956 | 7067.61 |
| NumCompaniesWorked | 0 | 1.00 | 1.0 | 2.96 | 5.00 | 9 | 2.70 |
| OverTime | 1 | 1.00 | 2.0 | 1.55 | 2.00 | 2 | 0.50 |
| PercentSalaryHike | 11 | 12.00 | 14.0 | 14.82 | 17.00 | 25 | 3.60 |
| PerformanceRating | 1 | 1.00 | 1.0 | 1.12 | 1.00 | 2 | 0.32 |
| RelationshipSatisfaction | 1 | 1.00 | 3.0 | 2.56 | 4.00 | 4 | 1.14 |
| StockOptionLevel | 1 | 1.00 | 1.0 | 1.54 | 2.00 | 4 | 0.86 |
| TotalWorkingYears | 0 | 3.00 | 7.0 | 8.28 | 10.00 | 40 | 7.51 |
| TrainingTimesLastYear | 0 | 2.00 | 2.5 | 2.62 | 3.00 | 6 | 1.25 |
| WorkLifeBalance | 1 | 2.00 | 3.0 | 2.68 | 3.00 | 4 | 0.80 |
| YearsAtCompany | 0 | 1.00 | 3.0 | 5.08 | 7.00 | 40 | 6.19 |
| YearsInCurrentRole | 0 | 0.00 | 2.0 | 2.90 | 4.00 | 15 | 3.14 |
| YearsSinceLastPromotion | 0 | 0.00 | 1.0 | 1.93 | 2.00 | 15 | 3.03 |
| YearsWithCurrManager | 0 | 0.00 | 2.0 | 2.83 | 5.00 | 14 | 3.16 |

* Employee Number is just for reference purpose it does not give any insights whether the employee has attrition or not.
* Employee count has value 1 for all the Five Number Summary & Standarhours also have 80 in all the five-number summary. Since they have the same value for all the observations it doesn't give any useful value for predicting the attrition.
* Over18 has the same class 'y' which is also not useful for prediction.

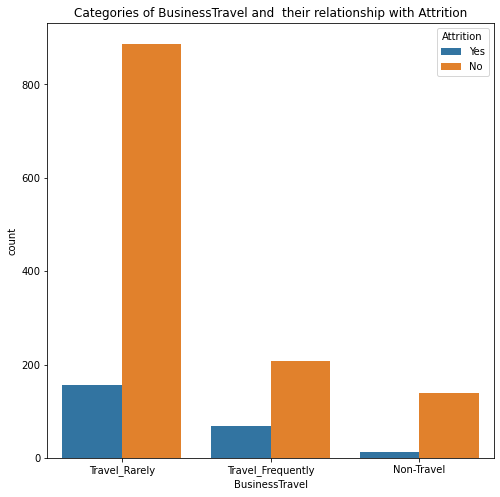
Thus Removing these columns from the dataset.

**Data visualization:**

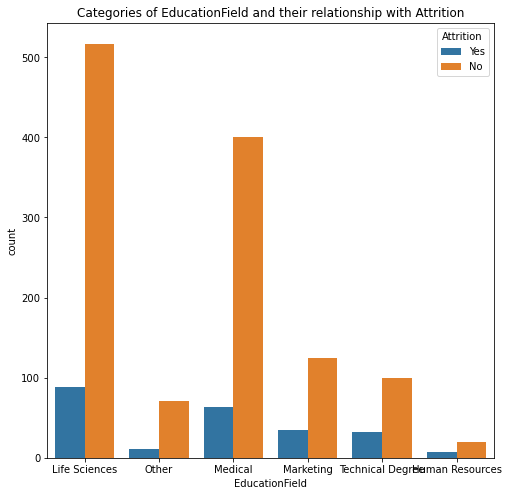
Let us understand the pattern we got in the dataset through visualization and get some insights.



