HR Analytic Project- Understanding the Attrition in HR



***Submitted By***

***TRIBENI SUBUDHI***

**INTRODUCTION:**

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.

**HR Analytic:**

Human resource analytic (HR analytic) is an area in the field of analytic 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 analytic 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 Analytic 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:**

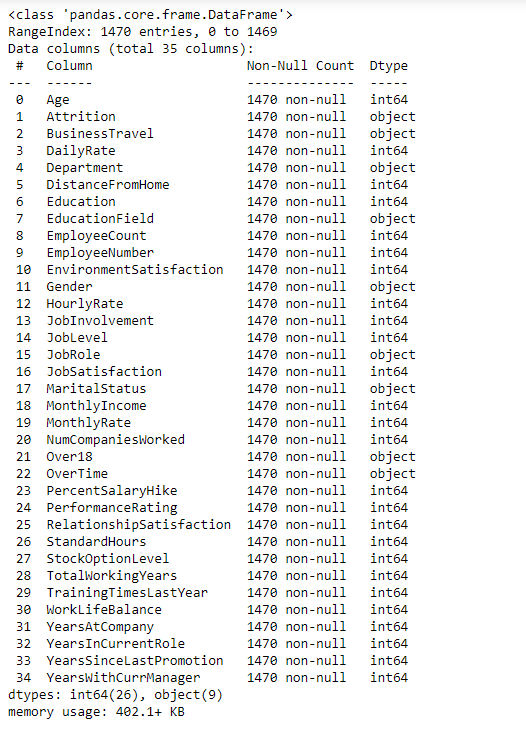
A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits the organization from increasing its collective knowledge base and experience over time.

This is especially concerning if the business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if organization constantly have new workers.

**Data Analysis:**

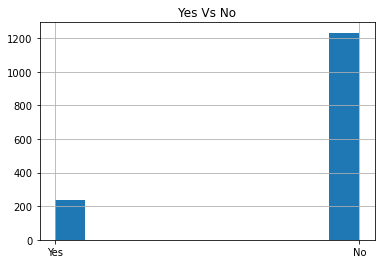
We have been given with a dataset which contains the details of the attrition along with the employee details.

The provided dataset contains 1,470 rows and 35 columns including the target column. Some of the columns name in the dataset are Age, Attrition, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, etc.

 As per our observations, we found that our dataset does not contain any missing or null values.

