**HR Analytics - Understanding the Attrition in HR**

**What is HR Analytics?**

A lot of HR professionals who want to be more data-driven in their work ask the question “what is HR analytics?”

HR analytics, also referred to as people analytics, workforce analytics, or talent analytics, involves gathering together, analyzing, and reporting HR data. It enables your organization to measure the impact of a range of HR metrics on overall business performance and make decisions based on data. In other words, HR analytics is a data-driven approach toward Human Resources Management.

In the past century, Human Resource Management has changed dramatically. It has shifted from an operational discipline towards a more strategic one. The popularity of the term Strategic Human Resource Management (SHRM) exemplifies this. The data-driven approach that characterizes HR analytics is in line with this development.

By using HR analytics, you don’t have to rely on gut feeling anymore. Analytics enables HR professionals to make data-driven decisions. Being able to use data in decision-making has been growing in importance throughout the global pandemic. Moving towards a post-pandemic world, there are many changes happening in employment – whether it is the growing popularity of hybrid work or the increased use of automation. In this age of disruption and uncertainty, it is vital to make the correct decisions in order to navigate our new realities. In this article we are trying to give a solution to the major problem in HR, that is employee attrition in a company.

**Business Problem definition:**

**What is Attrition in HR?**

Attrition is a problem that impacts all businesses, irrespective of geography, industry and size of the company. Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff. As such, there is great business interest in understanding the drivers of, and minimizing employee attrition.

**How does Attrition affect companies?**

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. 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.

**How does HR Analytics help in analyzing attrition?**

In this context, with the help of the data set given by the organization, we will build a classification model, which will predict whether the employee will quit the organization or not. This prediction could greatly increase the HR’s ability to intervene on time and remedy the situation to prevent attrition. While this model can be routinely run to identify employees who are most likely to quit, the key driver of success would be the human element of reaching out the employee, understanding the current situation of the employee and taking action to remedy controllable factors that can prevent attrition of the employee.

This data set presents an employee survey from IBM, indicating if there is attrition or not. The data set contains 1470 entries with 35 attributes. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition.

While some level of attrition in a company is inevitable, minimizing it and being prepared for the cases that cannot be helped will significantly help improve the operations of most businesses. As a future development, with a sufficiently large data set, it would be used to run a segmentation on employees, to develop certain “at risk” categories of employees. This could generate new insights for the business on what drives attrition, insights that cannot be generated by merely informational interviews with employees.

**The Data Set:**

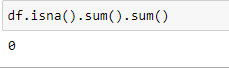
IBM has gathered information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees. Below is the list of attributes and definition of our dataset.

Our target variable: ATTRITION

|  |  |
| --- | --- |
| **Name** | **Description** |
| AGE | Age of the employee - Numerical Value |
| ATTRITION | Employee leaving the company |
| BUSINESS TRAVEL | How frequent the employee will travel (No Travel, Travel Frequently, Travel Rarely) |
| DAILY RATE | Salary Level - Numerical Value |
| DEPARTMENT | Department of the employee (HR, R&D, Sales) |
| DISTANCE FROM HOME | The distance from work to home - Numerical Value |
| EDUCATION | Education level - Numerical Value |
| EDUCATION FIELD | Education field (HR, LIFE SCIENCES, MARKETING, MEDICAL SCIENCES, OTHERS, TEHCNICAL) |
| EMPLOYEE COUNT | Numerical Value |
| EMPLOYEE NUMBER | Employee ID - Numerical Value |
| ENVIROMENT SATISFACTION | Satisfaction with the environment - Numerical Value |
| GENDER | Gender of the employee (FEMALE, MALE) |
| HOURLY RATE | Hourly salary - Numerical Value |
| JOB INVOLVEMENT | Job involvement - Numerical Value |
| JOB LEVEL | Job level - Numerical Value |
| JOB ROLE | Role of the employee (HC REP, HR, LAB TECHNICIAN, MANAGER, MANAGING DIRECTOR, REASEARCH DIRECTOR, RESEARCH SCIENTIST, SALES EXECUTIEVE, SALES REPRESENTATIVE) |
| JOB SATISFACTION | Satisfaction with the current job - Numerical Value |
| MARITAL STATUS | Marital status (DIVORCED, MARRIED, SINGLE) |
| MONTHLY INCOME | Monthly salary - Numerical Value |
| MONTHY RATE | Monthly rate - Numerical Value |
| NUMCOMPANIES WORKED | No. of companies worked at - Numerical Value |
| OVER 18 | Over 18 years? (YES, NO) |
| OVERTIME | Overtime (NO, YES) |
| PERCENT SALARY HIKE | Percentage increase in salary - Numerical Value |
| PERFORMANCE RATING | Performance rating - Numerical Value |
| RELATIONS SATISFACTION | Relations satisfaction - Numerical Value |
| STANDARD HOURS | Standard hours - Numerical Value |
| STOCK OPTIONS LEVEL | Stock options - Numerical Value |
| TOTAL WORKING YEARS | Total years worked - Numerical Value |
| TRAINING TIMES LAST YEAR | Hours spent training - Numerical Value |
| WORK LIFE BALANCE | Time spent between work and outside - Numerical Value |
| YEARS AT COMPANY | Total number of years at the company - Numerical Value |
| YEARS IN CURRENT ROLE | Years in current role - Numerical Value |
| YEARS SINCE LAST PROMOTION | Last promotion - Numerical Value |
| YEARS WITH CURRENT MANAGER | Years spent with current manager - Numerical Value |

