HR Attrition Prediction with Machine Learning Models

horizontal line

# 

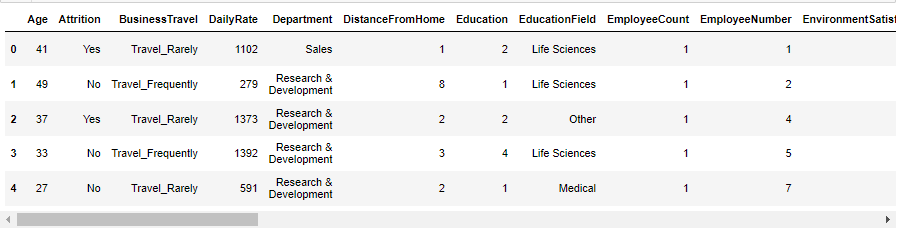
# Introduction

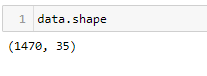
**Problem Definition:**

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. 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.

A lot of research work has been conducted on constructing models for predicting the outcome of Attrition. The accuracy in predicting the results of Attrition dependents greatly on the size of available datasets. Therefore, models built using machine learning methods are useful for predicting the outcomes (Yes/No) of employee Attrition. It is also very important to compare the differences between the models with respect to their performance. In this project, the dataset is utilized for building machine learning models that would predict the Attrition of employees. Based on the outcome of comparing the prediction accuracies of the models, the best one from among them will be finally picked and tuned further to improve its prediction accuracy.

**About the Dataset:**



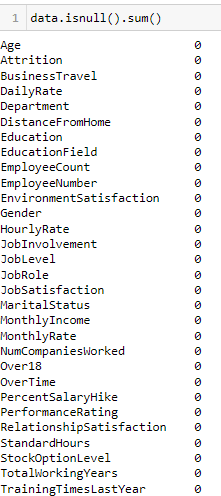


The given dataset consists of 35 columns and 1470 rows.

This dataset utilizes data in order to develop an algorithm that predicts the attrition for a given employee. There are 34 different features that will be used as the inputs to the machine learning and the output will be a value in term of 1 or 0 that represents the number Yes or No.

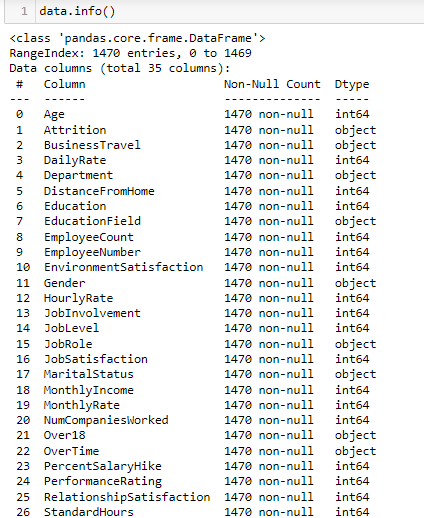
**Data Cleaning:**

Upon inspecting all the columns in the data frame, it is observed there are no null values / values missing from any of the columns in the data frame.

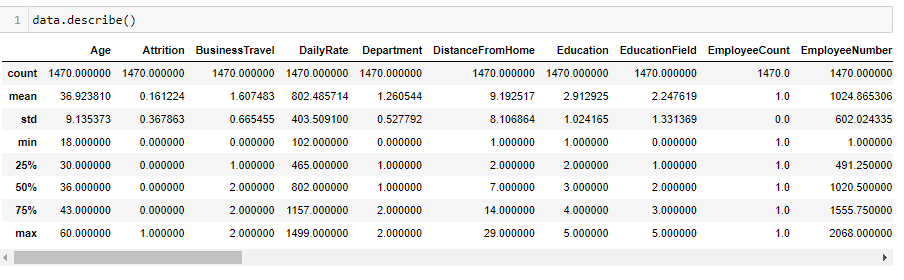


**Exploratory Data Analysis**

**Getting the basic summary and statistical information of the data.**



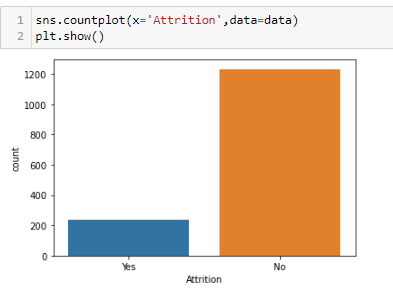
* No null value in datasets.
* Datatype is ok with respect to their column.



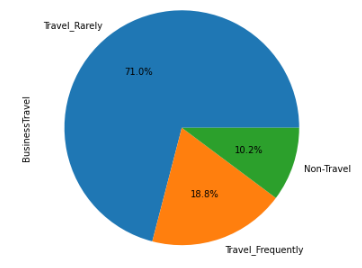
* Difference between 75% and max in Daily Rate, Monthly Income and Monthly Rate is high, columns is considerable indicating presence of outliers.
* Monthly Income Mean to Standard of deviation difference is close.

**This is a Classification Problem since the Target variable / Label column ("Attrition") has Categorical type of Data.**

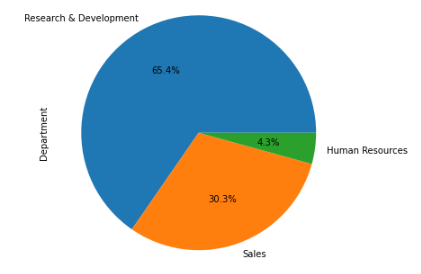
**Univariate Analysis:**



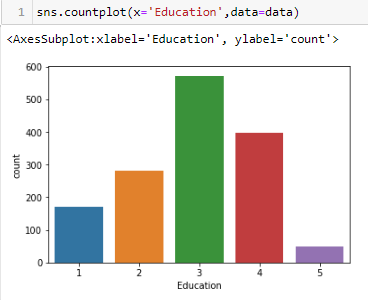
* We can see, there is low yes attrition, it means there is no problematic. If attrition is high then its problematic.
* There is less employees whose leave the job or their attrition is yes.



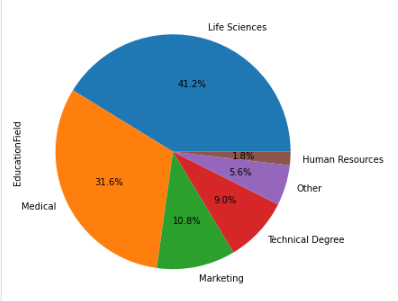
* We can see, there are 71% employees are Rarely doing business travel.
* 10.2% employees are not doing business travel.



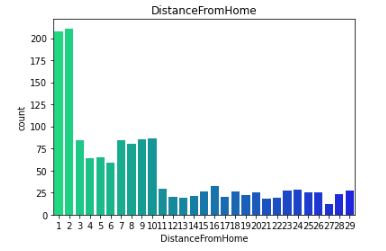
* We can see, R and D Department has 65.4% of manpower.
* HR has 4.3% of Employees are worked.



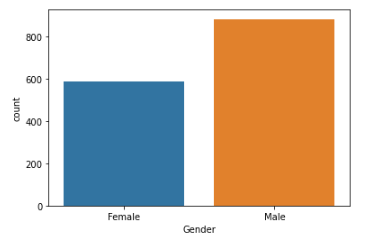
* Education 3 meant Bachelor degree
* Whose have Bachelor degree there are more followed by 4.
* Doctor degree has low employees.