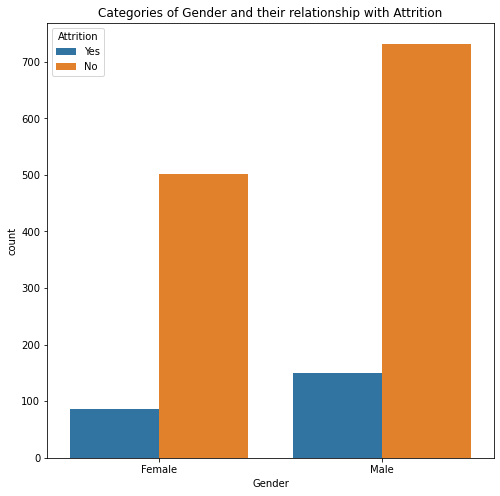
* The R&D department has the most attrition whereas HR and sales have the least.



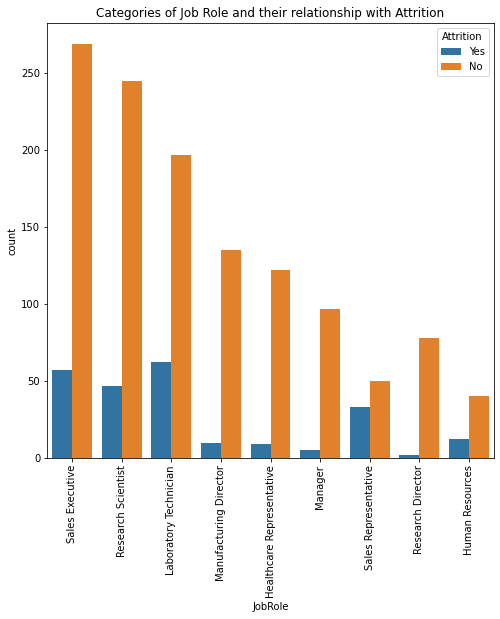
* Employees those travel rarely are the one who have the most attrition when compared with others.



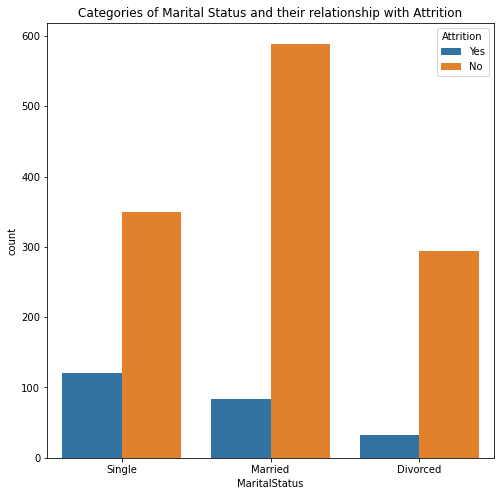
* Employees with education Life science and medical have the most attriton because of huge numbers.



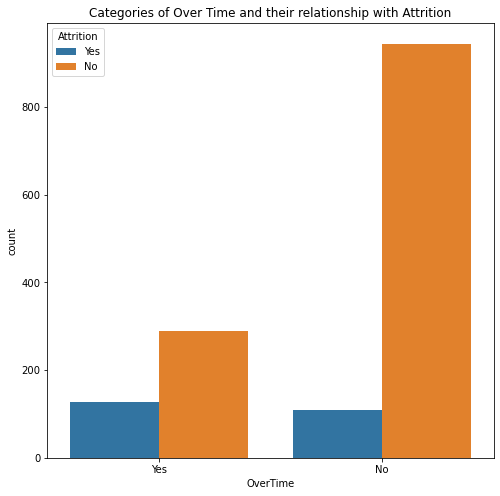
* Male employees have more attrition than female.



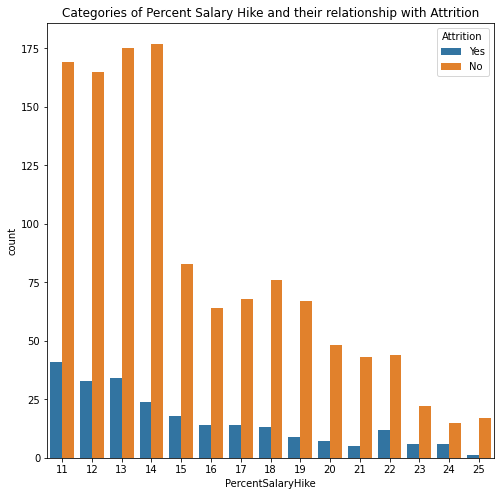
* Employees working in Laboratory Technician, Sales Executive, Research Scientist have more attrition than other job roles.