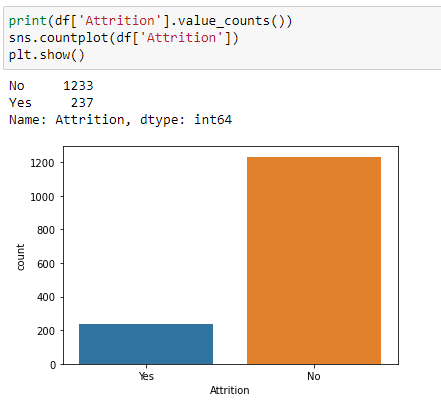
**Exploratory Data Analysis:**

* Observations from the features and values, we have features with categorical data and continuous data. So let’s split the categorical and continuous features for the further analysis.
  + **Categorical features**: 'Education', 'EmployeeCount', 'EnvironmentSatisfaction', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'PerformanceRating', 'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel', 'TrainingTimesLastYear', 'WorkLifeBalance', 'Attrition', 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'Over18', 'OverTime'
  + **Continuous features**: 'Age', 'DailyRate', 'DistanceFromHome', 'EmployeeNumber', 'HourlyRate', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'
* **Let’s look is there any missing value in the data set.**



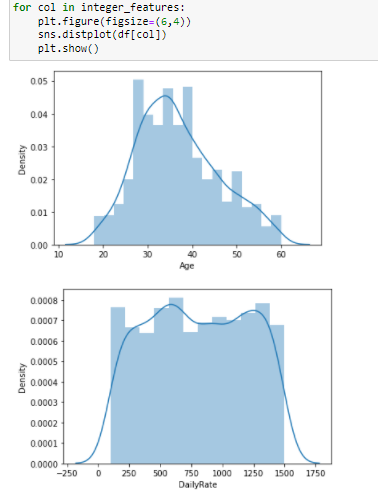
The above screenshot shows that there is no missing value in our dataset.

* **Let’s check how many records with target variable ‘Attrition’ YES and ‘Attrition’ NO.**



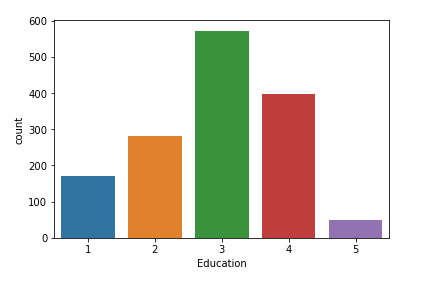
The plot shows that we have **imbalanced categorical data in our data set.** Which should be handled to build a non-biased model. The imbalanced categorical data will affect the model being biased to specific category. To avoid this, we will be using Oversampling technique to make the data balanced.

* **Let’s check whether our continuous data are normally distributed or not.**

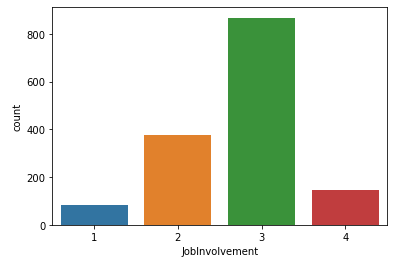


As we see the distribution diagrams, the feature ‘Age’ is normally distributed. But all other **continuous features are not normally distributed**. This will be handled by transforming the data of these features.

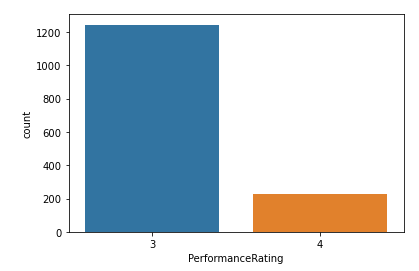
* **Let’s use count plot and visualize the categorical data.**
  + We have more employees whose education level is 3.



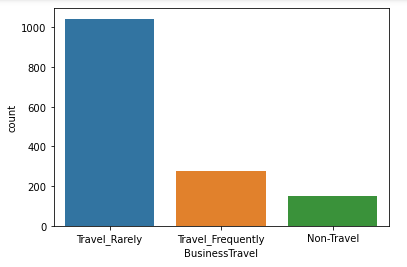
* + We have more number of employees whose Job Involvement is 3 and very less number of employees with Job Involvement 1.It shows that we have more employees with medium level of involvement. Not too involved and not too bad.



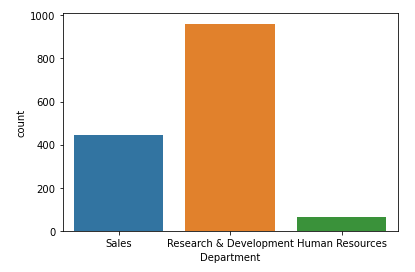
* + We have performance rating scale as 3 and 4. More employees got rating 3.



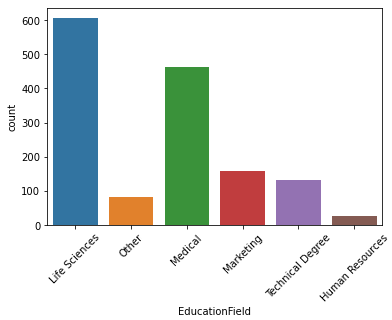
* + We have more employee who do business travel very rarely and less who does not travel at all.



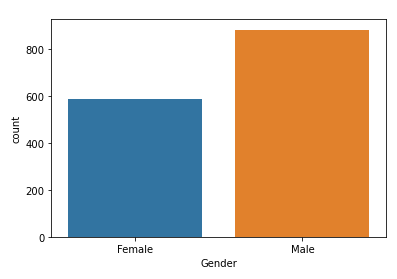
* + We have employees from 3 different departments. More employees are from Research&Development department, then followed by Sales and less employees are from HR department.



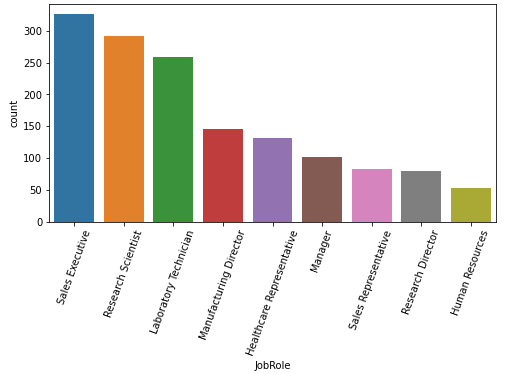
* + We have 6 categories of educational field. We have employees who studied LifeScience and followed by Medical are more than other categories.