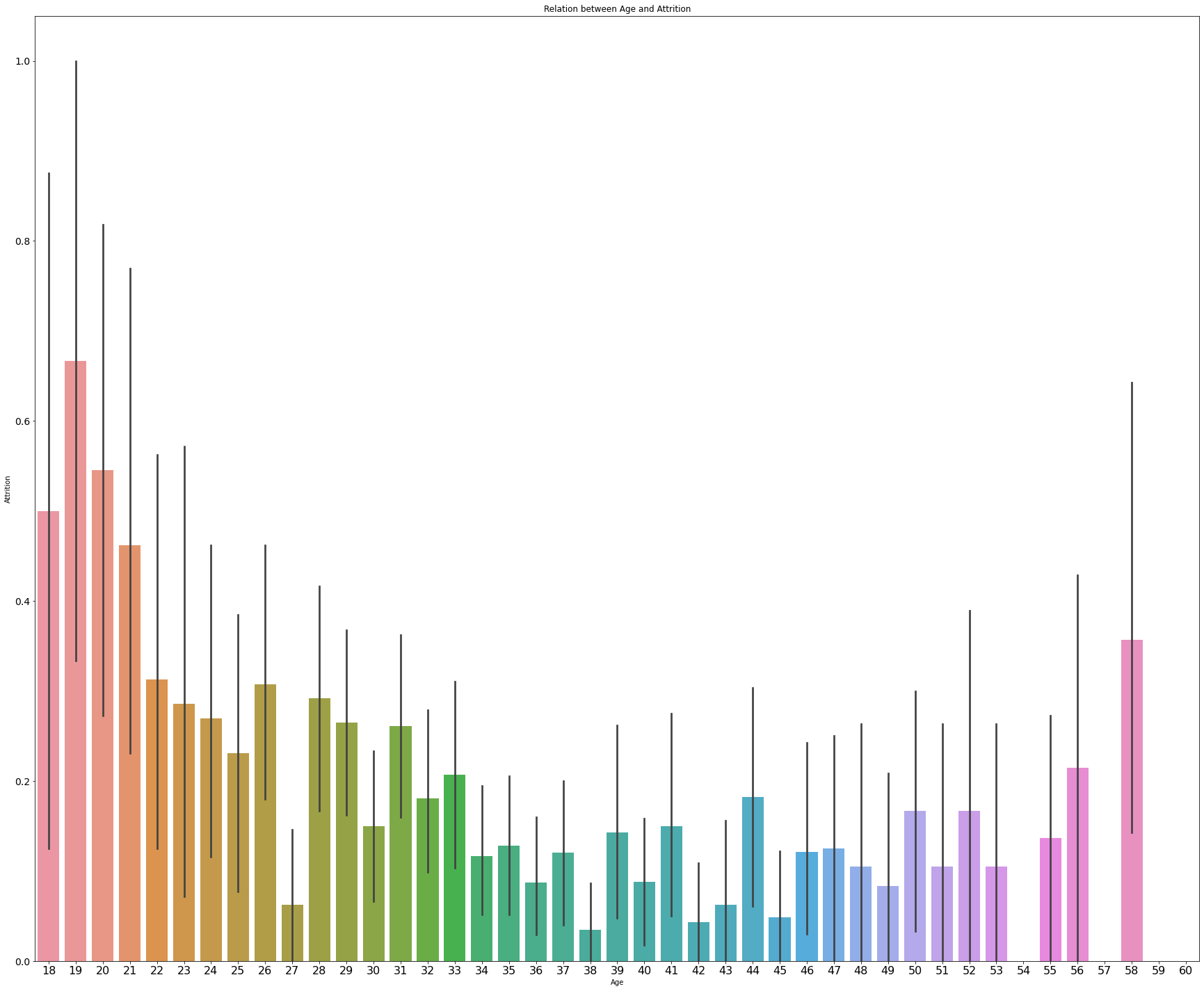
**Exploratory Data Analysis (EDA) :**

* **Target Column (Dependent variable):**



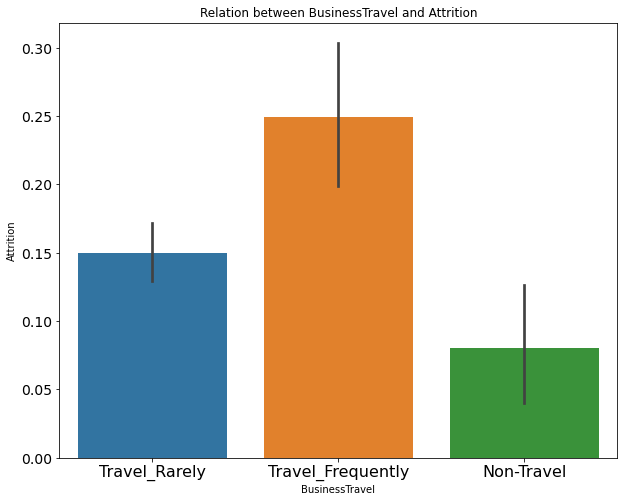
We found that the data is imbalance in the target column. We’ll work on this in further steps.

* **Effect of age on Attrition:**



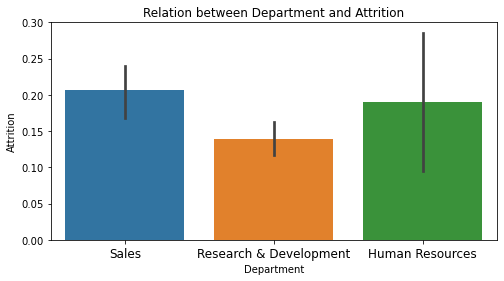
From the graph we can make the following observations:

* The rate of attrition is high among the age group of 18 to 21.
* The rate of attrition is low among the age group of 34 to 53.
* **Let’s check if business travel is affecting HR attrition:**



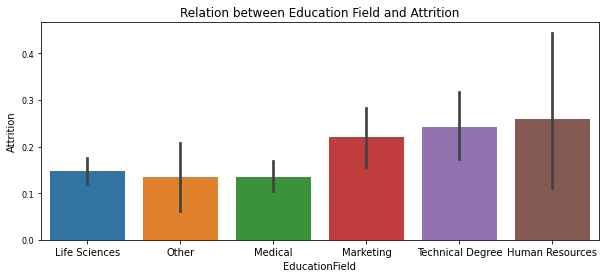
We can make following observations from the above graph:

* The rate of attrition is high among the employee who travel frequently for business.
* The rate of attrition is low among the employee who do not travel for business.
* **Let’s check if the department of the employee have any effect on the HR attrition:**

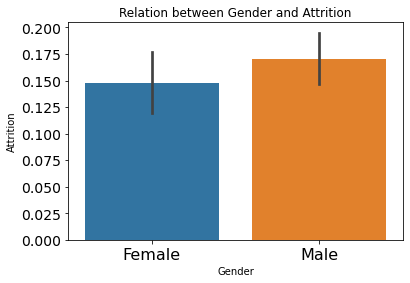


* The rate of attrition is low among the employee of Research and Development.
* The rate of attrition is almost same among the employee of Sales and HR department.