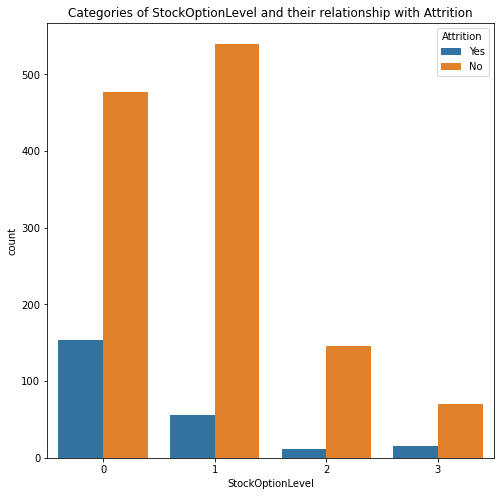
* Single employees have a higher chance to have attrition than married or divorced.



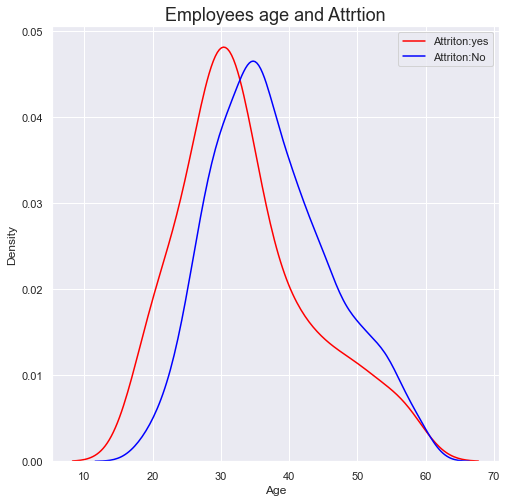
* People working overtime are mostly going to do leave the company.



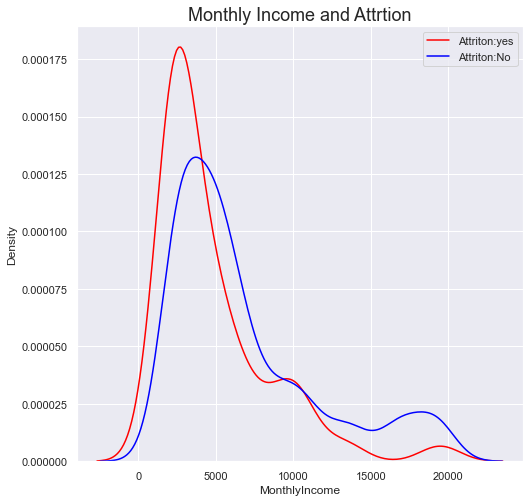
* The more the percentage salary hike less the employee leaving the company.



* The more the stock options level the employee has lesser the employee leaving the company.



* Employees of more age are least likely to leave the company than in the mid 20s.