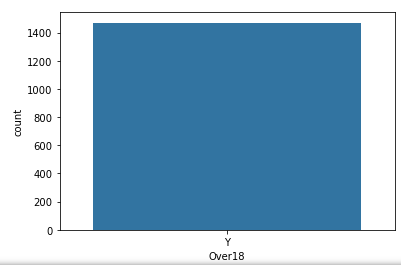
* + **More male employees than female employees**.



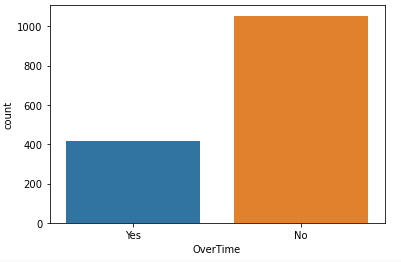
* + **We have 326 employees who are working as Sales Executive and 292 as Reserach Scientist and 259 as Laboratory Technician** and followed by few other job roles.



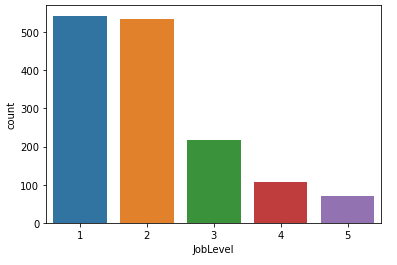
* + All the employees whose age are above 18 years.



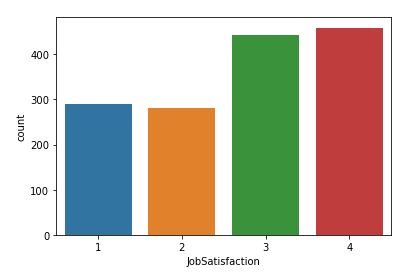
* + We have less employees who do overtime.



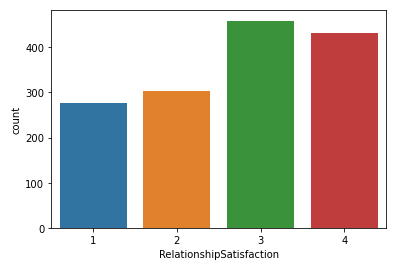
* + We have employees who work in 5 different job levels. More employees from level 1 and 2 and then followed by 3. This trend shows that, when the job level going higher, the number of employees are less.



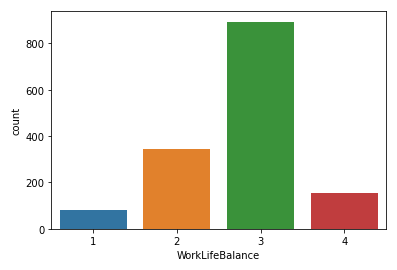
* + We have scale between 1 to 4 for employee job satisfaction and most of the employees are fully satisfied(4) with their job.



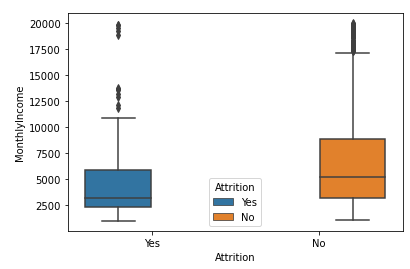
* + For relationship satisfaction we have scale between 1 to 4 and more employees are moderately satisfied (3), then followed by completely satisfied.



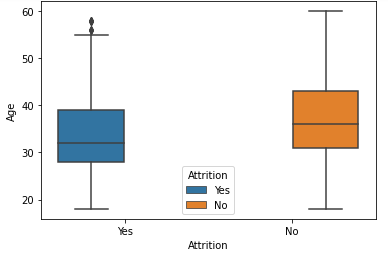
* + We have scale between 1 to 4 for work life balance of an employee and we have more employees who are moderately able to maintain their work life balance and we have very less employees who does not able to manage their work life balance.



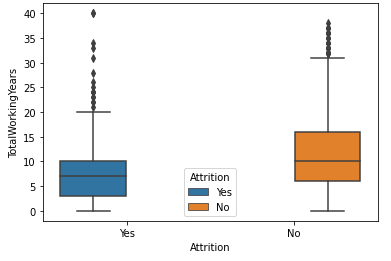
* **Let’s plot box plot and see, how the continuous data is scattered with our target variable which is categorical data.**
  + Employees whose monthly income is less than 7500 are mostly leaving the company.



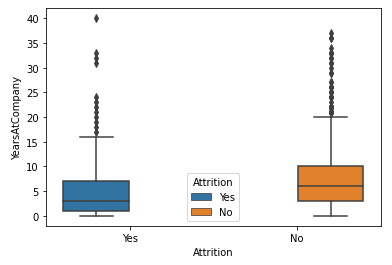
* + Employees whose age are between 25 to 40 are most likely to leave the company.



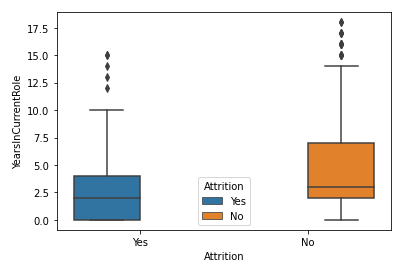
* + Employees whose total working years less than 10 are mostly leaving the company.



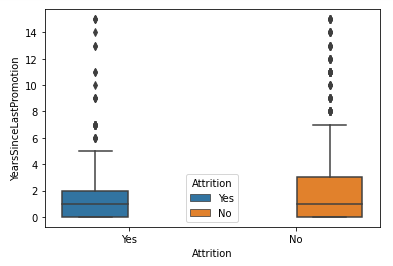
* + Employees whose years at company is less than 10 are mostly leaving the company.

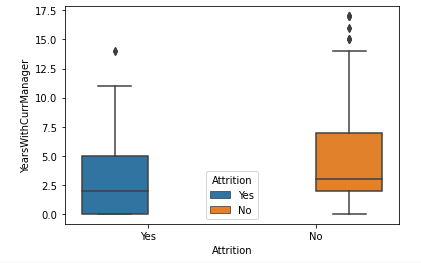


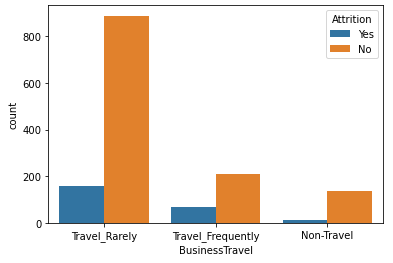
* + Employees who are working less than 5 years in current role are mostly leaving the company.