* We can say that, 41.2% employees is having life Sciences as education field.
* 31.6% employees are having medical as education field.
* 1.8% employees are having HR education field.

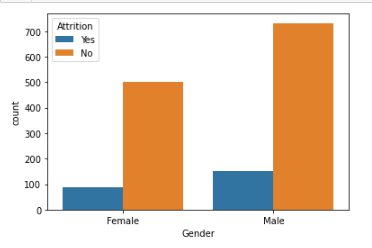


* As we can see, most of the employee may have come from 2km. Because it has higher in number followed by 1.
* Very less employee come from more distance 16km.

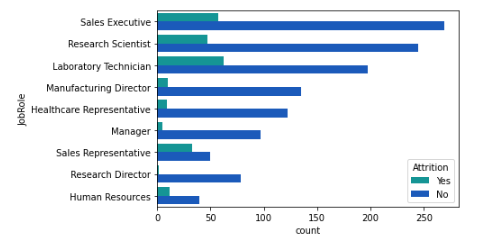


* We can see, male employees are more than female.

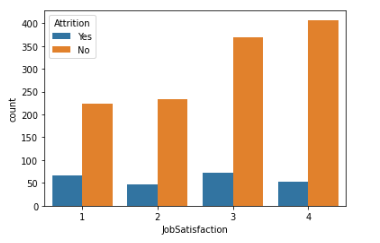
**Bivariate Analysis:**

****

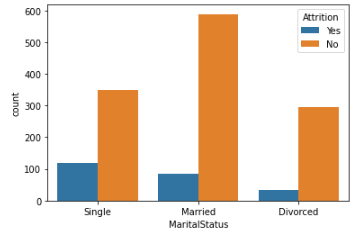
* We can see, around 170 male employees are wanting to leave a company. But more than 700 male employees want to stay.
* But in female up to 500 employees don’t want to leave a company.
* If we talking about both genders, the male has highest proportion whose want to leave.

****

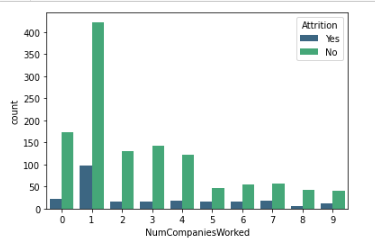
* We can see, laboratory technician has a higher Attrition Yes. It means they want to leave.
* Who’s having a sales Executive as job role, those least are willing to leave.
* Research Director has low proportion in job role.

****

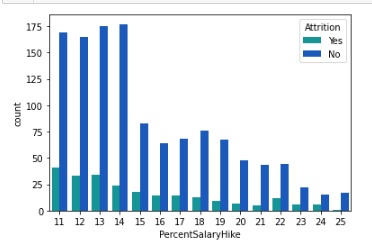
* In Job Satisfaction 1 have least difference between Yes and No.
* In Job Satisfaction 4 has highest employees are not want to leave.
* Job satisfaction whose says No is in increasing in nature.

****

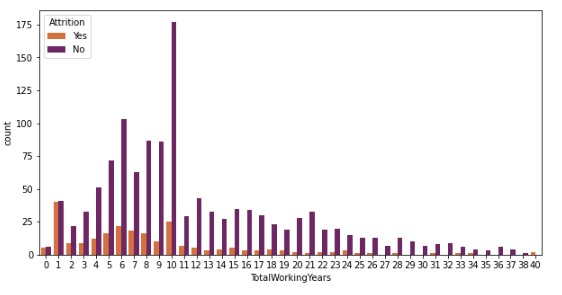
* We can see, those employees are single mostly they willing to leave. Their proportion is high.
* Married employee has least difference between Yes and No.
* Married employees are least interesting to leave a job.
* Divorced has low Yes among all.



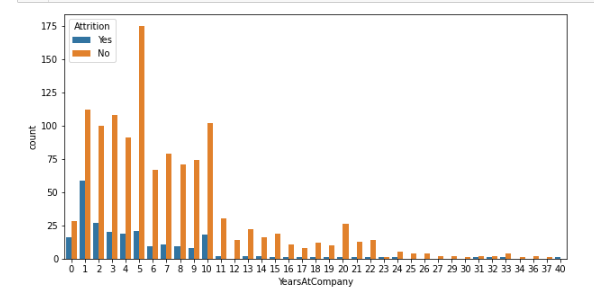
* Number of companies worked 1 is having highest proportion of Yes and No.
* Those employees who’s worked only 1 company they have highest chance to leave the company.
* We can see, those are worked more 6 they, are less willing to leave a company.



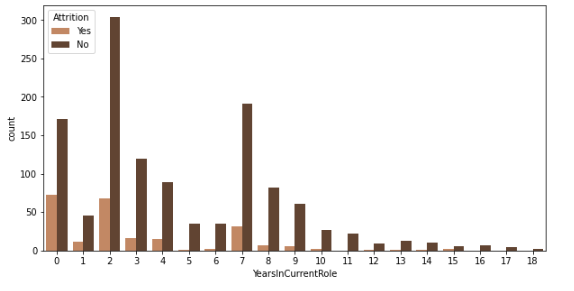
* We can see, Percent Salary Hike is a decreasing in nature.
* Employees have Percent Salary Hike less 14%, those employees are willing to leave.
* Percent Salary Hike > 17, those are less willing to leave a company.



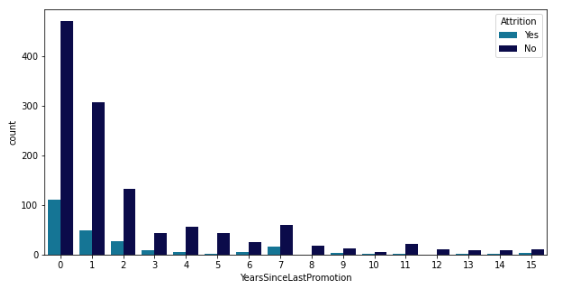
* We can see, up to 10 total working year is having same proportion of Attrition 'Yes' and 'No'. It means they are more willing to switch the company.
* But In total working year is 1, they have highest among all too willing switch a company.
* After the 25 year of total working, there are very less employees too willing to leave. And they are in decreasing in nature.



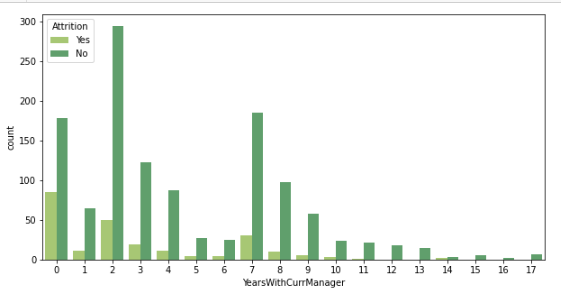
* Year at company 5 has a highest attrition No. It means that they don’t want to go.
* But up to 5 year at company, they are more willing to leave a company.
* After 17 year at company, they have almost all not willing to leave.



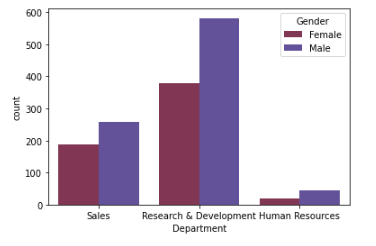
* Those employees have spent 0 and 2 year at company they have more chance to leave to company because they spend 2 year at same role/ same position.
* After the 7 year since current role they have less chance to leave the company.