**Let’s check the effect of Education field of the employee on the Attrition:**

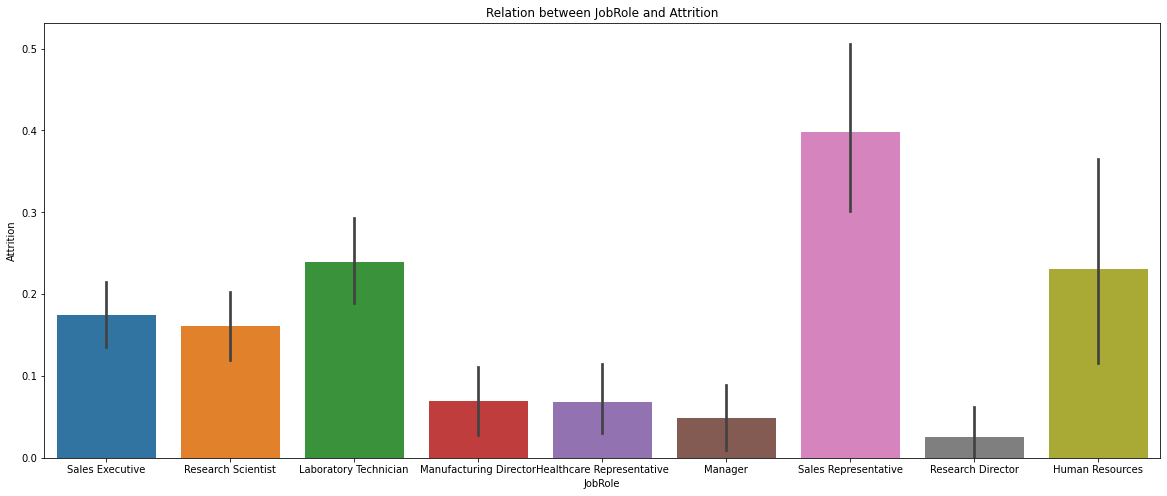
****

* The rate of attrition is high among the employee with the education field HR, technical and Marketing respectively.
* The rate of attrition is low among the employee with the education field Other, Medical and Life Sciences.
* **Let’s check for the rate of attrition gender-wise:**

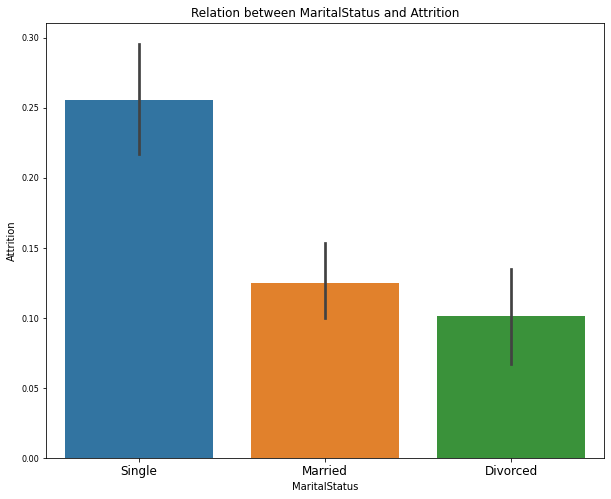


The observations shows that rate of attrition is high among the males as compared to the females.

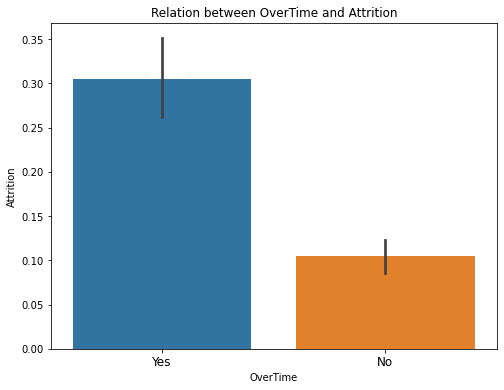
**Let’s check for the attrition job role wise:**

****

* The graph shows that the rate of attrition is high among the sales Representatives.
* The employees having the Job role Research Director has least chance of attrition.
* **Let’s check the effect of marital status of the employee on the attrition:**



* Unmarried employees have the maximum chance of attrition.
* Divorced employees have the minimum chance of attrition.
* **Does overtime make employee to think for attrition?**



From the graph it is clear that employees do not like to work overtime. So, the organization should look for this to avoid the chances of HR attrition.

**Data Pre-processing:**

Data pre-processing is very important step in Machine Learning to get highly accurate and insightful results. The greater quality of data will help us to make highly efficient model. This will lead to the highly reliability of the predicted results.

**Incomplete, Noisy, and Inconsistent data** will lead to the inherent nature of real-world problems. Data pre-processing help us in increasing the quality of data by treating the missing values, removing the noises, and resolving inconsistency.

* **Incomplete Data:** There can be a number of reasons for this. There may be some misunderstanding or technical fault which lead to the incomplete data.
* **Noisy Data:** This can be the incorrect feature values present in the dataset. There can be many reasons for this. The instrument used for the data collection might be faulty. There may be human error at the time of data entry.

**Stages of Data Pre-processing:**

* **Data Cleaning:** In this stage we work on the dataset to remove the unwanted feature columns (columns which are not contributing to the dataset), treating the missing or null values, removing the outliers and skewness if present.
* **Data integration:** integrates data from a multitude if source into single data warehouse.
* **Data transformation:** such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurement.
* **Data reduction:** Reduce the data size by dropping out redundant features. Feature selection and feature extraction technique can be used.

