|  |  |
| --- | --- |
| Employee Attrition  EMPLOYEE ATTRITION | “Your number one customers are your employees. Look after your people first and then customers after.” |

**Attrition Case Study**

**Objective**:

Attrition is widely understood to be one of the major problems affecting organizations today. In this project, we will study various factors that leads employees to leave the company. The goal of this project is to predict employee attrition using machine learning algorithms so that organization can take steps to keep their employees satisfied.

Before studying the factors affecting the employee attrition, we need to understand how employee attrition impact any organization.

**Impact of employee attrition on organization**:

Losing employees has many direct and indirect impacts across a company. The direct impacts are relatively easy to measure: costs must be incurred to recruit and train new employees. During the recruitment process, other employees generally see an increase in their workload which can result overtime costs, but also affect their own productivity. In roles where employees maintain a close relationship with customers, employee attrition can also lead to reduced revenues from these customers and in the worst case, customer attrition.

However, there are also indirect impacts of employee attrition which, although they are not always immediately visible, can affect the organization’s processes and profitability.

1. **Team dynamics:** Most employees today work in teams and play a role in ensuring effective collaboration. Losing an employee can affect the smooth functioning of a team and lead to productivity losses. For example, losing a key engineering lead can have impacts across the whole team working under him.
2. **Acquired knowledge:** Employees who stay with organizations for a long period of time often acquire skills and knowledge on the job. Losing an employee can lead to the loss of all this acquired knowledge, which is not easily replaceable. For example, an employee who has spent several years on a project will not be easily replaced by another person with the same years of experience, because the project knowledge acquired will be lost.
3. **Employee engagement and morale:**Employees leaving an organization can have negative impacts on the morale and engagement of their colleagues. This impact often goes unnoticed, even though it can have tangible effects on productivity.
4. **Culture:** Employees who stay with organizations for a long time tend to take on the culture of the organization. Losing such employees and having them replaced by others can lead to a culture clash within the organization, with impacts on effective collaboration and employee engagement.
5. **Extended productivity:** Employee productivity increases over a period of time as they become more familiar with their role. Although employees who leave are replaced by others, it is often difficult for organizations to reach the same level of productivity. This is especially true of employees who have had long stints in the organization.

Now that we understand that employee attrition is the major cost to the organization, we aim to build a model to get a better explanation of the pattern linked to the employee attrition. So that organizations can strategize accordingly to make their employees satisfied and reduce attrition rate.

**Dataset Description:**

The dataset consist of 1470 observations and 35 features about employees’ information, evaluation about the company and attrition result. The features of the dataset is as follows:

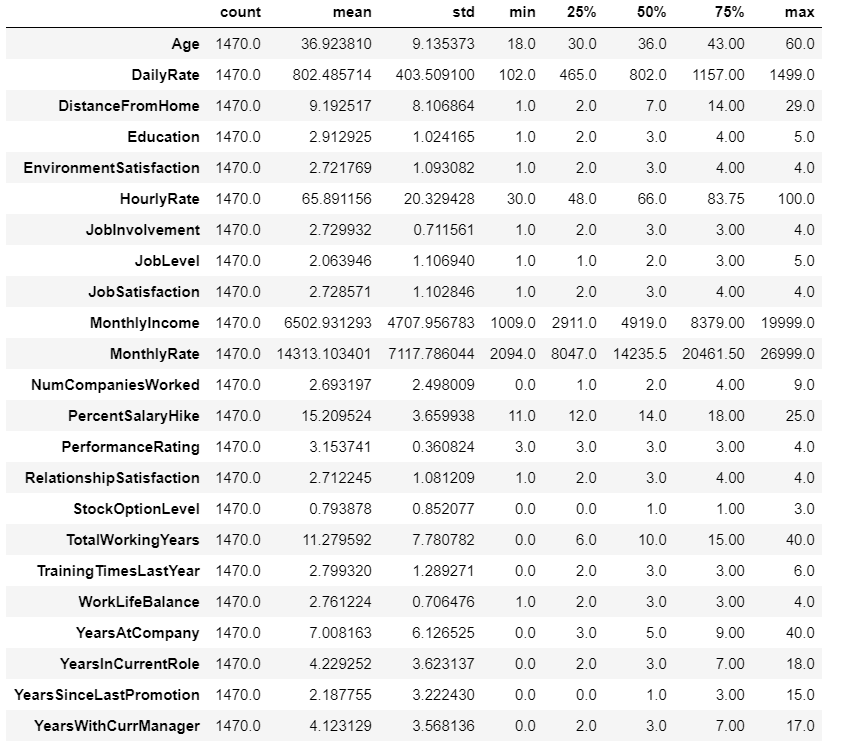
|  |
| --- |
|  |
|  |
|  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Field description** | **Variable type** | **String/****Numeric** | **Ordinal/****Nominal** |
| **Age** | **Age of employees in the company** | **Continuous** | **Numeric** |  |
| **Attrition** | **Target variable. 2 levels: Yes, No** | **Categorical** | **String** | **Nominal** |
| **BusinessTravel** | **Whether an employee travels frequently, rarely or doesn’t travel** | **Categorical** | **String** | **Nominal** |
| **DailyRate** | **Daily pay rate of an employee** | **Continuous** | **Numeric** |  |
| **Department** | **Department under which an employee works** | **Categorical** | **String** | **Nominal** |
| **DistanceFromHome** | **Distance of the work place from an employee’s home** | **Continuous** | **Numeric** |  |
| **Education** | **Education level of an employee. Value labels:****1- Below College 2- College 3- Bachelors 4- Masters 5- Doctor** | **Categorical** | **Numeric** | **Nominal** |
| **EducationField** | **Employee’s educational field like Life Sciences, Medical, etc.** | **Categorical** | **String** | **Nominal** |
| **EmployeeCount** | **Unary variable containing value 1** | **Unary** | **Numeric** |  |
| **EmployeeNumber** | **ID Number of an employee** | **Continuous** | **Numeric** |  |
| **EnvironmentSatisfaction** | **How satisfied an employee is with his work environment. Value labels: 1- Low 2- Medium 3- High 4- Very high** | **Categorical** | **Numeric** | **Ordinal** |
| **Gender** | **Gender of an employee: Male, Female** | **Categorical** | **String** | **Nominal** |
| **HourlyRate** | **Hourly pay rate of an employee** | **Continuous** | **Numeric** |  |
| **JobInvolvement** | **How heavily an employee is involved with his job. Value labels:****1- Low 2- Medium 3- High 4- Very High** | **Categorical** | **Numeric** | **Ordinal** |
| **JobLevel** | **Level of job of an employee: Value labels:****1- Low 2-Medium 3- High 4- Very high** | **Categorical** | **Numeric** | **Ordinal** |
| **JobRole** | **Current job role of an employee.** | **Categorical** | **String** | **Nominal** |
| **JobSatisfaction** | **How satisfied an employee is with his current job. Value labels:****1- Low 2- Medium 3- High 4- Very high** | **Categorical** | **Numeric** | **Ordinal** |
| **MaritalStatus** | **Whether an employee is married, single or divorced.** | **Categorical** | **String** | **Nominal** |
| **MonthlyIncome** | **Monthly income of an employee** | **Continuous** | **Numeric** |  |
| **MonthlyRate** | **Monthly pay rate of an employee.** | **Continuous** | **Numeric** |  |
| **NumCompaniesWorked** | **Total number of companies an employee has worked.** | **Continuous** | **Numeric** |  |
| **Over18** | **Whether an employee is over 18 or not.** | **Unary** | **String** |  |
| **OverTime** | **Whether an employee works overtime or not. 2 levels: Yes, No** | **Categorical** | **String** | **Nominal** |
| **PercentSalaryHike** | **Salary hike of an employee based on his performance** | **Continuous** | **Numeric** |  |
| **PerformanceRating** | **Rating given to each employee’s performance. Value labels:****1- Low 2- Good 3- Excellent 4- Outstanding** | **Categorical** | **Numeric** | **Ordinal** |
| **RelationshipSatisfaction** | **Satisfaction level of the customer-employee relationship.****Value labels:****1- Low 2- Medium 3- High 4- Very High** | **Categorical** | **Numeric** | **Ordinal** |
| **StandardHours** | **Standard hours an employee works (bi-weekly)** | **Unary** | **Numeric** |  |
| **StockOptionLevel** | **Level of stocks provided by the company: 0, 1, 2** | **Categorical** | **Numeric** | **Ordinal** |
| **TotalWorkingYears** | **Total number of years an employee has been working** | **Continuous** | **Numeric** |  |
| **TrainingTimesLastYear** | **The no. of trainings an employee has taken in the previous year** | **Continuous** | **Numeric** |  |
| **WorkLifeBalance** | **How well an employee can balance his work life. Value labels:****1- Bad 2- Good 3- Better 4- Best** | **Categorical** | **Numeric** | **Ordinal** |
| **YearsAtCompany** | **The total number of years an employee has spent with the company** | **Continuous** | **Numeric** |  |
| **YearsInCurrentRole** | **The number of years an employee has spent in his current job role** | **Continuous** | **Numeric** |  |
| **YearsSinceLastPromotion** | **The number of years since employee’s last promotion** | **Continuous** | **Numeric** |  |
| **YearsWithCurrManager** | **The number of years an employee has worked with his current manager** | **Continuous** | **Numeric** |  |