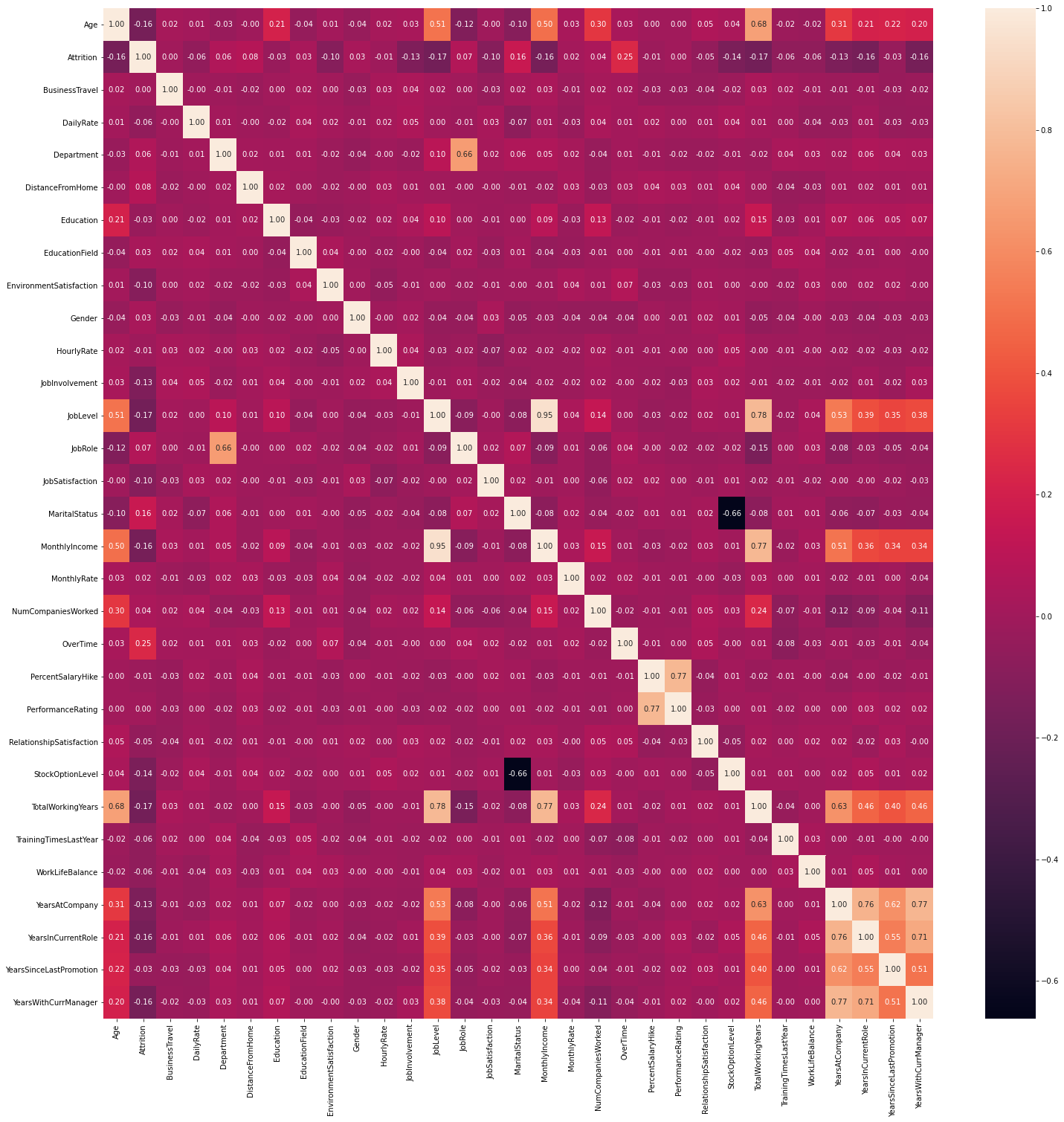
* Employees getting low monthly income are most likely to leave the job than others.

All the features have been represented in the form of a graph. We can also see the relationship with Attrition prediction.

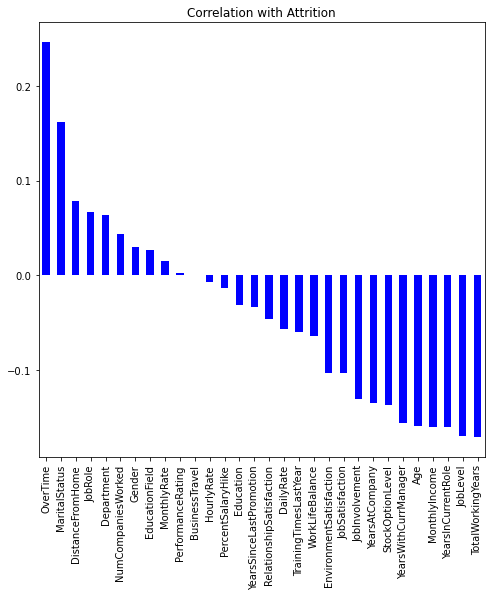
**4. Pre-processing:**

We need to check if there’s multicollinearity between the features. And also checking the correlation with the target variable to see which independent variable is positively correlated and negatively correlated. Before that, the target variable needs to be encoded by replacing “Yes” with 1 and “No” with 0 and creating dummies for the categorical features.

**Correlation between one column with others:**



* Some of the features are slightly multicollinear but it's not that much to worry about. We will leave it.