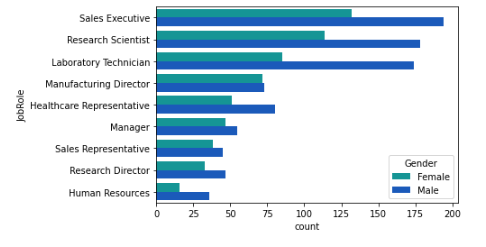
* Year since last promotion 0 has more willing to leave. But these are the fresher for the company. And Might they have started their career recently.
* After 3 year of last Promotion, they are less willing to switched.



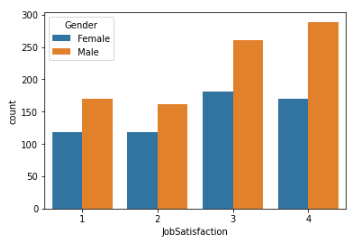
* We can see, those are spent more than 5 year with same manager, they are less willing to leave the company.
* But Those are spent 0 to 2 year with same manger they more willing to leave the company.



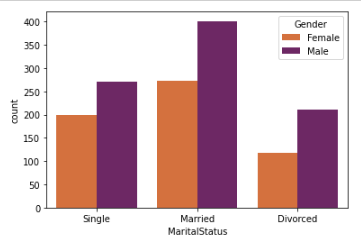
* We can see, In Research and development department, female employees are more. At same department male are also more.
* Least employees in HR department.



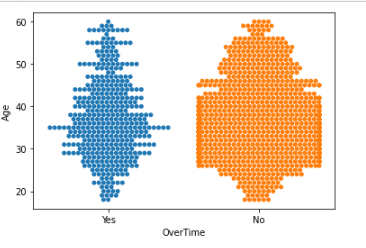
* We can see, all job role is having decreasing in nature with respect to gender.
* But in Manufacturing director male and female have almost equal proportion.



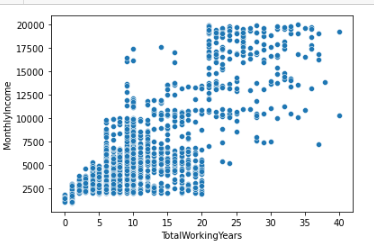
* male employees are having more job satisfaction then female.
* male job satisfaction is increasing in nature but female after 3 it slightly decreases.



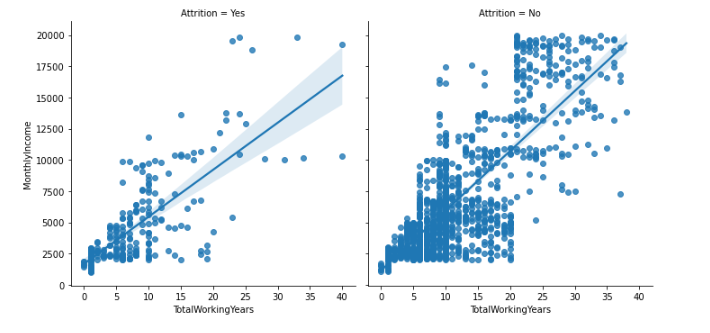
* we can see married male and female are more employees than other.
* Single marital status employees are equal up to 190.
* Divorced marital status employees are equal up to 110.



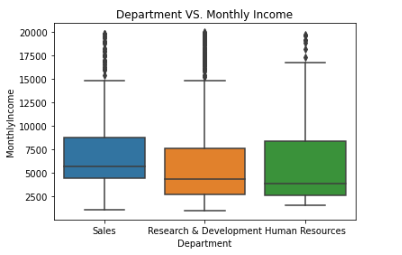
* We can see, 27 to 42 at this age group the employees willing to do overtime.
* But at same age group, the employee doesn’t want to do overtime.
* Proportion of 'No' is higher than 'Yes'.



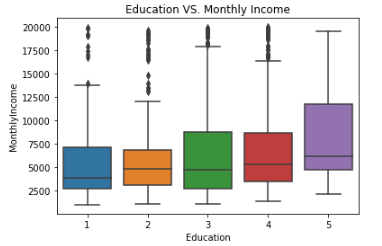
* As we can see, Monthly income and total working year has linear relationship with each other.
* As total working year increases then monthly income also increases.



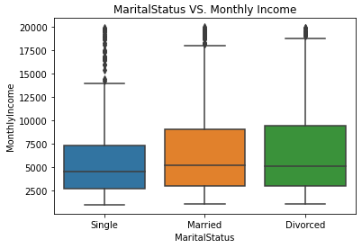
* We can see that, the employees who had more working years in the company they are more attached to their jobs, as they also have a higher monthly income.
* We can see, Both Attrition has linear relation to both features.



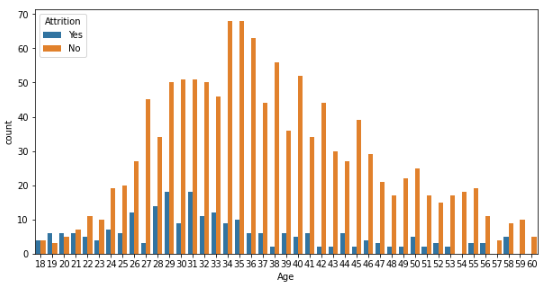
* We can see, Sales has highest monthly income. Whose working in sales department those are having higher salary then other.
* Research and Development department have lower salaries then HR.



* We can see, +2 employees are having lower monthly income than others.
* Doctor degree holding employees is having highest monthly income.

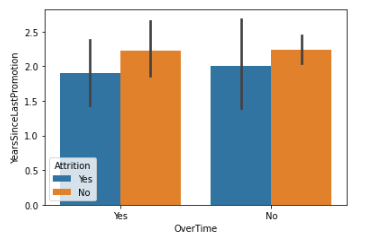


* we can see single employees have low Monthly Income compared by other, so that's why they leave more.
* Divorced employees have high monthly income than others.

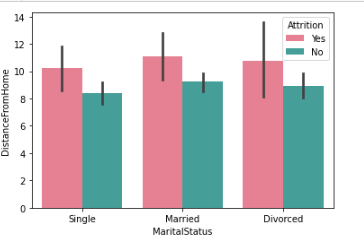


* We can see that the most common age of employees is around 30 - 40
* We can see, age from 26 to 32 they are more willing to leave the company. But in the same age group they are less employees are willing to leave.
* Up to the age of 30 Attrition 'Yes' is in increasing in nature. After the age of 30 Attrition 'Yes' is decreasing in nature.

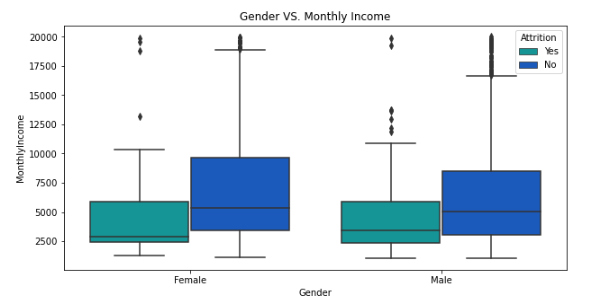
**Multivariate Analysis:**



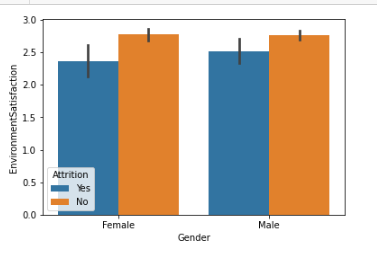
* Employees who put more efforts by working overtime seems to take an average more years to get promoted.
* And such treatment by the company could demotivate the hardworking employees and they why they leaving the company.