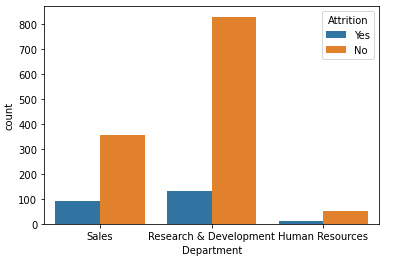
* + Employees who got their last promotion between 0 to 2 years are mostly leaving the company.



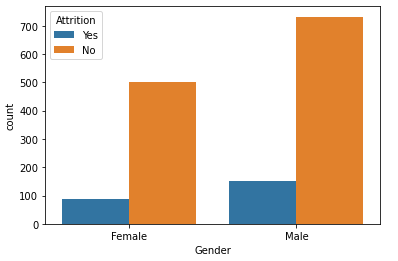
* + Employees who are working with current manager for less than 5 years are mostly leaving the company.
  + 
* **Let’s plot bar plot and see, how the categorical data is categorized with our target variable which is categorical data.**
  + Employees who travel rarely are mostly leaving the company.



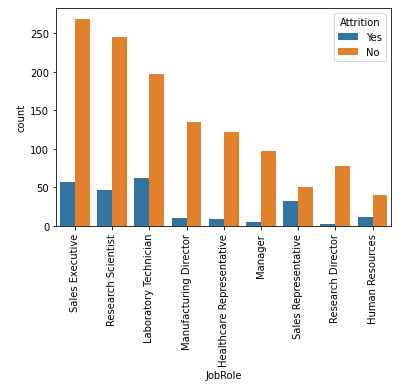
* + Employees who work from Research&Development department are leaving the company more and followed by Sales department.



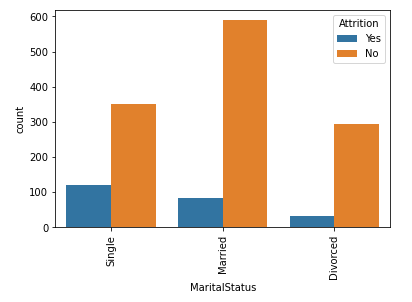
* + More male employees are leaving the company than female employees.



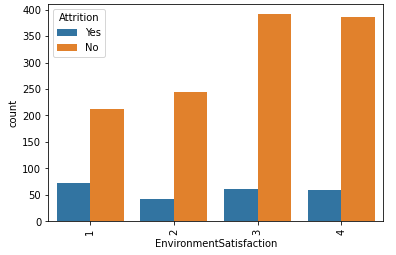
* + Employees who work as Laboratory technicians are leaving the company more and followed by Sales Executives and Research scientist.



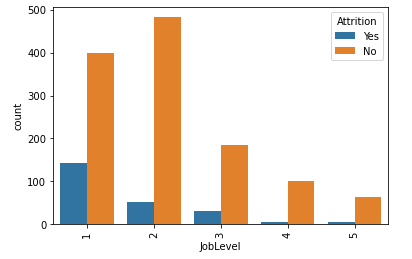
* + Employees whose marital status is single are leaving the company more.



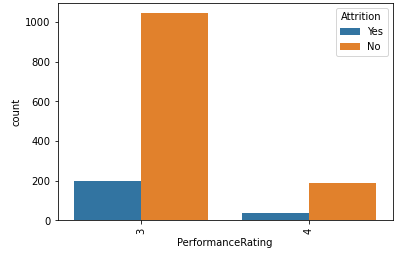
* + Employees who does not have good environmental satisfaction are leaving the company more.

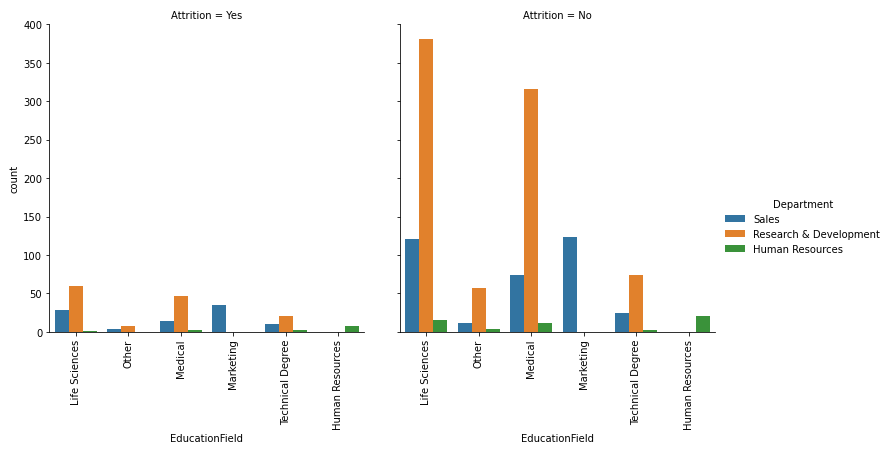


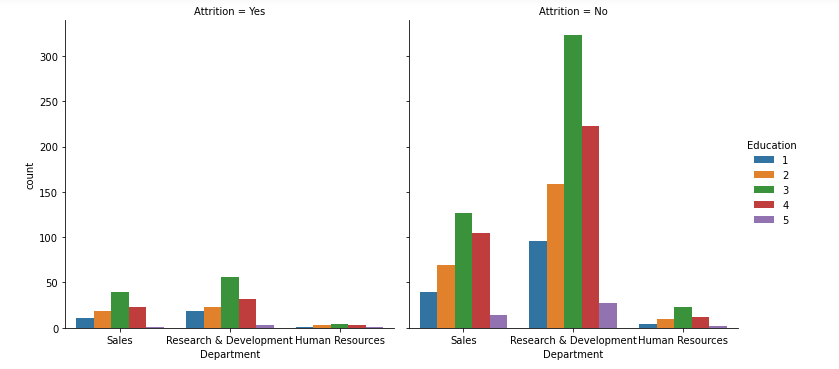
* + Employees who works in level 1 jobs are leaving the company more.