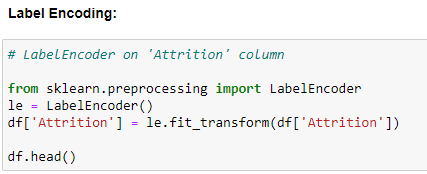
**Label Encoding: (Converting the categorical features to numerical)**

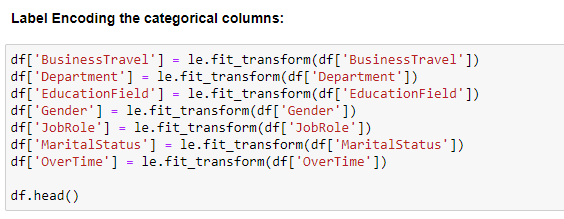
In dataset there can be some columns which will be containing the categorical columns (columns with the datatype of object). For machine learning model we need to convert the categorical data to numerical. We can use label encoding or one hot encoding methods for the conversion.

In our data there are columns with categorical values. The categorical columns are Attrition, Business Travel, Department, Education Field, Gender, Job Role, Marital Status, Over18, and

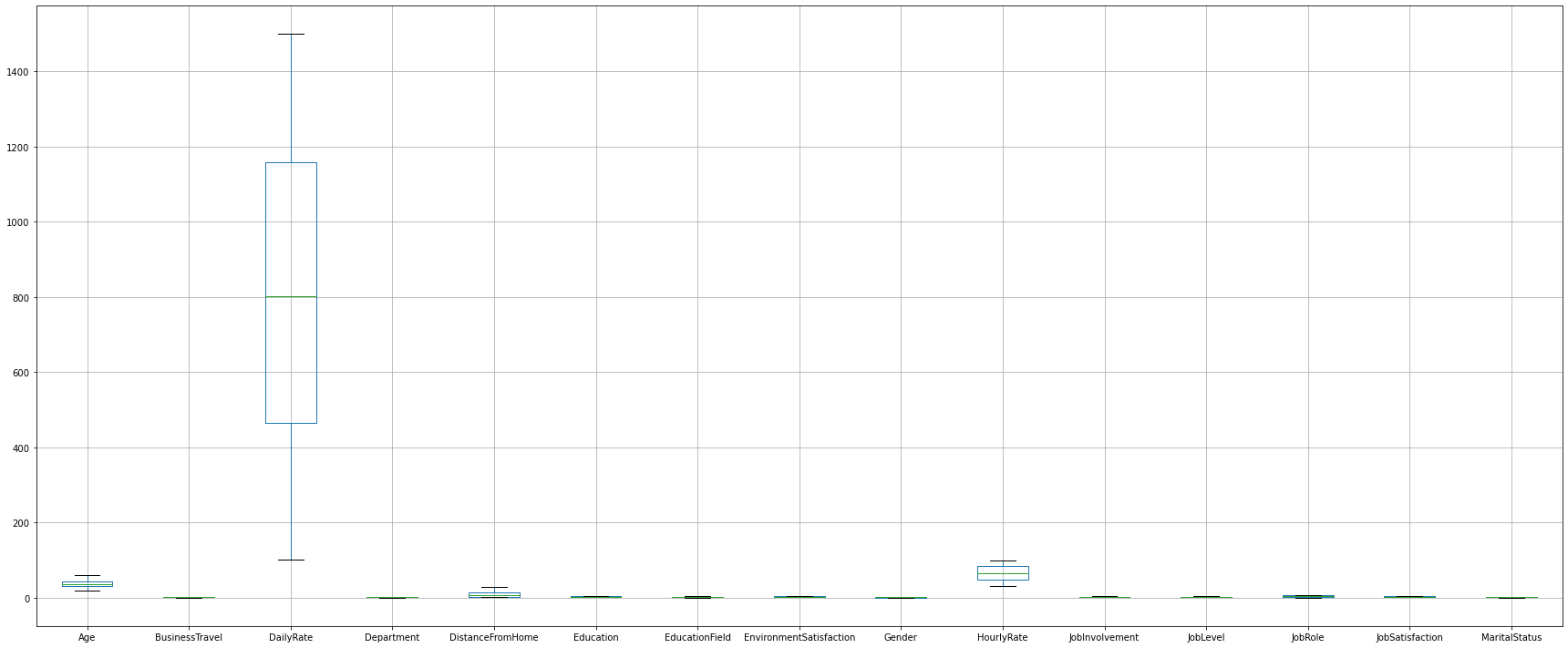
Overtime. We are treating these columns with the label encoding method.

