**EMPLOYEE ATTRITION RATE ANALYSIS – INSIGHTS FROM IBM HR DATA**

**Introduction:**

**What is an Attrition?**

Attrition is defined as a voluntary and involuntary reduction of a company’s workforce. This reduction is due to retirements, transfers, resignations, terminations, or deaths.

Employees are the backbone of the organization. Organization's performance is heavily based on the quality of the employees. Challenges that an organization has to face due employee attrition are:

1. Expensive in terms of both money and time to train new employees.
2. Loss of experienced employees
3. Impact in productivity
4. Impact profit

**Types of Attrition:**

There are 4 types of attrition which are as follows:

**Voluntary Attrition**

Voluntary Attrition means employees willingly leaving the company due to personal or professional reasons. It is one of the most common types of attrition. Voluntary attrition includes leaving jobs for better prospects, retirement, or relocation.

**Involuntary Attrition**

Involuntary Attrition is when an employer initiates the termination of employees for misconduct, merger, or acquisitions. Often structural changes or production line changes also initiate involuntary attrition. Even economic slowdown is also a cause of involuntary attrition.

**Internal Attrition**

The movement of employees from one department to another refers to internal attrition. This includes moving of employees within the company for growth. It includes moving of employees to higher designations or other departments that fit their talent.

**Demographic Attrition**

Demographic Attrition refers to the loss of a specific group of employees.  This includes specific age, sex, ethnic minorities, people with disabilities, veterans, or older professionals. The cause of such attrition is due to bad work culture or structural changes in the organization.

**Attrition Rate:**

The attrition rate means calculating the proportion of employees leaving an organization over a specific period. A normal rate of attrition is expected in normal business operations. But a high rate of attrition leads to many problems and a lack of workforce. HR professionals design and implement company compensation programs and motivation systems. This helps them to keep the employees happy and attrition rates low. Keeping the attrition rates as low as possible helps to save money.

**Problem Definition:**

This study is based on a fictional dataset from IBM HR Analytics Employee Attrition & Performance. It has 1470 data points and 35 features describing each employee’s background and characteristics; and labelled with whether they are still in the company or whether they have gone to work somewhere else.

Machine Learning models can help to understand and determine how these factors relate to workforce attrition.

This article provides in-depth analysis as well as predictive modelling to understand important factors and make accurate predictions.

Table of Content:

Data Preparation and Understanding.

Feature Selection.

Model Fitting

Model Comparison

Recommendations & Conclusion.

**Data Preparation and Understanding:**

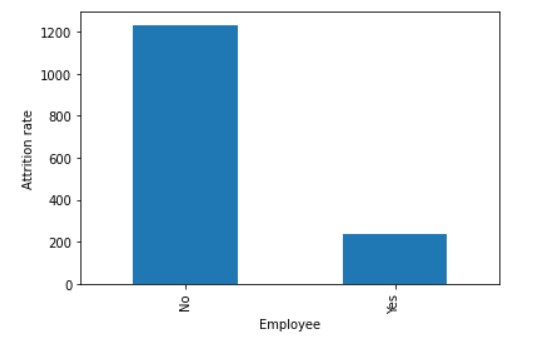
The [dataset](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)that is published by the Human Resource department of IBM is made available at Kaggle.

IBM has gathered information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees.

To understand the data, we will do some analysis like shape of data, dtypes to determine what kind of data is present in data set, whether the null values are present or not and statistical analysis as well for more clarity.

**Target Variable: Attrition**

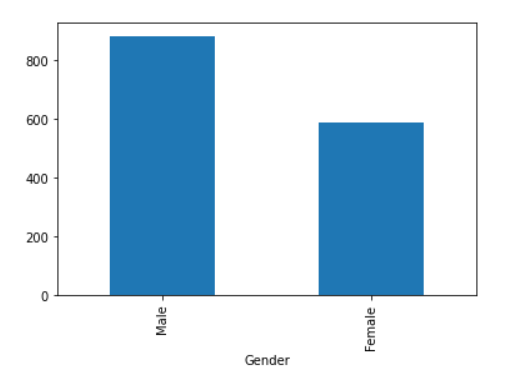
The Attrition dataset had 1470 observations with 35 variables. Out of the 35 variables, there exists one target variable Attritionwith possible outcomes ***Yes***and ***No***. The other 34 variables are independent variables but one, that was, *Employee Number*which denotes the employee number or the identification number.