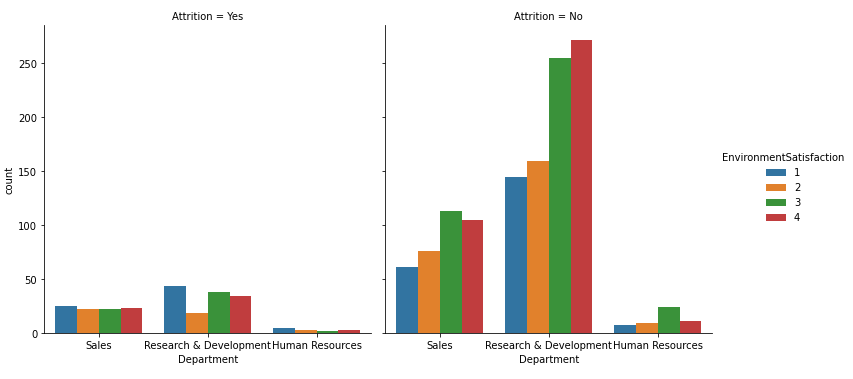
* + Employees who got performance rating 3 are leaving the company more.

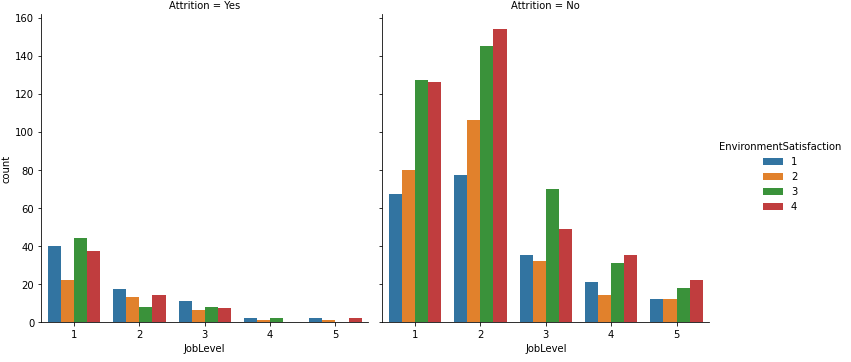


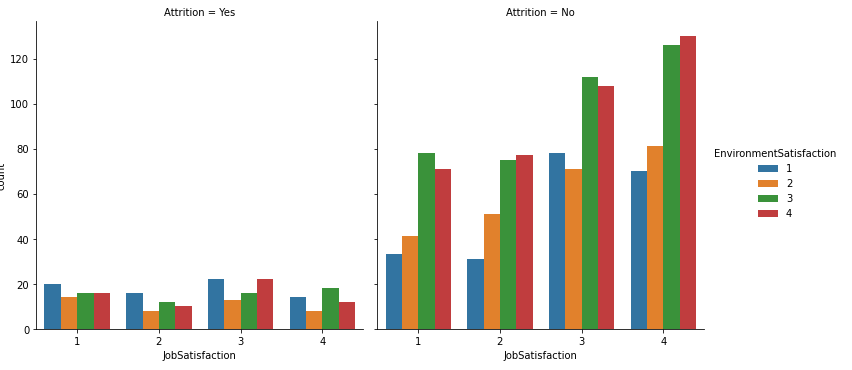
* **Let’s plot few categorical plot and see, how the categorized between the features.**
  + Employees who studied Life Science and Medical are working more in Research&Development department.
  + Employees who studied Marketing are working more under Sales department and they are most likely to leave the company.
  + 
  + Employees who work under Research& development department and whose education level is 3 are more likely to leave the company.



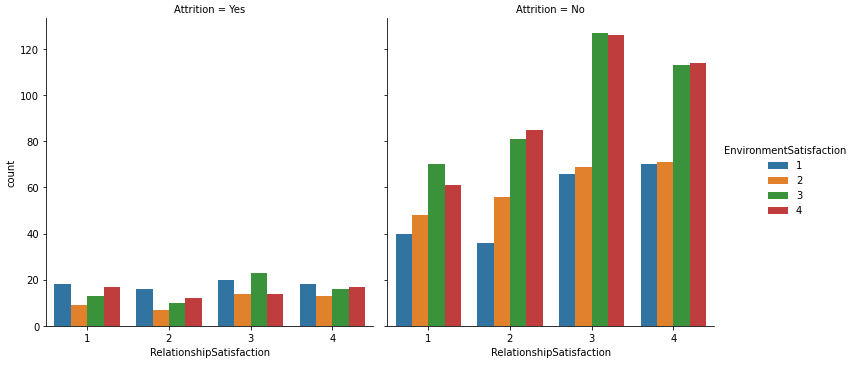
* + **Employees who works in Research&Developement department and who are having bad environment satisfaction are more likely to leave the company.**
  + **In all the departments, employees with less Environment Satisfaction are leaving the company more.**



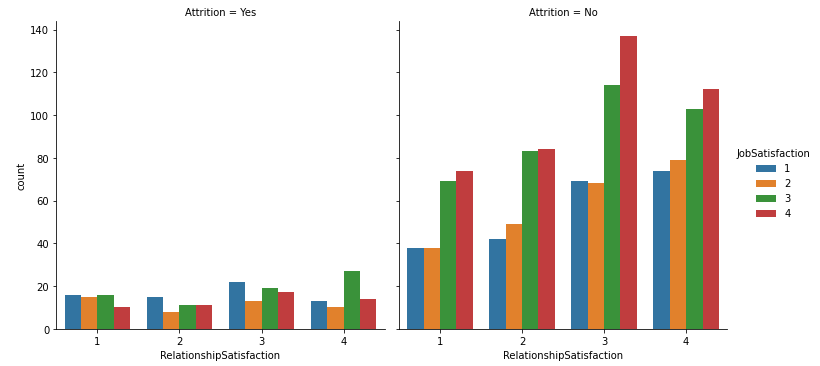
* + **Employees who work with level 1 jobs are not having good environment satisfaction and those employees are most likely to leave the company.**
  + 
  + Employees who have less Environment satisfaction and moderate job satisfation are leave the company more, followed by employees who have very less satisfaction in both environment and job satisfaction are leaving the company.