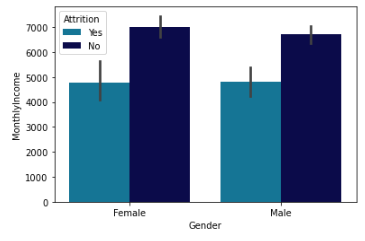
* Married employees have more attrition rate then other.
* Married employees have distance from home is high they why they might will to left job.



* As we can see, Female has little more monthly income then male.
* We can see, monthly income is affecting the attrition.
* Male Attrition is higher than female.

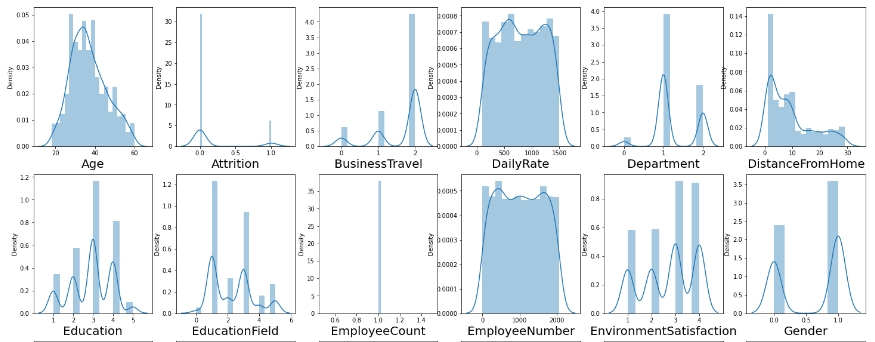


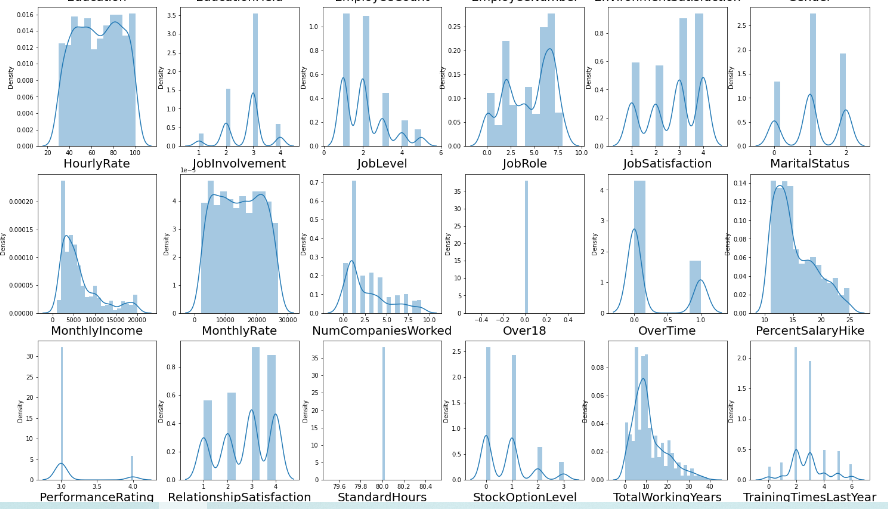
* In male have more environment satisfaction then female but both are having very close difference.
* Attrition 'Yes' is high in male then female.



* We can see, those are having monthly income low they are more willing to leave.
* Females are more comparatively male are willing leave because their attrition Yes is high.
* Both male and female are having same pattern.

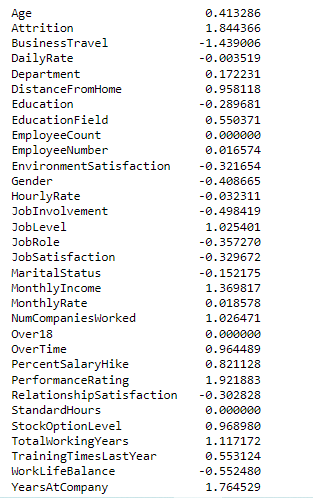
### **Checking Distribution of data:**





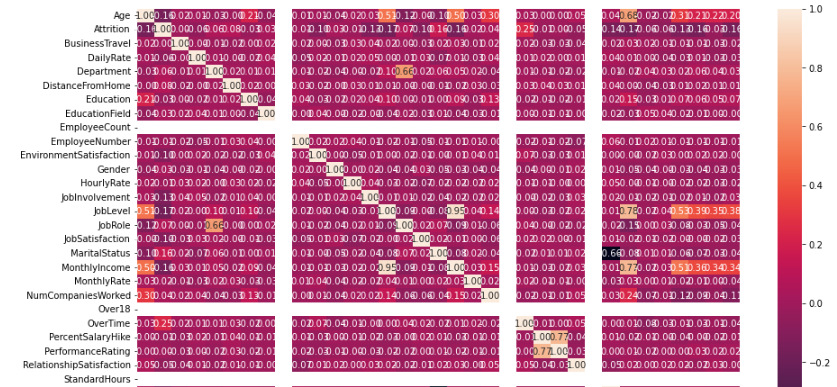
* All feature seems to skewed.
* We can see, Age and Monthly Income are positively skewed.
* Job Role and Work Life Balance are Negative skewed

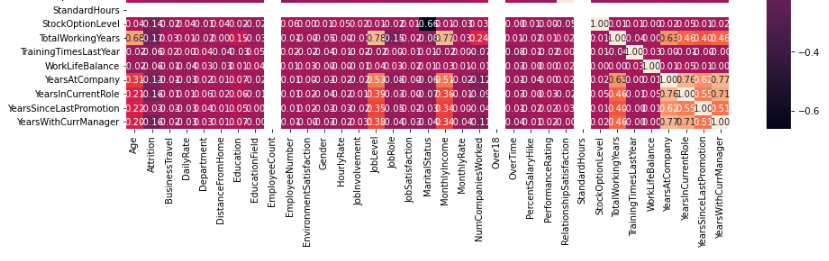
**Checking Skewness in data:**

****

* Outcome of the skewness: Skewness threshold is taken is +/-0.65.
* Almost all continuous data are having skewness.