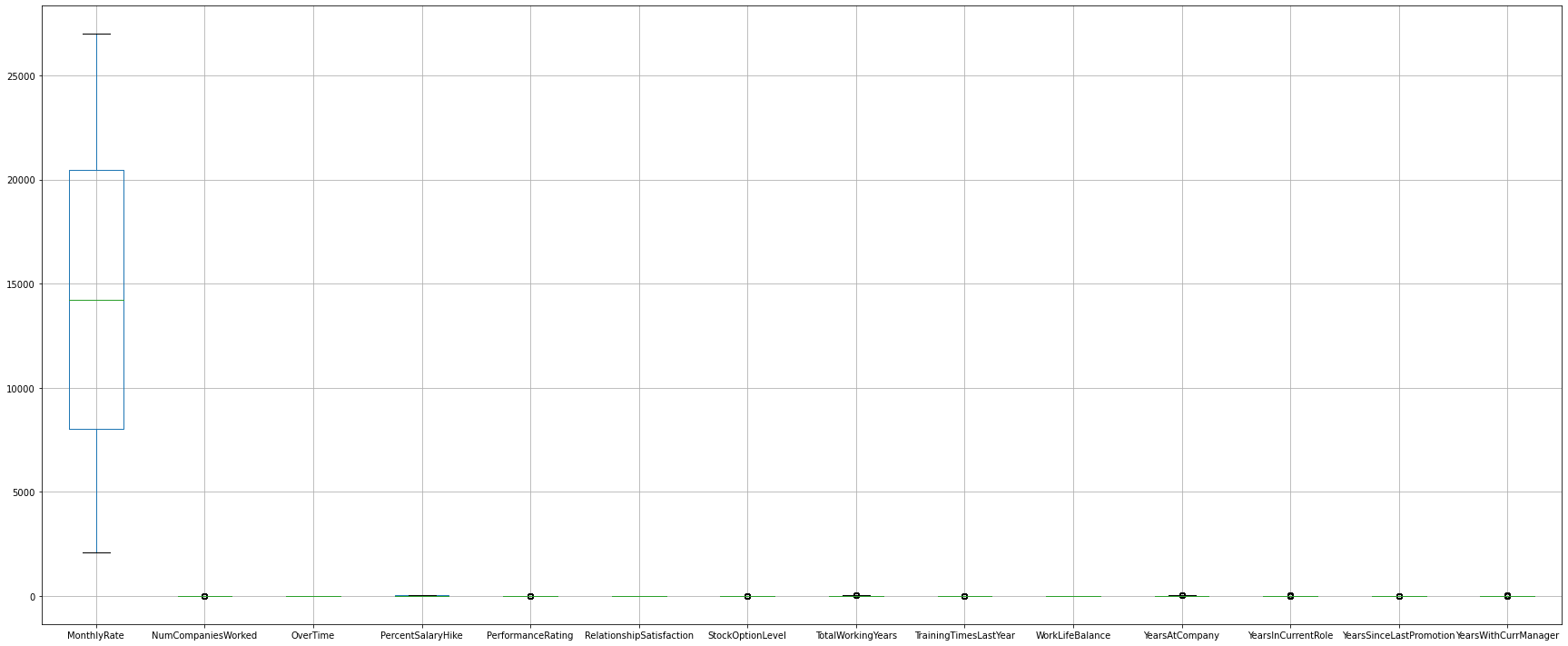
We can find the Label Encoding in skLearn library, so we can import it. skLearn provides very efficient tool for encoding. Label encoder encode labels with a value between 0 and n\_classes-1. Let’s see an example from our dataset:



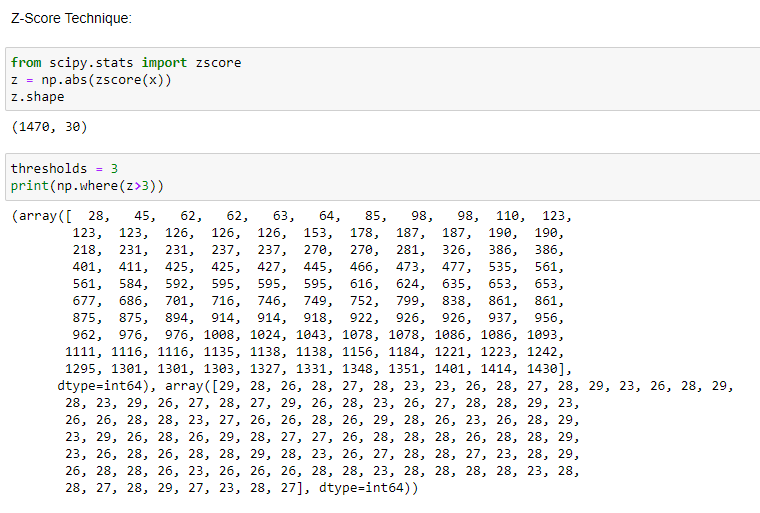


**Outliers:** We can define outliers as the data point which is distant from the other similar points. They may be due to variability in the measurement or may indicate experimental errors. If possible, outliers should be excluded from the data set. However, detecting that anomalous instance might be very difficult, and is not always possible.





We have removed the outliers using z-score method.



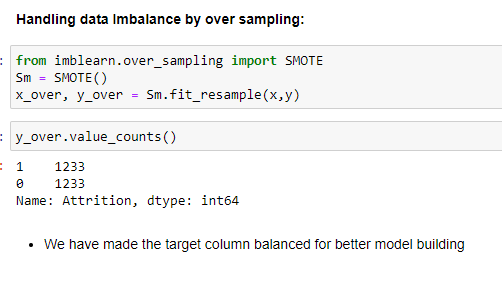
**Balancing the imbalanced dataset:**

We have seen earlier that; our dataset was imbalance. There are different methods available to make the data balanced. We have used SMOTE method.