* + Employees who are having moderate Relationship satisfaction and less environment satisfaction are leaving the company more.



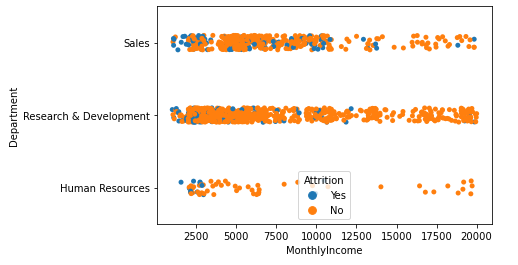
* + Employees who are having moderate Relationship satisfaction and less job satisfation are leaving the company more.



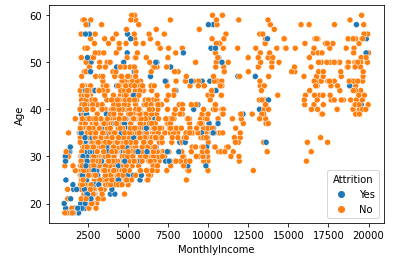
* + Employee from Sales department and whose job level is 1 are leaving the company more.
  + Employees whose salary is less than 15000 from Sales department are mostly leaving the company.



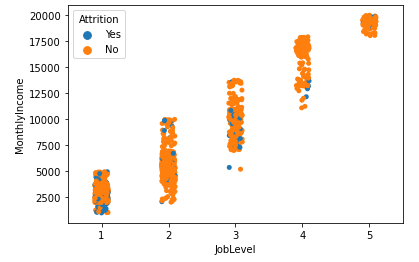
* + This plot shows that employees who are getting low salary from Sales department are leaving more.



* + From this plot we can observe that employees whose age between 18 to 35 and who are getting salary below 5000 are most likely to leave the company.



* + **When the job level increases the salary of the employee is increases. Employees works under job level 1 and 3 and getting low salary are leaving the company more.**

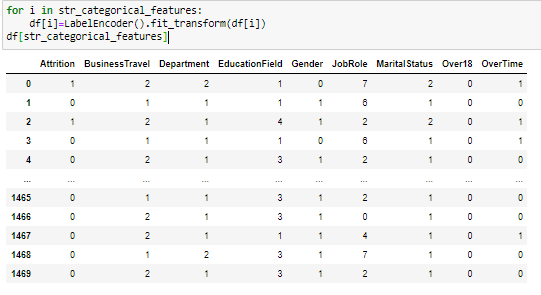


**EDA Concluding Remark:**

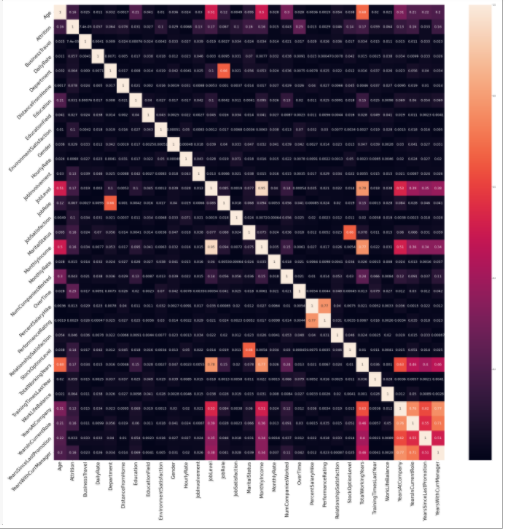
* **Insights/Observation from the above analysis:**
  + **From all the above observations we can see, employees who are leaving the company are mostly falls under the below category.**
    - **Employees with low monthly income,**
    - **Employees who are having less environment satisfaction,**
    - **Employees from job level 1,**
    - **Employees who got performance rating 3,**
    - **Employees with age between 25 to 40 ,**
    - **Employees who are having total work experience less than 10 years,**
    - **Employees who are travelling rarely.**

**Pre-Processing Pipeline:**

* **Encoding the non - numeric features:**
  + We have multiple categorical data with non-numeric type and these needs to be encoded as numeric type as the systems will understand only the numeric values.
  + We used LabelEncoding method to convert all the non-numeric features to numeric features.
  + **LabelEncoding will give the numeric value to each category value within the feature and convert the categories with the respective numbers.**



* **Checking Multi collinearity between features using heatmap**



**Observations from above heatmap**

1. Job level is strongly correlated with monthly income.

2. Job level is 78% correlated with Total working years

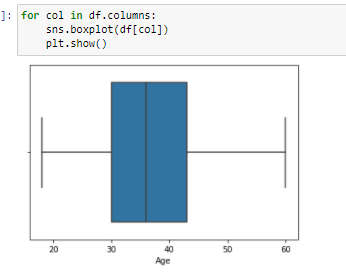
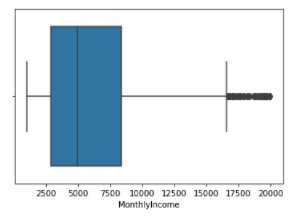
3. Monthly Income is 77% correlated with Total working years.

4. Performance rating is 77% correlated with Percentage of salary hike.

5. Years at company is 77% correlated with Years with current manager and 76% correlated with Years in current role.

**Checking Outliers:**

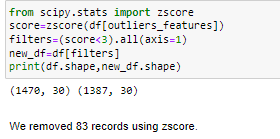
* **Let’s use box plot to see the outliers in each feature.**

As result of the box plots we identified there are outliers in the features ‘TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'.

