Blog on HR Analytics

## 1. Introduction

Employee attrition (turnover) is a major cost to an organization, and in many organizations the predicting turnover is at the frontend of needs of Human Resources (HR). In machine learning (ML), we can get both better predictive performance and better explanations of what critical features are linked to employee attrition.

Attrition in industry can mean the reduction in staff and employees in a company through normal means, such as retirement means loss of customers or clients and resignation means growing out of the company target. Another reason of attrition is when a company removes or eliminates a person completely. When person think about own investment in recruiting and training the employees and only having them stay on for a short period of time, people are not getting back a return on own investment. Consider what’s the reason of lost when a productive employee quits: new project or product ideas, good project management, or customer or client’s relationships. Customer or company attrition generally gives a negative effect on the company’s profits and the growth.

## 2. Literature Review

The main objective of this dataset is to predict if an employee is going to resign or not that means if the Attrition is Yes then person wants to resign otherwise Person don’t want to resign. The goal of this dataset is reduction in attrition using the analysis of data collected. A good idea or knowledge about the reasons are concerning the attrition of employees that helps the human resource team eliminates the skepticism. It majorly helps in cutting down the costs of person, that a company checks when its employees resign. It’s rightly said that “It takes a lot of resources to build the Human Resource”. The output variable “Attrition” is a target variable with values “Yes/No”. The analysis of data uncovers the factors that leads to the employee attrition.

Methodology:

1. We shall be looking at all features or variables through some criteria and infer about it in the exploratory analysis.

2. Through the analysis we intend to build a model which can predict if an employee is about to resign or not.

3. After the exploration we shall build some features based on the variables at hand and take a call on inclusion/exclusion of few variables.

We will see the implementation of Classification model and logistic Regression which is part of a larger class of algorithms known as Generalized Linear Model.

3. Data Description

We collected data from IBM Watson HR Analytics website (<https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/>). This is a fictional data set created by IBM data scientists. The original dataset is in csv form. The dataset consists of 1470 employees (rows) and 35 features (columns) portion of which have along with Attrition target variable left the organization (Attrition = “Yes”). The data set contains data like age, gender, job satisfaction, Education, Department, Monthly Income, environment satisfaction, job role, income, overtime, percentage salary hike, tenure, training time, years in current role, relationship status, and more. Some attributes such as EmployeeNumber, EmployeeCount, Over18 and StandardHours, being same for each employee are not related to our data.

Therefore among the remaining relevant attributes, there are 16 categorical variables in the dataset i.e. **(Attrition,BusinessTravel, Department, Education, EducationField, EnvironmentSatisfaction, Gender,JobInvolvement, JobLevel, JobRole, JobSatisfaction, ‘’ MaritalStatus, OverTime, PerformanceRating, RelationshipSatisfaction, WorkLifeBalance);**

and 11 numeric variables namely **(Age, DistanceFromHome, MonthlyIncome, NumCompaniesWorked, PercentSalaryHike, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, ‘’ YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager)**. Some categorical variables are coded with numeric values. Such variables are called dummy variables. For example

**Education** 1 ‘Below College’ 2 ‘College’ 3 ‘Bachelor’ 4 ‘Master’ 5 ‘Doctor’

**EnvironmentSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’

**JobInvolvement** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’

**JobSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’

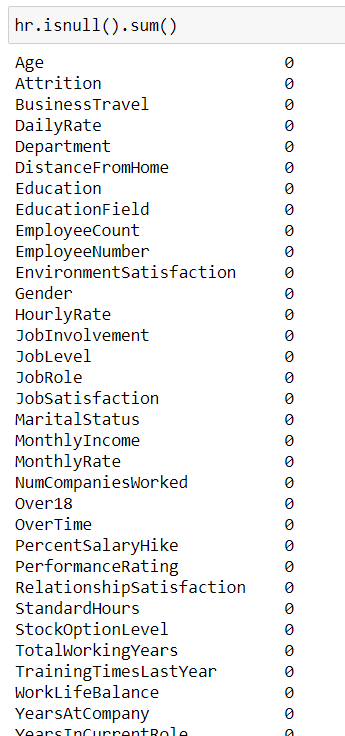
**PerformanceRating** 1 ‘Low’ 2 ‘Good’ 3 ‘Excellent’ 4 ‘Outstanding’

**RelationshipSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’

**WorkLifeBalance** 1 ‘Bad’ 2 ‘Good’ 3 ‘Better’ 4 ‘Best’

4. Data Analysis

1. Identify Features with NULL values: There is no missing values



2. Exploratory Data Analysis - Categorical Features: Converting Categorical to Numeric i.e. Attribute: Yes – 1 and No – 0

Some categorical variables are coded with numeric values. For example

**Education** 1 ‘Below College’ 2 ‘College’ 3 ‘Bachelor’ 4 ‘Master’ 5 ‘Doctor’

**EnvironmentSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’

**JobInvolvement** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’

**JobSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’

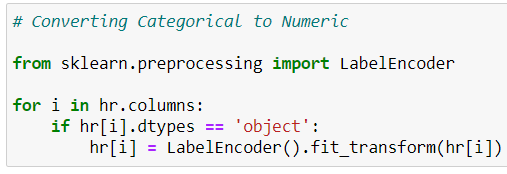
**PerformanceRating** 1 ‘Low’ 2 ‘Good’ 3 ‘Excellent’ 4 ‘Outstanding’

**RelationshipSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’

**WorkLifeBalance** 1 ‘Bad’ 2 ‘Good’ 3 ‘Better’ 4 ‘Best’



Converting the above categorical to Numeric



5. EDA Concluding Remark

1. Converted all the categorical data to the numeric data.

2. Check the correlation between the features or columns in the dataset

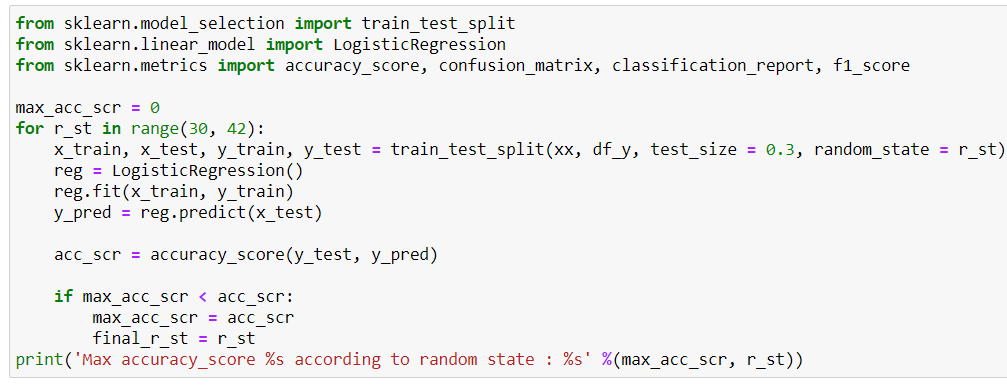
3. Removed the Outliers with high skewness