**SMOTE** (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors of this point. The synthetic points are added between the chosen point and its neighbors.

SMOTE algorithm works in 4 simple steps: -

1. Choose a minority class as input vector.
2. Find its k-nearest neighbors.
3. Choose one of these neighbors and place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbors.
4. Repeat the step until the data is balanced.



The SMOTE algorithm has made the target column Attrition to 1233 (the highest value) for both the values (‘1’ for YES and ‘0’ for NO).

**Machine Learning Model Building:**

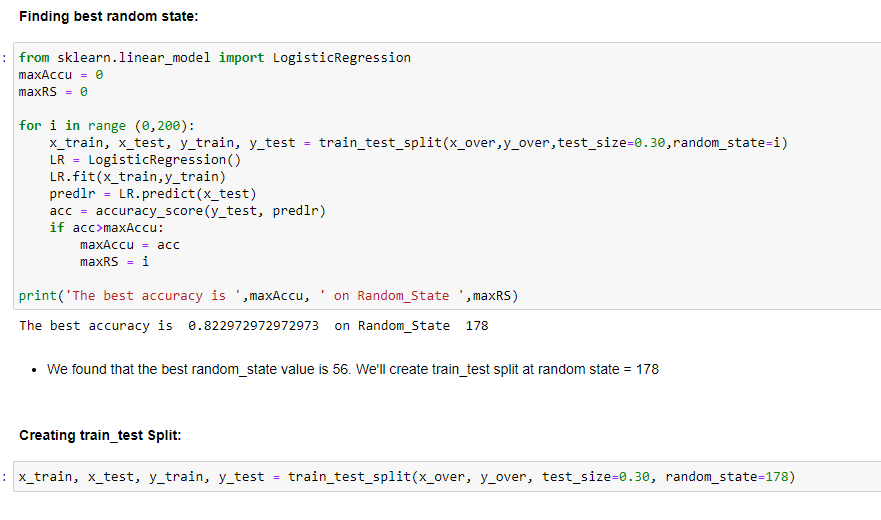
We can find many types of methods for model building in the skLearn library.

There are two types of models which are present in the skLearn library: 1. Regression and 2. Classification

In the case of our dataset, we have to make machine learning model to predict whether the claim made by the customers are fraud or not. Since we have only two values, so we will use classification models to build the machine learning model.

Before fitting the data to the model, we will need to first separate the dependent variable and independent variables, then we will pass these values to train\_test\_split to generate random training and testing subset of the data.

**About train\_test\_split method:** It is a function, which we can find in skLearn model selection library. We use this function for splitting data arrays into two subsets for training data and testing data. By using this function, we can avoid the manual task to separate the dataset. By default, sklearn train\_test\_split will make random partitions for the two subsets. However, we can also specify a random state for the operation. It gives four output x\_train, x\_test, y\_train and y\_test. The x\_train and x\_test contains the training and testing predictor variables while y\_train and y\_test contains the training and testing target variable.



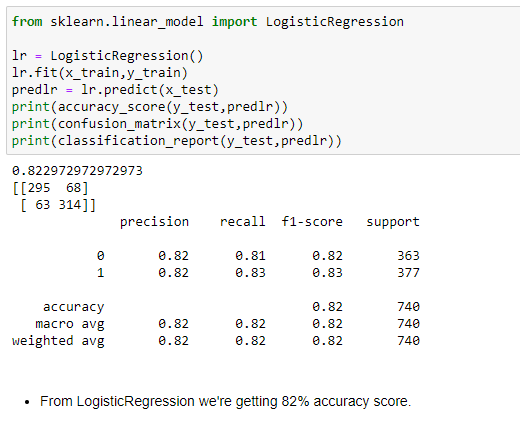
After performing train\_test\_split we have to choose the models to pass the training variable.

We can build as many models as we want to compare the accuracy given by these models and to select best model among them.

We have made 4 models to get the closest possible accuracy score to predict the fraud claims.

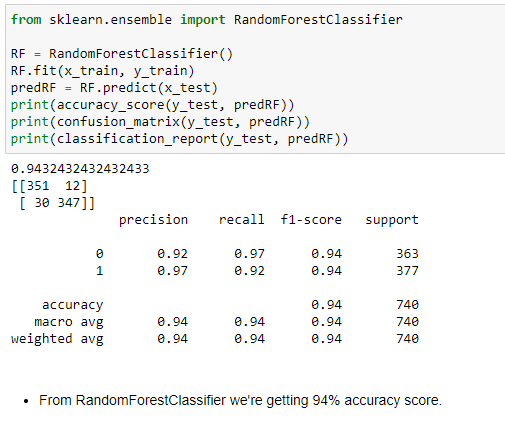
1. **Logistic Regression:**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is binary, which means there would be only two possible classes 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.