**Finding the Correlation:**

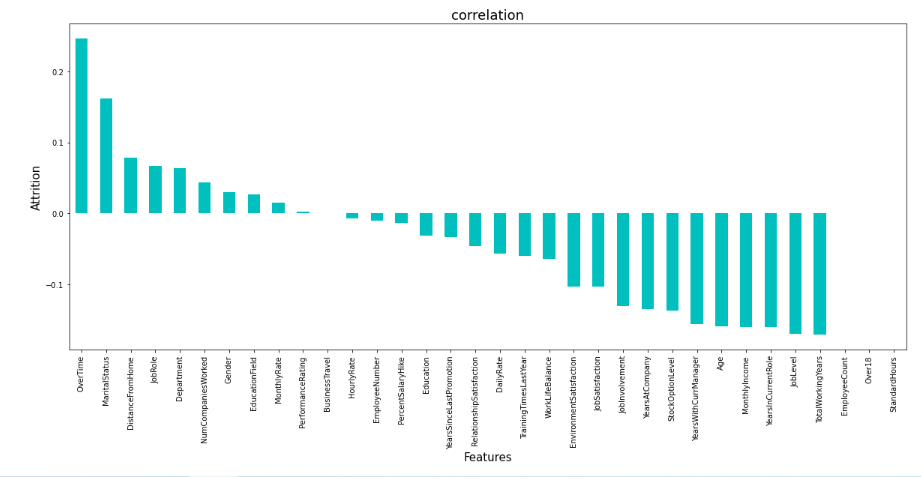
****

****

**Outcome of Correlation**

1. Job Level has 51% positive correlation with target column.
2. Monthly Income has 51% positive correlation with target column.
3. Total Working Year has 68% positive correlation with target column.
4. Job Satisfaction, Percentage Salary Hike and Performance Rating has 0% correlation with target column.
5. Age has 68% positive correlation with Total Working Year.
6. Department has 66% positive correlation with Job Role.
7. Job Level has 95% positive correlation with Monthly Income and 78% positive correlation with Total Working Year.
8. Monthly Income has 77% positive correlation with Total Working Year.
9. Percent Salary Hike has 77% positive correlation with Performance Rating.
10. Total Working Year has 63% positive correlation with Year at Company.
11. Year at Company has 77% and 76% positive correlation with Year With Current Manager and Year Sin Current Role.

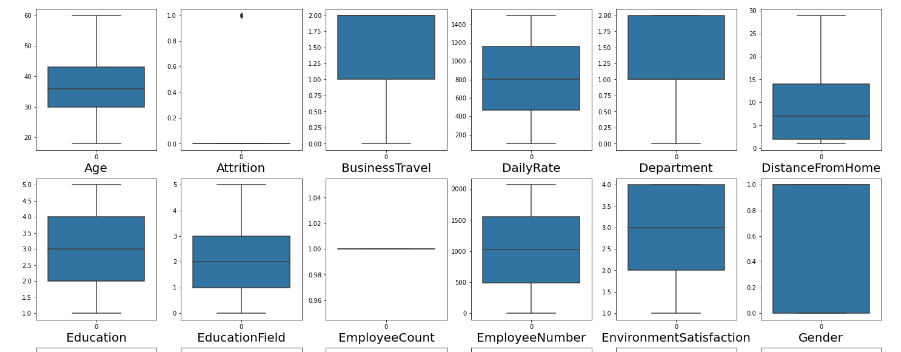
**Visualizing correlation of feature columns with label column.**

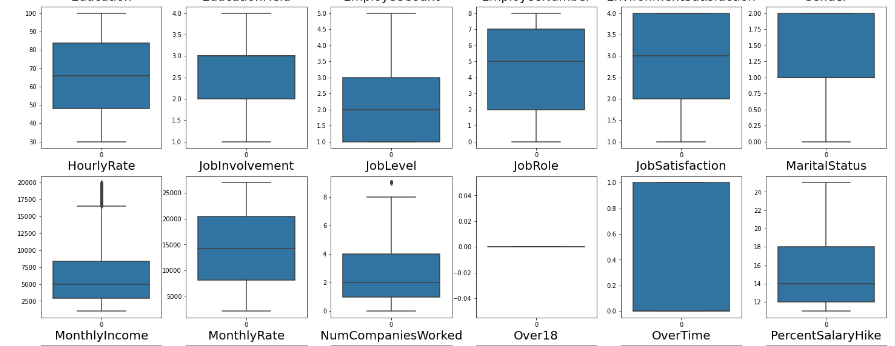
****

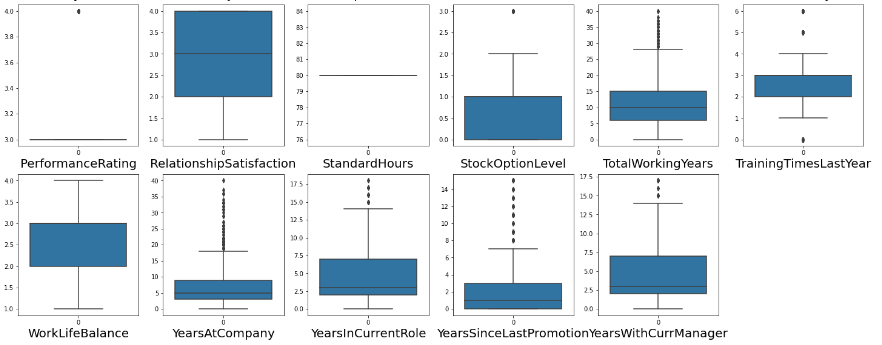
* It is observed that Over Time, Marital Status and Distance from Home have the highest positive correlation with Attrition.
* While Total Working Year, Job Level and Year Sin Current Role have the highest negative correlation with Attrition.

**Data Pre-Processing**

**Checking for Outliers in columns**





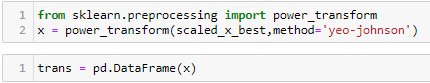
****

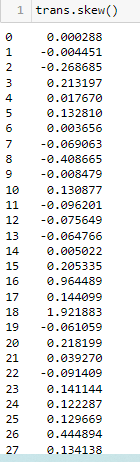
* Monthly Income, Total Working Year, Year Sin Current Role, Year at Company, Year Since Last Promotion are have a more Outliers.
* Some features are not having outliers but those are categorical features.
* But some categorical features are having outliers.

### **Normalizing Data Distribution using Power Transformer:**

The Yeo-Johnson power transformer method is used to transform the values of the columns whose data distributions are skewed. The optimal parameter for stabilizing variance and minimizing skewness is estimated through maximum likelihood.

Using the code below the data distribution was normalised.



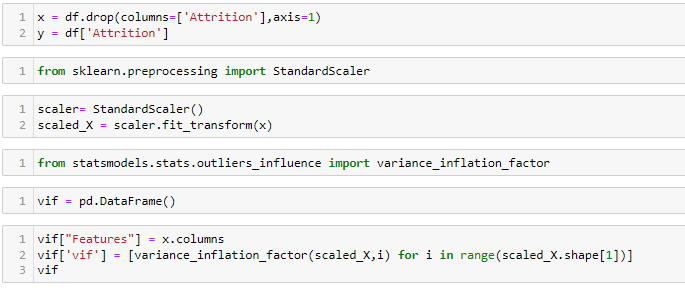


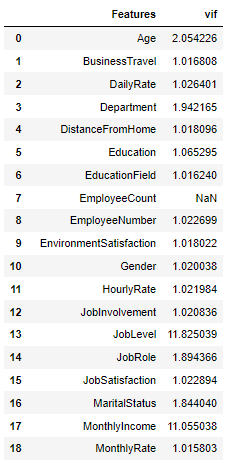
It is observed that Skewness has been greatly reduced.

Next step is to select the best features which would build the most accurate Machine Learning Models to predict the target variable.

### **Checking for Multicollinearity using Variance Inflation Factor:**

Variance inflation factor measures how much the variance of an independent variable is influenced / inflated, by its interaction/correlation with other independent variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error.