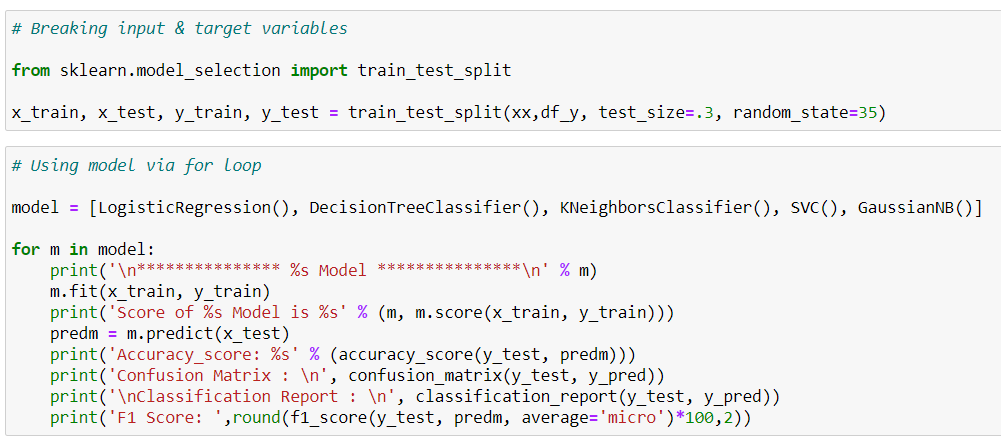
6. Model Building and Validation

We are going to use the classification models for training & predicting Employee Attrition. We need to check that we should be to have a better accuracy of predicting attrition i.e. Attrition = 1 which in confusion matrix will be "True Positive". However, we should not forget the prediction accuracy of not qualifying for attrition i.e. Attrition = 0 which will be "True Negative" in confusion matrix.

So, we need to focus on four parameters:

* **Accuracy:** Overall, how often is the classifier correct? i.e {(TP+TN)/Total}
* **True Positive Rate:** When it's yes, how often does it predict yes? default\_ind = 1, {TP/Actual YES}, this is also known as "Sensitivity" or "Recall"
* **Precision:** When it predicts yes, how often is it correct? i.e. {TP/(TP+FP)}
* **Specificity:** When it's actually no, how often does it predict no? default\_ind = 0, {TN/actual NO}
* **Cross Validation Score:** Cross Validation is a technique which involves reserving a particular sample of a dataset on which you do not train the model. Later, you test your model on this sample before finalizing it. Do this for k folds and take mean of accuracy scores of the k fold models.
* **F1 Score:** This is a weighted average of the true positive rate (recall) and precision.
* **ROC Curve:** This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class.





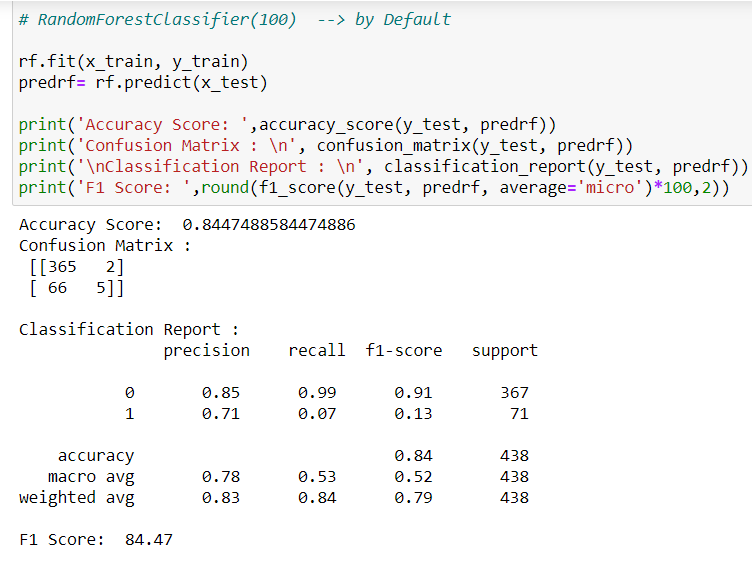
'Precision'-- (true positives)/(true positives+false positives)