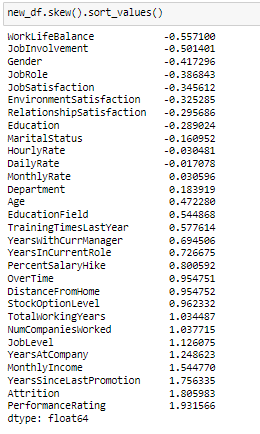
**Zscore Method:**

Zscore method is used to remove the outliers from the features. As part of outliers removal we removed 83 records from the dataset.



**Checking data skewness:**

Skewness is the amount of deviation from the normal distribution of each features. Generally, the features are expected with skewness of +/-0.5. If the skewness score is beyond this range, then it should be handled using any of the transformation techniques.



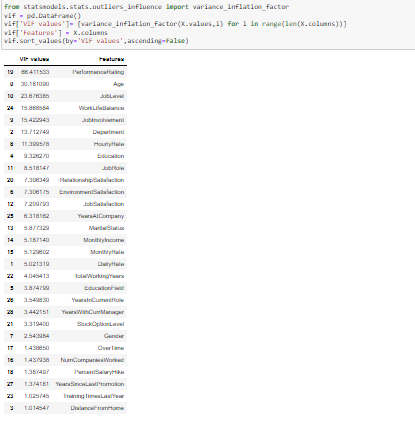
We have skewness in the features EducationField, TrainingTimesLastYear, PercentSalaryHike, YearsWithCurrManager, YearsInCurrentRole, OverTime, StockOptionLevel, JobLevel, NumCompaniesWorked, TotalWorkingYears, MonthlyIncome, YearsAtCompany, Attrition, PerformanceRating, YearsSinceLastPromotion.

Skewness should be handled for continuous data, there is no logic of skewness with categorical data.

**Here, in this project we handled the skewed features with PowerTransformer. Applying PowerTransformation on the skewed features helped us to remove the skewness completely.**

**Checking Multi collinearity using VIF score:**

VIF score tells us about the correlation between the features. Generally, when the VIF score if greater than 10, it is considered as having strong correlation with other feature/features. So lets check for the VIF score as below.

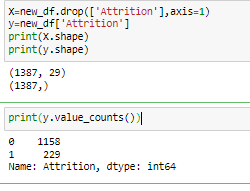


As we see in the screenshot, we have multi-collinearity problem with the features, this should be handled. The features ‘Age’, ’PerformanceRating’, ’JobLevel’, ’WorkLifeBalance’,’JobInvolvement’, ‘Department’, ’HourlyRate’ are having VIF score greater than.

As part of removing multi-collinearity problem, we dropped the features, ‘Age’, ’PerformanceRating’, ’JobLevel’, ’WorkLifeBalance’,’JobInvolvement’, ‘Department’. Now the data set is free from multi-collinearity problem.

**Building Machine Learning Models:**

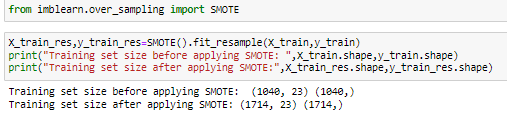
We are done with our analysis and preprocessing steps, now we are ready with data to build our model. Let’s split the list of features(X) and label (y).



As we see in above screenshot, our target label y has imbalanced data which needs to be handled.

**Applying SMOTE:**

SMOTE is one of the oversampling techniques, it will increase the number of minority class data to match with the majority class. We apply SMOTE technique on the training data alone.



As we see in the above screenshot, after applying SMOTE technique on the tarin data, the number of data has been increased from 1040 to 1714.

**Standard Scalar:** We will scale our feature data to the standard scale in such a way that the mean is 0 and standard deviation is 1.



**Building Base models:**

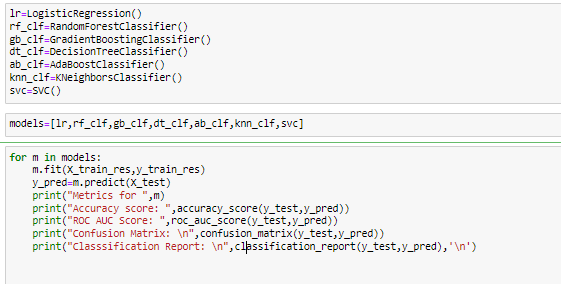
As we are working with classification problem and we not sure which algorithm will work better for our model, we are going to build the base models with multiple classification models and will select the best model among them. In this project, I am going to train the data with the below algorithms and see their performance.

Machine Learning Algorithms used:

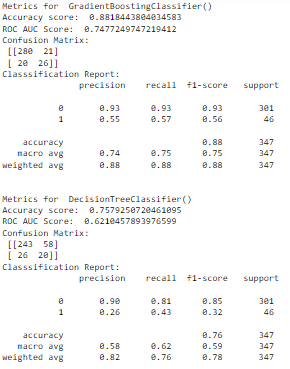
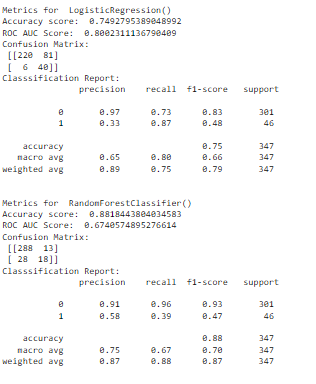
* LogisticRegression
* DecisionTreeClassifier
* KNeighborsClassifier
* SVC(Support Vector Classifier)
* Ensemble models:
  + RandomForestClassifier
  + AdaBoostClassifier
  + GradientBoostingClassifier

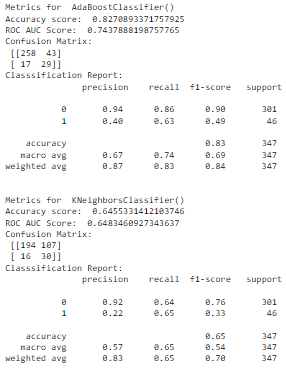
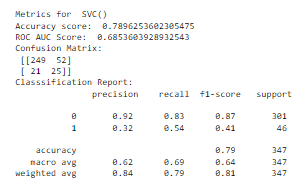


We instantiated the object for each model and trained the model with training data as below.