**Observation:**

* There are some multicollinearity between Age, MonthlyIncome, TotalWorkingYears,YearsAtCompany,YearsInCurrentRole,YearsSinceLastPromotion,YearsWithCurrManager.
* Attrition has a high positive correlation with OverTime,MaritalStatus & highly negatively correlated with TotalWorkingYears, YearsInCurrentRole,JobLevel,MonthlyIncome,Age etc
* Attrition has less correlation with BusinessTravel.

**Skewness:**

We need to check skewness in the variables that are numerical data types.

* Age 0.413286
* DailyRate -0.003519
* DistanceFromHome 0.958118
* Education - 0.289681
* EmployeeCount 0.000000
* EmployeeNumber 0.016574
* EnvironmentSatisfaction - 0.321654
* HourlyRate - 0.032311
* JobInvolvement - 0.498419
* JobLevel 1.025401
* JobSatisfaction - 0.329672
* MonthlyIncome 1.369817
* MonthlyRate 0.018578
* NumCompaniesWorked 1.026471
* PercentSalaryHike 0.821128
* PerformanceRating 1.921883
* RelationshipSatisfaction - 0.302828
* StandardHours 0.000000
* StockOptionLevel 0.968980
* TotalWorkingYears 1.117172
* TrainingTimesLastYear 0.553124
* WorkLifeBalance - 0.552480
* YearsAtCompany 1.764529
* YearsInCurrentRole 0.917363
* YearsSinceLastPromotion 1.984290
* YearsWithCurrManager 0.833451

There are some skewness in the dataset. But those variables are categorical data type so we don’t need to remove those skew. Let's remove the skewness for variables that are numerical data types. Removed skewness for numerical dataset variables is below.

* Age -0.007603
* DailyRate -0.196160
* DistanceFromHome -0.007468
* Education -0.289681
* EnvironmentSatisfaction -0.321654
* HourlyRate -0.106461
* JobInvolvement -0.498419
* JobLevel 1.025401
* JobSatisfaction -0.329672
* MonthlyIncome 0.000000
* MonthlyRate -0.184087
* NumCompaniesWorked 1.026471
* PercentSalaryHike 0.821128
* PerformanceRating 1.921883
* RelationshipSatisfaction -0.302828
* StockOptionLevel 0.968980
* TotalWorkingYears 1.117172
* TrainingTimesLastYear 0.553124
* WorkLifeBalance -0.552480
* YearsAtCompany 1.764529
* YearsInCurrentRole 0.917363
* YearsSinceLastPromotion 1.984290
* YearsWithCurrManager 0.833450

**Outliers:**

Outliers may lead to some kind of bias and make the model worse with less accuracy score etc. Thus we need to see if there are any outliers in the dataset.

There are no observations with Zscore of more than 3 standard deviations i.e There are no outliers in the features with numerical datatypes.

**5. Building Machine Learning Models:**

First, we need to split the dataset into dependent and independent variables for building the model. Keeping the independent columns in one variable and the attrition variable in other. Then select the best random state with the best score to train our model with so that each time it's picking the same observation for training and testing dataset. By running the score we are Selecting the best random state as 41 which gives an accuracy of about 89% in both test and train datasets using Logical regression.

Next, the dataset will be split into x\_train,x\_test,y\_train,y\_test with a test sample of 25% and a random state of 41.

Since the dataset is imbalanced which we have seen above. We need to rectify it by Synthetic Minority Oversampling Technique (SMOTE). Creating oversampling for the target variable YES with 75% of the No target variable. And selecting the best k-fold for cross-validation score which is 4.

We are going to predict the attrition by using 5 to 6 algorithms and select one of them which gives a higher accuracy score. Basically, we will train the model with the training dataset and the algorithm will find the pattern between those features and target. Then we will send the independent variable of the testing dataset and tell the model to predict the target variable of the testing dataset and we will compare it with the actual testing target variable to find the accuracy of the model.