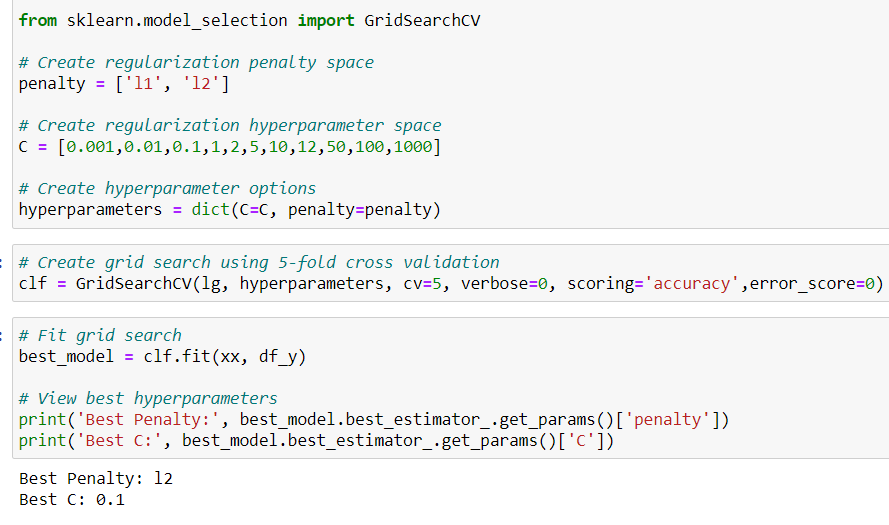
'Recall'-- (true positives)/(true positives+false negatives)

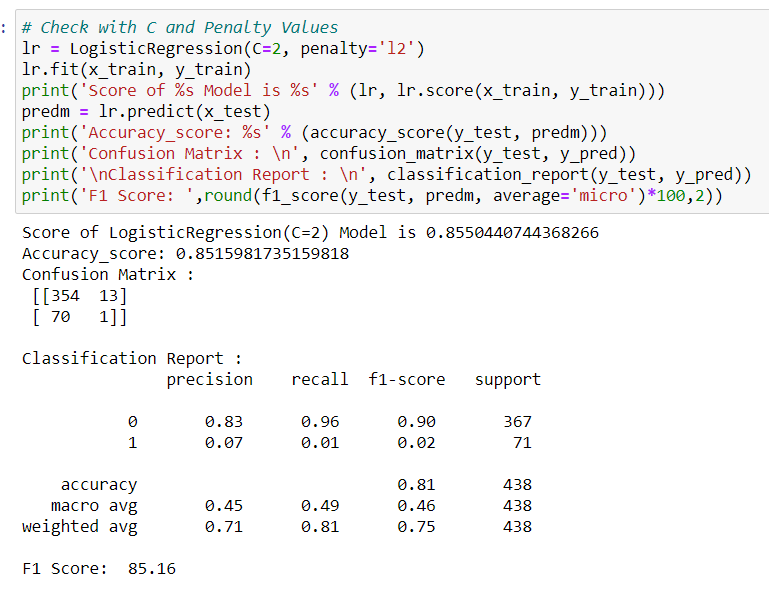
'F1 Score'-- The harmonic mean of 'precision' and 'recall'

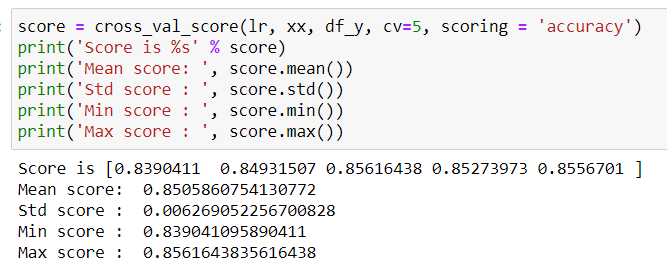
'AUC ROC'-- ROC curve is a plot between 'senstivity' (Recall) and '1-specificity' (Specificity=Precision)

'Confusion Matrix'-- Plot the entire confusion matrix









7. Conclusion:

You can see that there is approx 1.5% Accuracy difference when we use the Ensemble Method, however, the F1 Score, which is the trade Off which we should be looking seems more desirable for each class!

We tried a Grid Search then using the **Logistic Regression Method**, to find the best Hyperparameters! Here, our Hyperparameter Tuning was more focused on Getting a Better Macro **F1 Score! -> 85% overall.**