Above graph shows the distribution of target variable. Among 1470 observations 1233 is No whereas 237 is Yes. In dataset there is no null values.

**Gender vs Attrition**

Attrition rate of male employees is more than that of the female employee. Below graph shows the comparison of attrition rate between male and female employees.



**Outliers:**

Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyRate, PercentSalaryHike tend not to have any outliers.

NumCompaniesWorked, TrainingTimesLastYear, YearsWithCurrManager, YearsInCurrentRole have a moderate number of outliers.

MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion have large number of outliers.

One way to counter this problem is by scaling the variables so as to reduce its effect on the model. The standardization technique StandardScaler().

**Education vs Attrition**

Education is most valuable aspect for attrition. For better understanding we will distribute the education of employees in five score ranges as below:

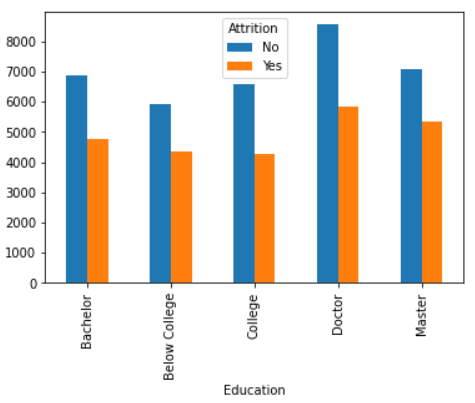
Education: 1 'Below College'

2 'College'

3 'Bachelors'

4 'Master'

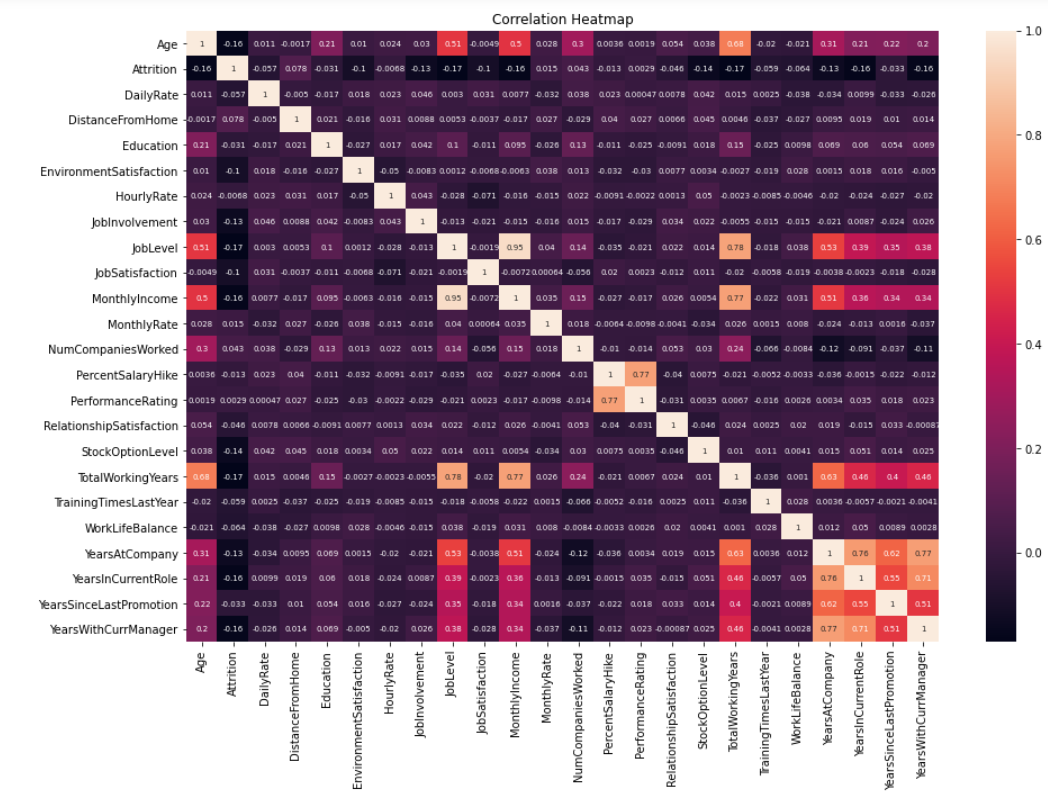
5 'Doctors'



Above graph shows the relationship between education and attrition rate of an employees. We can see that the employees who have doctorate or master’s degree have high attrition rate as compare to another employee.

**Correlation between independent variables:**

Before moving forward, we need to check multicollinearity, so we will plot correlation matrix.