1. **LogisticRegression:**

Accuracy 0.8858695652173914

----------------------------------------------------------------------------------------------------

Cross Validation Score 0.8829931972789116

----------------------------------------------------------------------------------------------------

Confusion Matrix

[[303 15]

[ 27 23]]

----------------------------------------------------------------------------------------------------

precision recall f1-score support

0 0.92 0.95 0.94 318

1 0.61 0.46 0.52 50

accuracy 0.89 368

macro avg 0.76 0.71 0.73 368

weighted avg 0.88 0.89 0.88 368

1. **Random Forest Classifier:**

Accuracy 0.8940217391304348

----------------------------------------------------------------------------------------------------

Cross Validation Score 0.8564625850340135

----------------------------------------------------------------------------------------------------

Confusion Matrix

[[314 4]

[ 35 15]]

----------------------------------------------------------------------------------------------------

precision recall f1-score support

0 0.90 0.99 0.94 318

1 0.79 0.30 0.43 50

accuracy 0.89 368

macro avg 0.84 0.64 0.69 368

weighted avg 0.88 0.89 0.87 368

1. **Decision tree classifier:**

Accuracy 0.7880434782608695

----------------------------------------------------------------------------------------------------

Cross Validation Score 0.7843537414965985

----------------------------------------------------------------------------------------------------

Confusion Matrix

[[266 52]

[ 26 24]]

----------------------------------------------------------------------------------------------------

precision recall f1-score support

0 0.91 0.84 0.87 318

1 0.32 0.48 0.38 50

accuracy 0.79 368

macro avg 0.61 0.66 0.63 368

weighted avg 0.83 0.79 0.81 368

1. **Gradient Boost Classifier:**

Accuracy 0.8831521739130435

----------------------------------------------------------------------------------------------------

Cross Validation Score 0.8721088435374149

----------------------------------------------------------------------------------------------------

Confusion Matrix

[[306 12]

[ 31 19]]

----------------------------------------------------------------------------------------------------

precision recall f1-score support

0 0.91 0.96 0.93 318

1 0.61 0.38 0.47 50

accuracy 0.88 368

macro avg 0.76 0.67 0.70 368

weighted avg 0.87 0.88 0.87 368

1. **Support vector classifier:**

Accuracy 0.8804347826086957

----------------------------------------------------------------------------------------------------

Cross Validation Score 0.8387755102040817

----------------------------------------------------------------------------------------------------

Confusion Matrix

[[303 15]

[ 29 21]]

----------------------------------------------------------------------------------------------------

precision recall f1-score support

0 0.91 0.95 0.93 318

1 0.58 0.42 0.49 50

accuracy 0.88 368

macro avg 0.75 0.69 0.71 368

weighted avg 0.87 0.88 0.87 368

1. **XGBClassifier:**

Accuracy 0.8831521739130435

----------------------------------------------------------------------------------------------------

Cross Validation Score 0.8666666666666666

----------------------------------------------------------------------------------------------------

Confusion Matrix

[[304 14]

[ 29 21]]

----------------------------------------------------------------------------------------------------

precision recall f1-score support

0 0.91 0.96 0.93 318

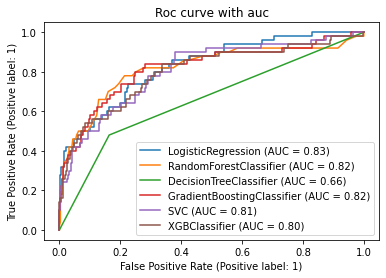
1 0.60 0.42 0.49 50

accuracy 0.88 368

macro avg 0.76 0.69 0.71 368

weighted avg 0.87 0.88 0.87 368

| **Model** | **Accuray Score in %** |
| --- | --- |
| **LogisticRegression** | **87** |
| **Random Forest Classifier** | **88** |
| **Decision tree classifier** | **78** |
| **Gradient Boost Classifier** | **88** |
| **Support vector classifier** | **87** |
| **XGBClassifier** | **88** |



**6. Conclusion:**

By considering the accuracy score and the area under the curve Gradient Boost Classifier gives the best in both of them with an Accuracy score of 88% and AUC of 82%.We will do hyperparameter tuning for the Gradient Boosting Classifier model. By doing the hyperparameter tuning with the Gridsearch library the best parameter is selected from different values. The best parameters are 'learning\_rate': 1, 'max\_depth': 9, 'min\_samples\_leaf': 5, 'min\_samples\_split': 4, 'n\_estimators': 50 which gives an improved accuracy score of 89% which is good improvement. The model can be saved for future prediction after some more testing the model can be deployed.