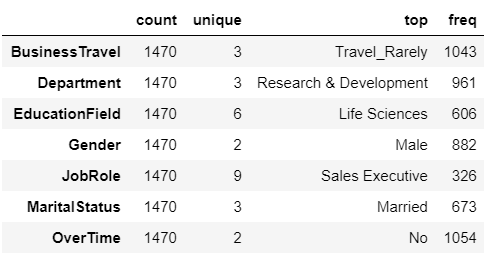
**Data Preprocessing and Exploratory Data Analysis:**

Before implementing machine learning algorithm it is important to understand the original data and various trends that data is following.

**Descriptive Statistics**: Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. With descriptive statistics you are simply describing what is or what the data shows.



The above table shows descriptive statistics for numerical data type. It returns mean, std, count and IQR values. The little difference between mean and median of the numerical variables show that there is little or no skewness in these variables.



The above table shows summary for object datatype. It returns thecount, unique value it contains, the value with higher frequency.

**Data Cleaning**:

**Detecting missing values:** Our dataset does not consist of any missing value. So we can proceed further.

**Detecting Outliers:** The boxplots of the variables showed that there is no outliers present in the data fortunately.

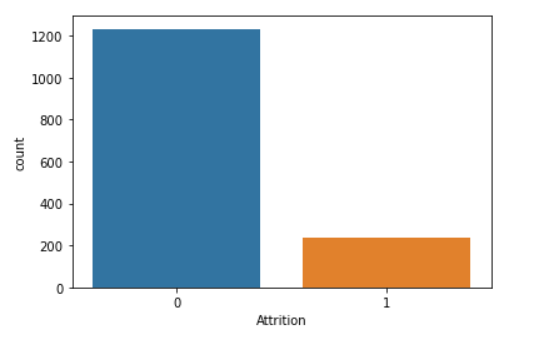
**Removing insignificant variables:** The following variables consists of unary values and contributes nothing to our analysis, so we directly remove them:

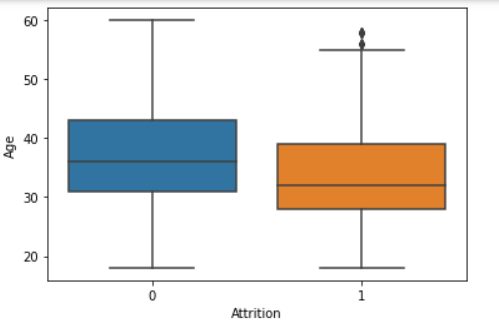
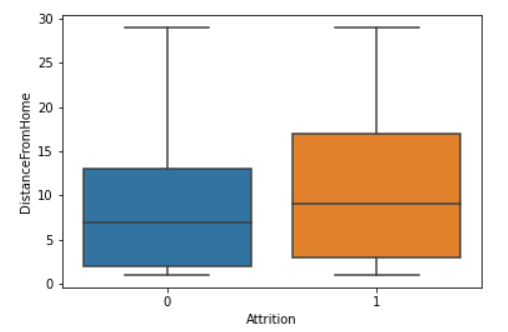
* EmployeeCount
* Over18
* StandardHours