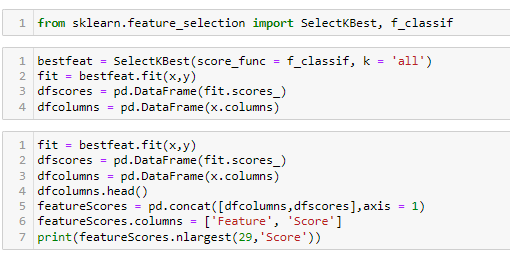


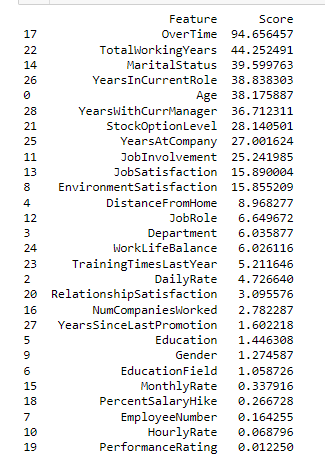
****

* If VIF > 10, It means multicollinearity is present.
* Multicollinearity exists in Job Level, Monthly Income, Based on ANOVA F scores.
* Those features are having lowest correlation with target column that feature will be dropped.

### **Selecting K best Features:**

Based on the respective ANOVA f-score values, the feature columns are selected that would best predict the Target variable, to train and test machine learning models.

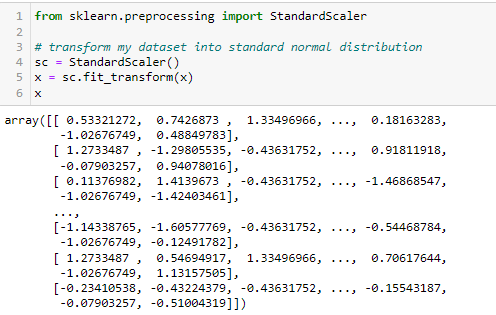


Upon analyzing the scores of each column, it is decided that the columns with the lowest scores, as well as the highly collinear column “Business Travel” will be dropped.

**Feature scaling:**

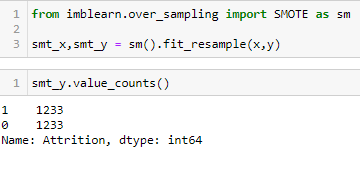
Scaling the values in the feature columns using Standard Scaler in order to normalize the range of data.



**Balancing Dataset:**

Imbalanced classification involves developing predictive model on classification datasets that have a severe class imbalance. But our data is imbalanced so we need to balanced it, for this I used SMOTE technique to balancing the data.

SMOTE first selects a minority class instance a at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the features space. The synthetic instances are generated as a convex combination of the two chosen instances a and b.

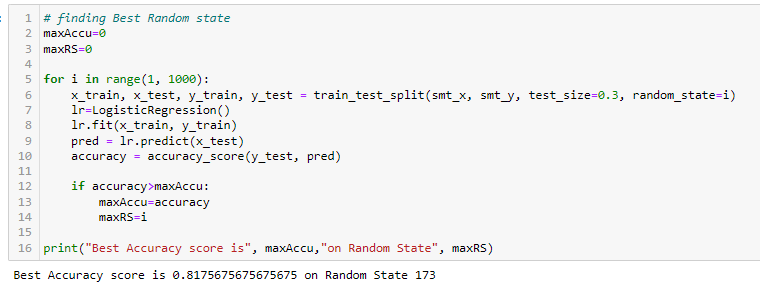
****

**Classification Model Building:**

**Logistic Regression**

**Finding the Best Random State**

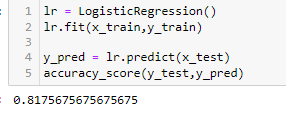
The best random state has to be determined, which will then decide the splitting of data into train and test indices in the most optimal way, that yields maximum model prediction accuracy.



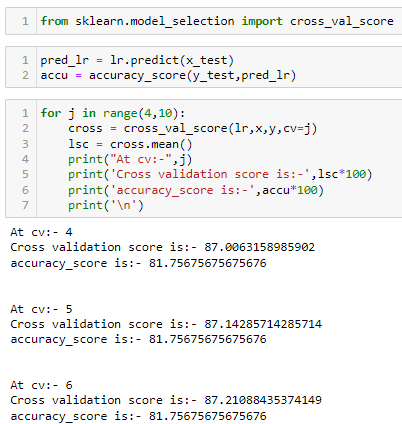
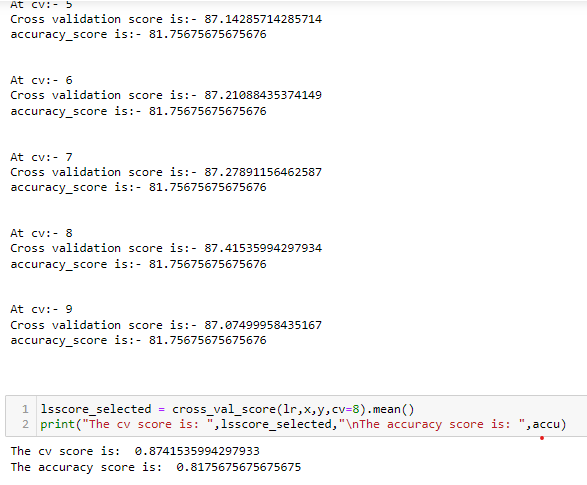
**Creating Train-Test split based on random state obtained above**



### **Training the Models and getting accuracy**

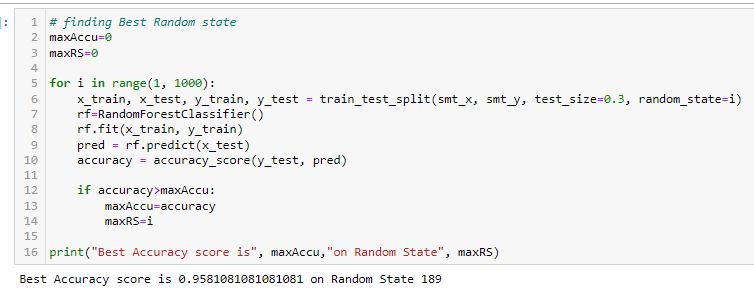


**Cross Validation for Logistic Regression:**

**** ****

**Random Forest Classifier**

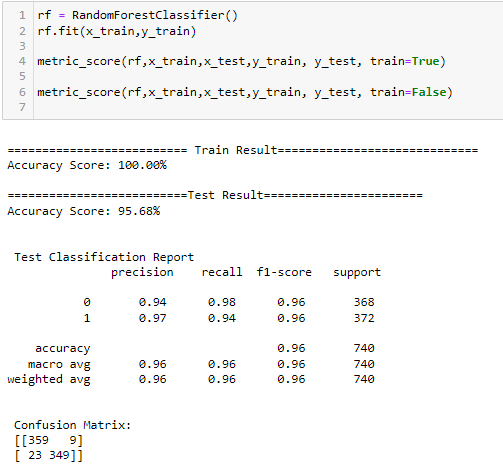
**Finding Best Random State**

****

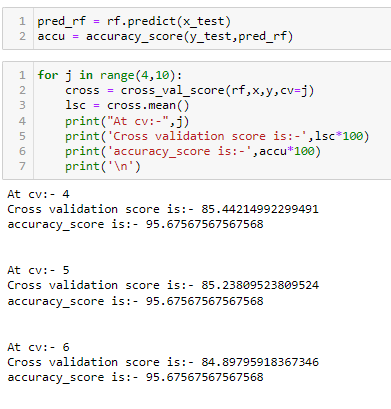
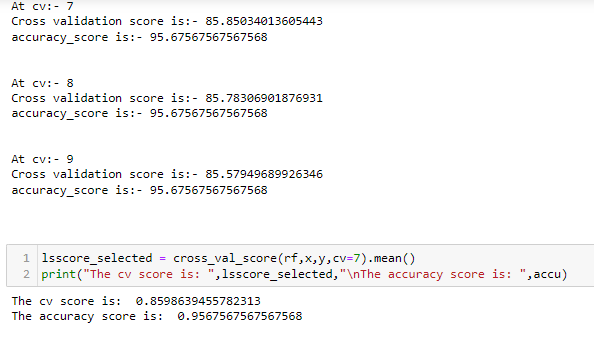
**Creating Train-Test split based on random state obtained above**