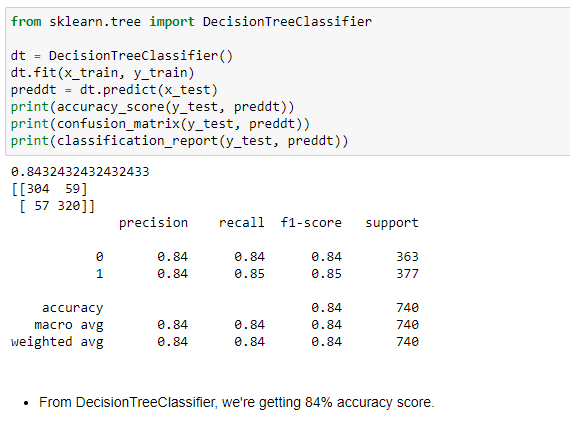
1. **Random Forest Classifier:**

As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.



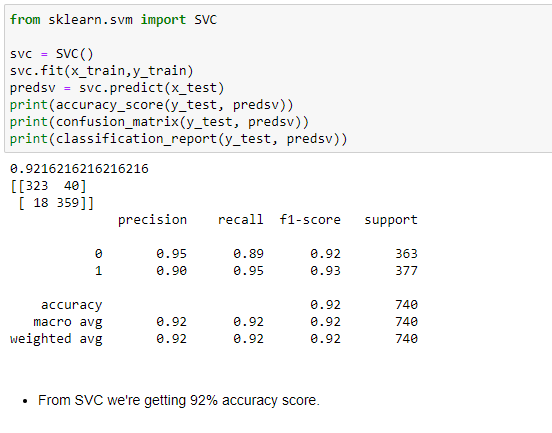
1. **Decision Tree Classifier:**

Decision Tree Classifier is a class capable of performing multi-class classification on a dataset. As with other classifiers, [DecisionTreeClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html" \l "sklearn.tree.DecisionTreeClassifier" \o "sklearn.tree.DecisionTreeClassifier) takes as input two arrays: an array X, sparse or dense, of shape (n\_samples, n\_ features) holding the training samples, and an array Y of integer values, shape (n\_samples), holding the class labels for the training samples. It is capable of both binary (where the labels are [-1, 1]) classification and multi-class (where the labels are [0, …, K-1]) classification.



1. **SVC: (Support Vector Classifier)**

SVC is a non-parametric clustering algorithm that does not make any assumption on the number or shape of the clusters in the data. In SVC data points are mapped from data space to a high dimensional feature space using a kernel function. In the kernel's feature space the [algorithm](http://www.scholarpedia.org/article/Algorithm) searches for the smallest sphere that encloses the image of the data using the [Support Vector Domain Description](http://www.scholarpedia.org/w/index.php?title=Support_Vector_Domain_Description&action=edit&redlink=1) algorithm. This sphere, when mapped back to data space, forms a set of contours which enclose the data points. Those contours are then interpreted as cluster boundaries, and points enclosed by each contour are associated by SVC to the same cluster.



If we compare all the models, we found that we are getting maximum accuracy score from Random Forest Classifier. But we cannot decide the best model on the basis of accuracy score only, since this might be possible that our data is over-fitted.

So, to decide the best fit model we will check for the cross-validation score of each model. We can import cross\_val\_score from the sklearn library (sklearn. model selection).



We will first calculate the difference between the accuracy score and cross validation to decide the best fit model for our prediction. The model having the lowest difference between the accuracy score and cross validation score will be the best model for our machine learning algorithm.

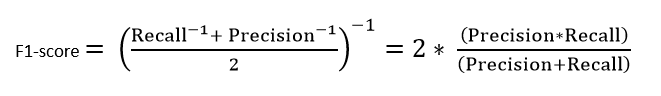
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S. No. | Model Name | Accuracy Score | Cross Validation Score | Difference |
|  |  |  |  |  |
| 1 | Logistic Regression | 82.29 | 86.59 | -4.3 |
| 2 | Random Forest Classifier | 94.32 | 86.05 | 8.27 |
| 3 | Decision Tree Classifier | 84.32 | 79.18 | 5.14 |
| 4 | SVC | 92.16 | 86.32 | 5.84 |

From the above table we can say that the best fit model is Logistic Regression since it is showing minimum difference between the accuracy score and cross validation score.

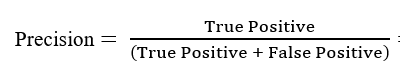
* Our model is showing approx. 82% accuracy score.
* It has predicted 295 true positive cases out of 363 positive cases and 314 true negative cases out of 377 cases.
* It has predicted 68 false positive cases out of 363 positive cases and 63 false negative cases out of 377 cases.
* It has given the f1 score of approx. 82%.

Let’s understand what does precision recall and f1 score and accuracy means.

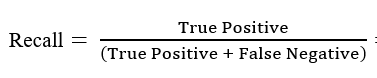
* **F1 score:** It is the harmonic mean of precision and recall and gives a better measure of the incorrectly classified cases than the accuracy matrix.



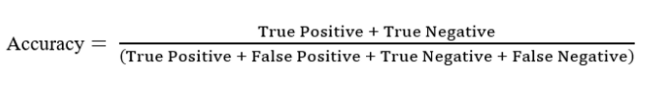
* **Precision:** It is implied as the measure of the correctly identified positive cases from all the predicted positive cases. Thus, it is useful when the costs of False Positives are high.