The above graph shows the correlation between independent variables. We perform this to necessarily check for multi collinearity. A rule of thumb that is followed for multi collinearity is if correlation coefficient (r) is close to 0.80. Based on that we identify the following variables to have a high correlation:

Correlation between MonthlyIncome and JobLevel is 0.95. This is a very high correlation.

Correlation between TotalWorkingYears and JobLevel is 0.78 which is also very close to 0.80.

All other variables seem to have a correlation which is less than 0.80.

Having got an understanding of the data along with preliminary analysis, now we will do some feature extraction on our data.

**Feature Selection**

Based on the Exploratory Data Analysis and Descriptive Statistical Analysis, we select and deselect certain variables which do not significant contribute to our model. The variables are deselected based on the following:

1. Variable type
2. Invariability in the Data point
3. Multicollinearity

**1.Variable Type:**

The nominal variable is not used in the analysis as it does not provide any input to the model building process. It is however kept so as to identify the employees on whom the study is done.

**2.Invariability in the Data point:**

Certain variables do not have any variability in them. Such variables are:

Employee Count: This is just a count of employee and the value it takes is always 1.

Over18: This variable describes if an employee is over 18 years of age. It takes the value ‘Yes’ in all cases.

Standard Hours: The standard number of hours an employee works in a week. Its constant value is 80

**3.Multi collinearity:**

Multi collinearity refers to the strong relationship or correlation between to input variables. There is said to be multi collinearity between two variables if there exist a correlation coefficient of more than 0.80. It is important to remove such variables as this leads to an inflated variance in the model which also increases the error in the model.

Based on our analysis, we remove the following variables:

*JobLevel*: 0.95 correlation coefficient with *MonthlyIncome*and 0.78 with *TotalWorkingYears*.

*TotalWorkingYears*: 0.77 correlation with *MonthlyIncome.*

**Model Fitting**

After pre-processing, we split our data into training, and testing dataset. From a total of 1470 observations, we choose:

80% observation for *Training Dataset.*

20% observation for *Test Dataset.*

By using different algorithms, we will compare the accuracy score and select the best performing model for our data.

Here I am using below algorithms for the comparison.

1. Decision Tree
2. Random Forest
3. Support Vector Machine

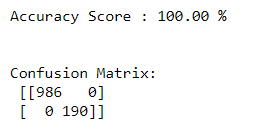
**Decision Tree:**

Decision trees are a popular supervised learning method for a variety of reasons. Benefits of decision trees include that they can be used for both regression and classification, they are easy to interpret and they don’t require feature scaling. They have several flaws including being prone to overfitting.

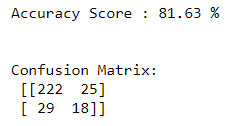
Let’s try it on our dataset for find the result.

The next process is to predict the values based on the fitted model for the train and test set. To calculate the accuracy, precision, recall, true positive and true negative, we create a confusion matrix.

Confusion matrix for Train Data given below:



Confusion matrix for Test Data given below:

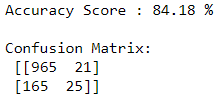


**Support Vector Machine:**

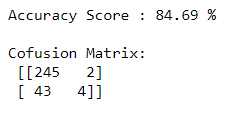
Support vector machines (SVM) is a supervised machine learning technique. And, even though it’s mostly used in classification, it can also be applied to regression problems.

SVMs define a decision boundary along with a maximal margin that separates almost all the points into two classes. While also leaving some room for misclassifications.

Confusion matrix for Train Data given below:



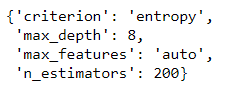
Confusion matrix for Train Data given below:



**Random Forest:**

Random forest is an ensemble machine learning model. An ensemble machine learning model is a model which is a collection of several smaller models. The Random Forest model of machine learning is nothing but a collection of several decision trees. These trees come together to a combined decision to give the output.

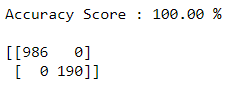
The parameters chosen for Random Forest is as below:



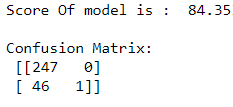
The important parameters such as criterion is chosen to be ‘entropy’ in this case. The max\_depth is taken to be 8 and n\_estimators are 200.

The next process is to predict the values based on the fitted model for the validation and test set. To calculate the accuracy, precision, recall, true positive and true negative, we create a confusion matrix.