**And below are the results from building base models.**



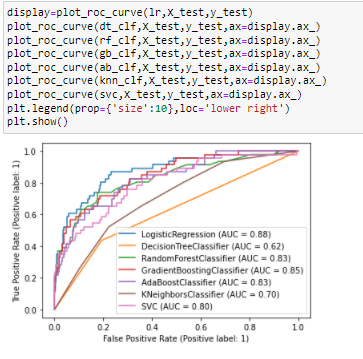
From the above screenshots, we could see RandomForestClassifier and GradientBoostingClassifier gives better base model accuracy than other models.

**Model Selection process:**

In this project, I am going to use **ROC AUC Plots and Cross validation score metrics for selecting the best model among the multiple models.**

Plotting ROC AUC Curve:

ROC AUC curve will help us to find the model which covers the maximum area under the curve. When the AUC (Area Under Curve) score is high, it means the model works better.

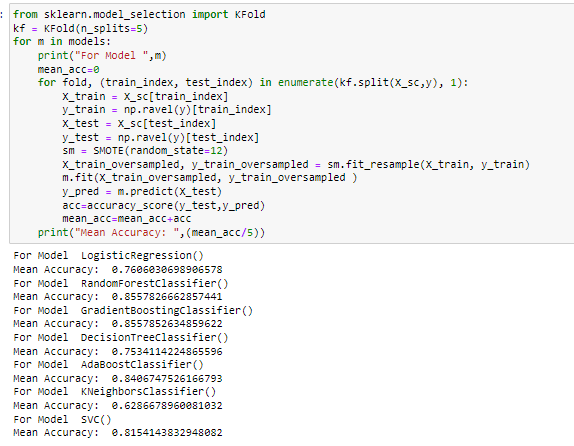


The above ROC AUC plot shows that Logistic Regression has high AUC Score followed by GradientDecentBoostingClassifier.

Cross Validation Score:

Cross validation score helps us to make sure our model is not overfitting and not underfitting.

So, let’s see cross validation score for each model. Here, for every iteration, I have used SMOTE oversampling technique on the training data to make sure the better performance on the model.



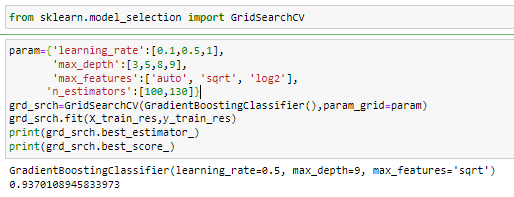
From the above screenshot we can see RandomForestClassifier and GradientBoostingClassifier has the highest cross validation score than other models.

**Concluding Remarks:**

**Based on the base model score, cross validation score and ROC AUC score, GradientBoostingClassifier works better than other models. So, I will conclude the GradientBoostingClassifier as the final model for this project.**

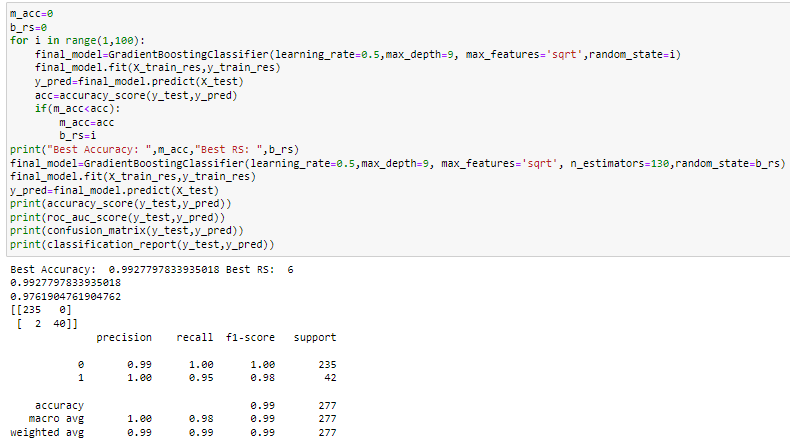
Hyper Parameter tuning:

Hyper parameter tuning will help us to increase the accuracy of the model by tuning different parameters of the machine learning algorithm.I am using GridSerachCV ensemble model to tune the final model with different parameters.

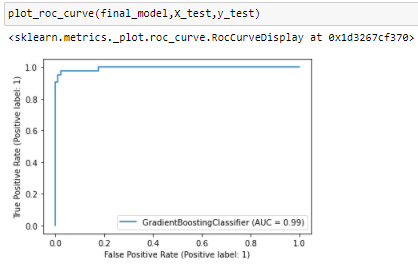


As we see in the screenshot, we able to increase the accuracy up to 93 by tuning few of the hyper parameters (learning\_rate, max\_depth, max\_features) of the GradientBoostingClassifier algorithm.

We still be able to increase the accuracy by finding out the best random state for the final model as below.



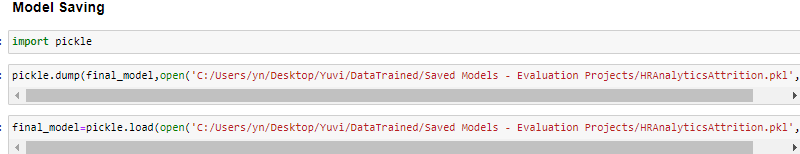
**As we see in the above screenshot, by finding the best random state and applying the best hyper parameter values found, we able to get the model’s test accuracy up to 99%.**



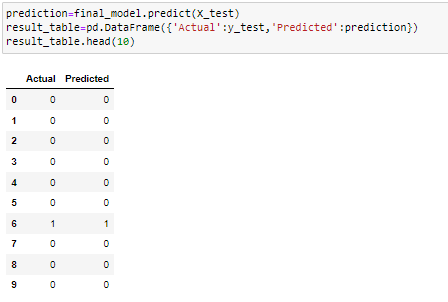
**Above is the AUC score for our final model, it can cover 99% of area under the curve.**

Model Saving:

We successfully build our final model with 99% of test accuracy and 99% of AUC score. Let us save the final model using the pickle to reuse the model for later for the predictions.



Final prediction of our test data:



As we see in the above screenshot, our final model can predict the test data with very good accuracy.