* **Recall:** It is the measure of the correctly identified positive cases from all the actual positive cases. It is important when the cost of False Negatives is high.

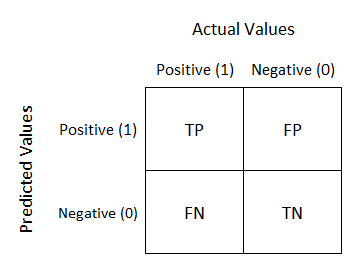


**Accuracy:** One of the more obvious metrics, it is the measure of all the correctly identified cases. It is most used when all the classes are equally important.



**Let’s understand what confusion matrix is:**

A confusion matrix is a table that is often used to describe the performance of a classification model (or ‘classifier’) on a set of test data for which the true values are known.



* **TN/True Negative:** the cases were negative and predicted negative.
* **TP/True Positive:** the cases were positive and predicted positive.
* **FN/False Negative:** the cases were positive but predicted negative.
* **TN/True Negative:** the cases were negative but predicted positive.

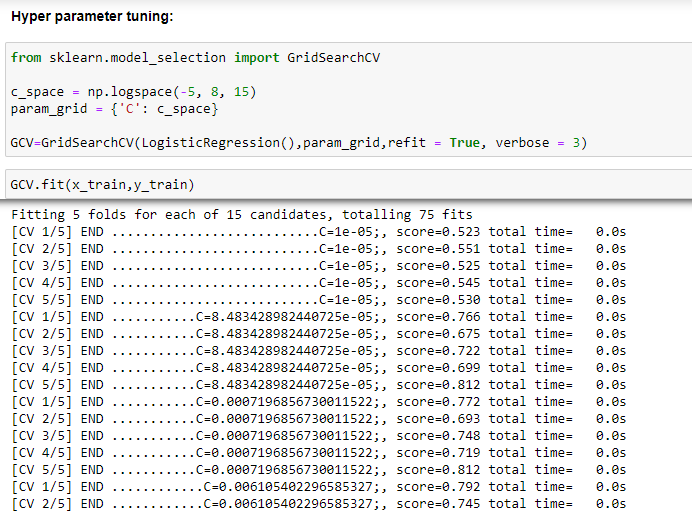
**Hyper parameter tuning:**

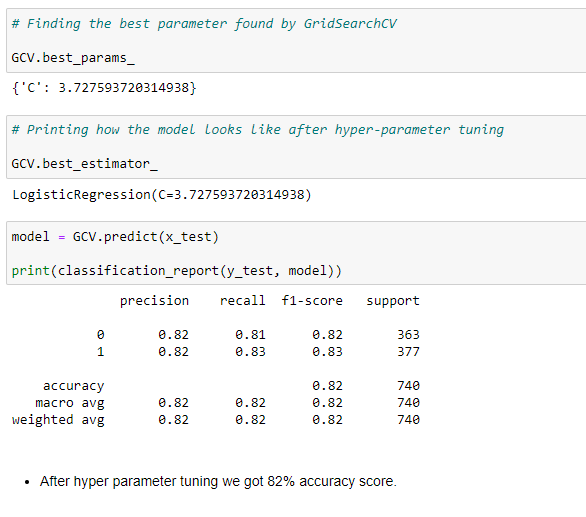
Hyper parameter optimization in machine learning intends to find the hyper parameters of a given machine learning algorithm that deliver the best performance as measured on a validation set. Hyper parameters, in contrast to model parameters, are set by the machine learning engineer before training. The number of trees in a random forest is a hyper parameter while the weights in a neural network are model parameters learned during training. I like to think of hyper parameters as the model settings to be tuned so that the model can optimally solve the machine learning problem.

We will use **GridSearchCV** for the hyper parameter tuning.

**GridSeatchCV:**

In GridSearchCV approach, machine learning model is evaluated for a range of hyper parameter values. This approach is called GridSearchCV, because it searches for best set of hyper parameters from a grid of hyper parameters values.



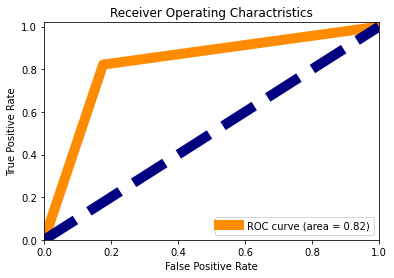
 We can see that after hyper parameter tuning our model accuracy score has been increased to approx. 82%.

**AUC ROC Curve**

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

We are plotting AUC ROC curve for the final model.



**Final Remarks:**

We have successfully built a model to predict the HR Attrition. Our observations and finding can help organizations to avoid attrition and retain the employees for longer period. This will benefit the organization to utilize their assets in better way.

Companies can avoid major problem in high employee attrition is its cost to an organization using the Machine Learning algorithm.

***References:***

[*https:/**/scikit-learn.org/*](https://scikit-learn.org/)

[*https://towardsdatascience.com/*](https://towardsdatascience.com/)