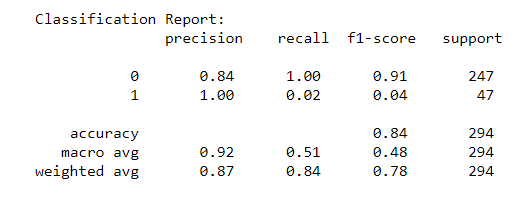
The confusion matrix for Train set is given below:



The confusion matrix for Test set is given below:

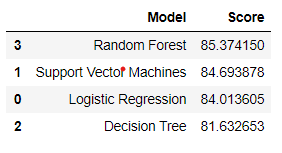


The classification report for random forest is as below:



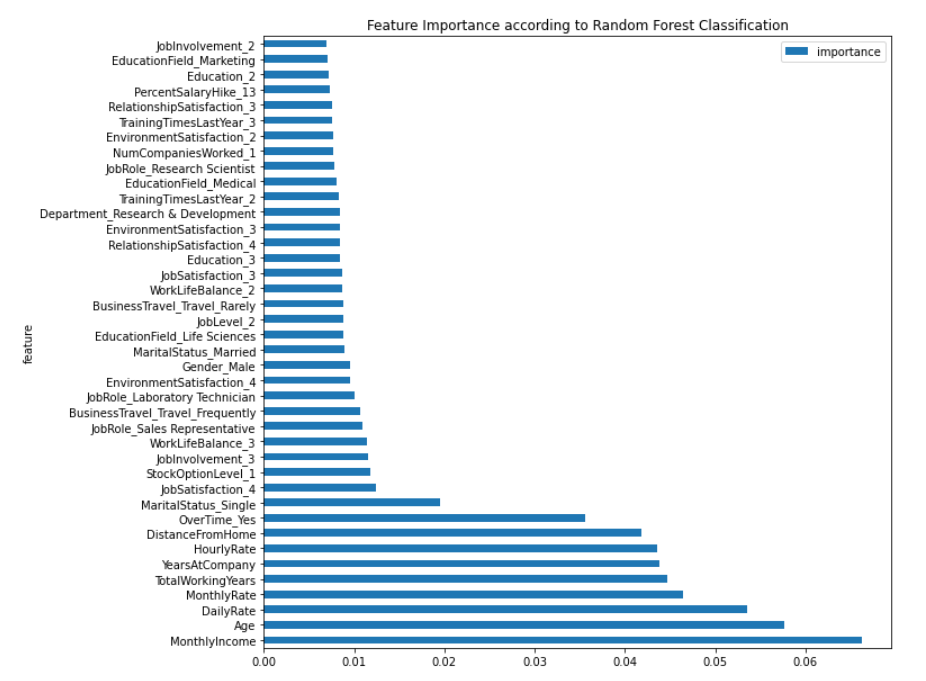
The above set of information pertains to the Test set. We get a weighted accuracy of 0.84 with precision as 0.84 and recall as 1.00.

**Model Comparison:**

****

As mentioned, several classifiers were used and the best one was selected. However, it is worth looking at the Benefits and Weaknesses of each classifier. *Random forest is gives highest accuracy (85.37%)* so we have chosen Random Forest here it’s also helpful with new data points in classification. Over fitting can be a concern with the best model as it might succumb to new data points. Also, it takes time in model-fitting. SVM is also useful for this type of problems.

**Recommendation and Conclusions:**

****

Based on the above chart, we can conclude that *MonthlyIncome* plays a very important role in deciding the attrition of the employee. Apart from that, *Age, DailyRate, TotleWorkingYears,DistanceFromHome* also are among the top contributors.

The HR Department can focus on the important variables that contribute significantly in determining if an employee is going to leave an organization. Such variables are:

* MonthlyIncome
* Overtime
* JobSatisfaction
* JobInvolvement
* WorkLifeBalance
* EnvironmentSatisfaction
* DistanceFromHome
* YearsAtCompany
* Age

Based on the above variables, one can clearly notice a pattern. The employees are more concerned with the materialistic objects that they get directly in hand. Then comes the psychological variables that determines if an employee might leave the organization.

Hence, the HR can focus on such aspects and understand from the viewpoint of the employees. Once that is followed, the project that is called Attrition project can be used as a Retention project. This can immensely help the organization.

Secondly, the model needs to be tuned from time to time as and when new dataset is received. In case any new input variable is introduced, it is important that the information is retrieved for the employees who participated in the initial study.

We hence conclude this project.