****

**Train and Test the model**

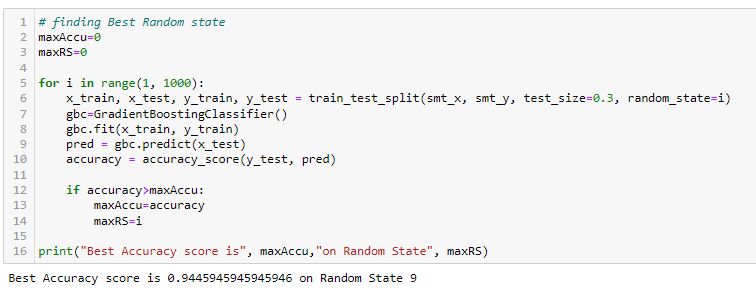
****

**Cross Validation of the model:**

**** ****

**Gradient Boosting Classifier**

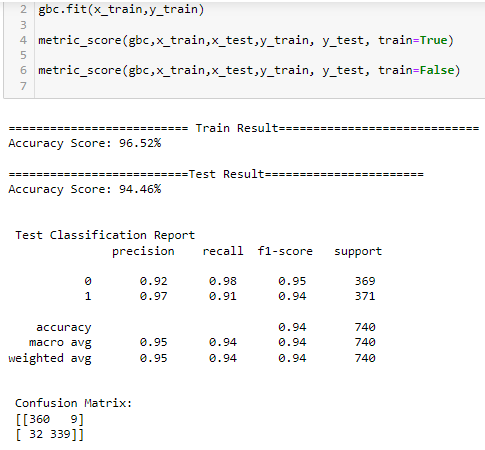
**Finding Best Random State**

****

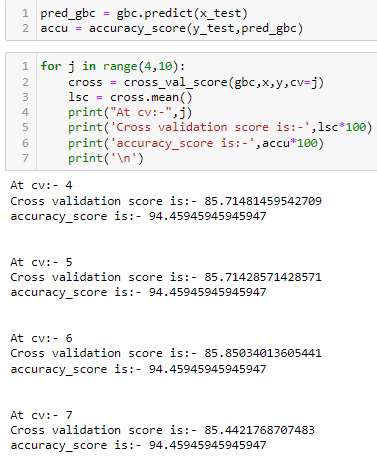
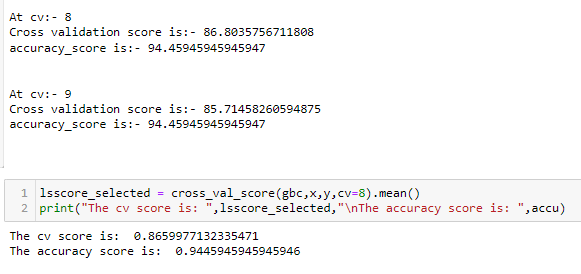
**Creating Train-Test split based on random state obtained above**

****

**Train and Test the Model**

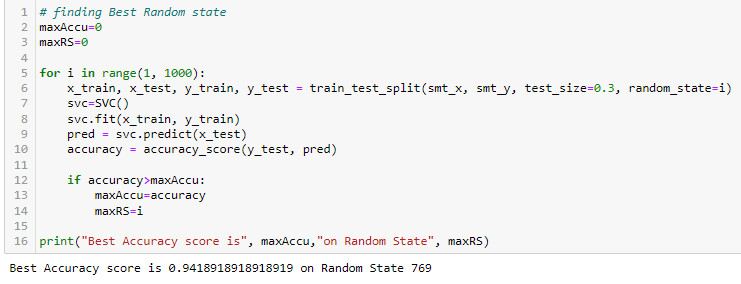
****

**Cross Validation of the model**

**** ****

**SVC:**

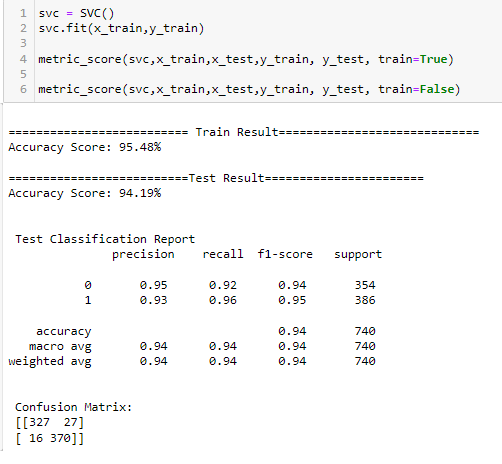
**Finding Best Random State**

****

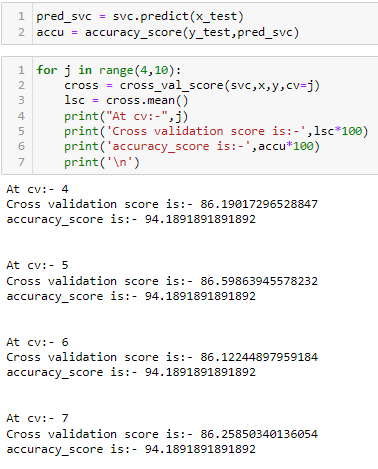
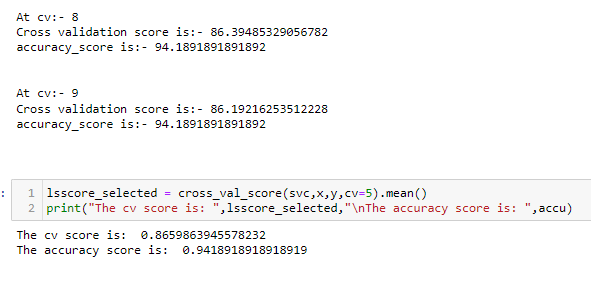
**Creating Train-Test split based on random state obtained above**

****

**Train and Test Model**

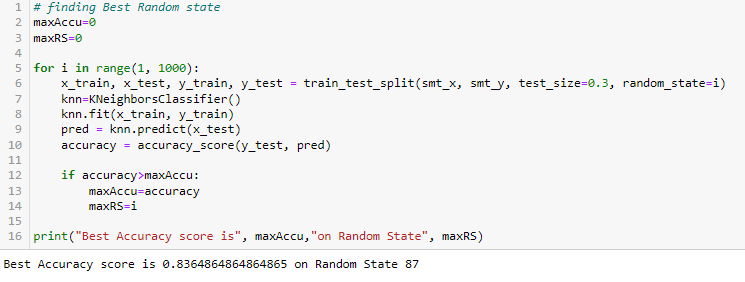
****

**Cross Validation of the Model**

**** ****

**K Nearest Neighbors:**

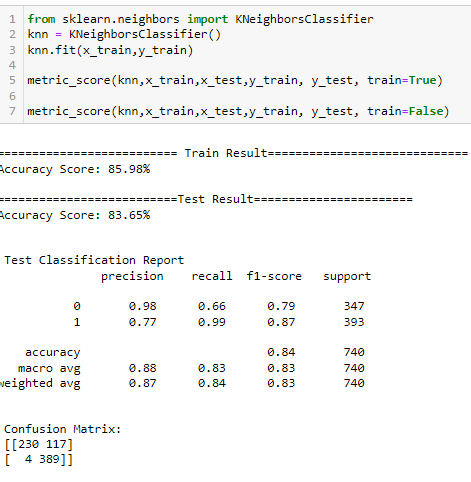
**Finding Best Random State**

****

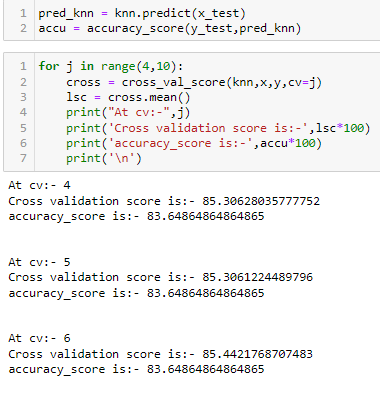
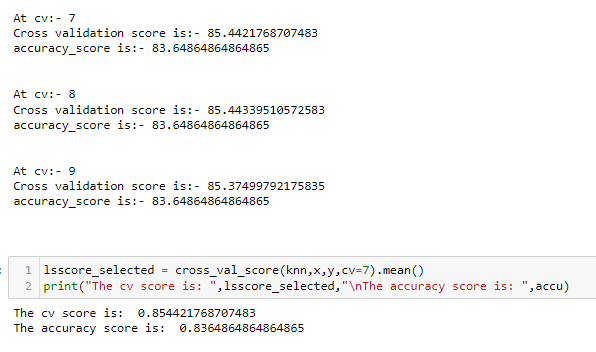
**Creating Train-Test split based on random state obtained above**

****

**Train and test the Model**

****

**Cross Validation of the Model**

**** ****

**Selecting Best Model**

**Based on comparing Accuracy Score results with Cross Validation results, it is determined K Neighbors Classifier is the best model. It has least difference between accuracy score and cross validation.**

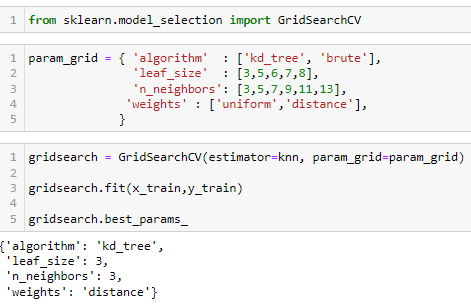
### **Hyper Parameter Tuning:**

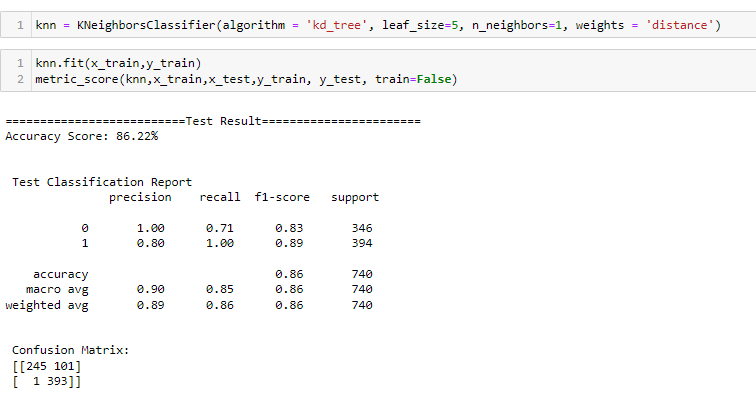
Grid Search CV is used for Hyper Parameter Tuning of the KNN Classification model.

Based on the input parameter values and after fitting the train datasets,

The KNN Classification Model was further tuned based on the parameter values yielded from Grid search CV.

The Tuned KNN Classification Model displayed an accuracy of 86.22%

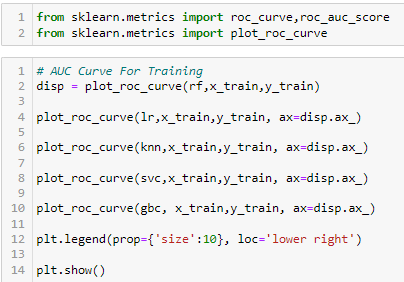
****

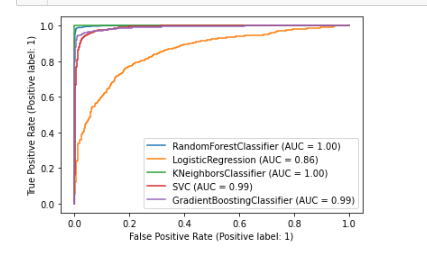
****

### **ROC AUC Curve:**

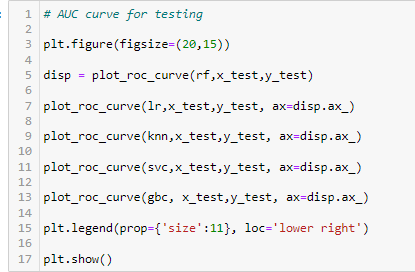
An ROC curve plot TPR Vs FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positive and True Positive.

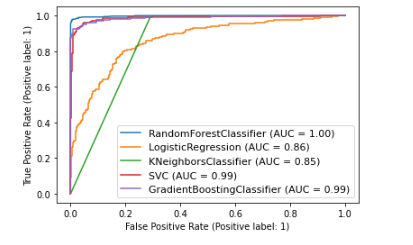
**Roc AUC Curve for Training**

****

****

**ROC AUC Curve for Testing**

****

****

**Concluding Remarks:**

* It’s better for company to get employees at young age as they are more likely to stay.
* It’s better for company to get single employees as they are more likely to stay. But have some consequences i.e. They are very free bird; they don’t have that much of responsivities as married employees have. So, need to have balance this by giving some incentive to single (marital status) employees.
* It’s better for company to hire more female employees due this we will avoiding the gender-based discrimination. As we find that men are more likely to leave the company and increase the attrition rate.
* Company should have to hire employees at any education degree except bachelor because bachelor degree holder has a higher rate of attrition. But it’s very difficult task to company to get employees other than bachelor. So, avoiding this company need to implement contract system where employees are bound to be company asset for particular year.
* Company should have to give more focus on research and development department because it is backbone of company. This department have higher attrition among the other department. For avoiding attrition company should have to increases salaries of this department. And one more thing that, company need to spend more in R&D apart from increasing salaries like instrument, budget etc.
* It’s better for company to keep maintain relationship of employees with their managers without replacing them because some employees are very comfortable to their manager. As new managers are come, employees will take a time to adjust with or make bond with new manager.
* Company should have to take care of, more employees’ satisfaction in job level, environment. It has a highly affected to the attrition rate.
* Company really should appreciate their employees who stayed more than 10 years by still giving promotion and incentives.