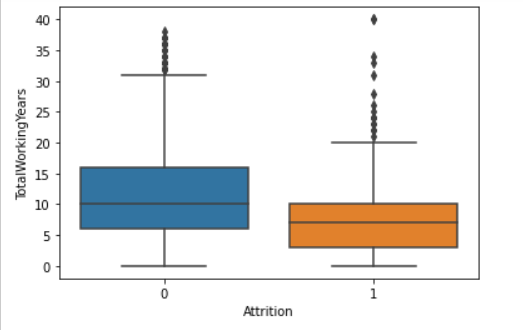
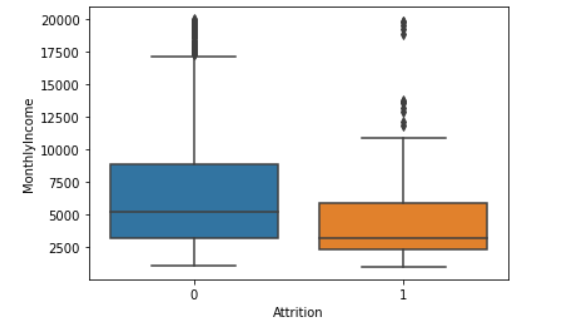
**Data Visualisation:**

Inference from data visualisation:

* The distribution of attrition is highly imbalanced i.e 237 out of 1470 employees tend to leave the company which is approximately 16% of total number of employees

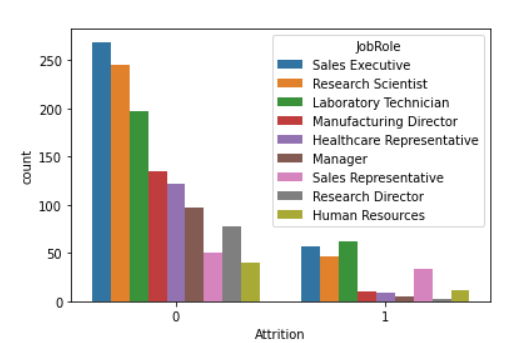
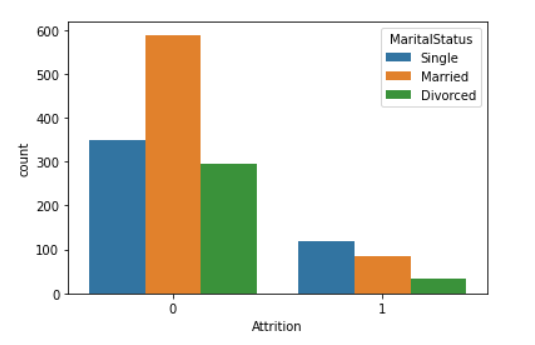


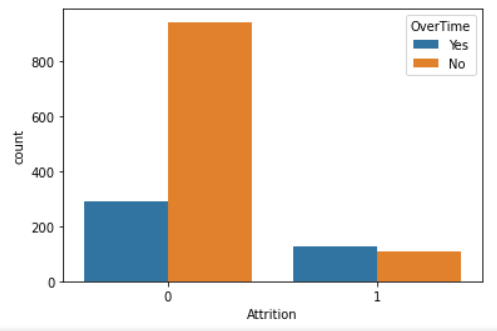
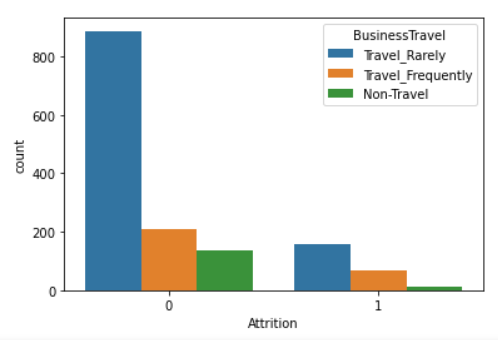
* Distribution of Attrition against continuous variable: 

The above boxplots shows that how attrition changes with different variables:

* The attrition rate is more for employees in their young age. This can be justified as young employees are clear for their career path and thrives for learning. Whereas the older employees are settled in their comfort zone.
* The attrition rate is more for employees who lives at a far distance from home which is obvious as employees prefer the work location near their homes.
* The attrition rate is more for employees who worked for less no of years.
* The attrition is more for employees with less monthly income which is justifiable as sometimes changing jobs is about money.
* Distribution of attrition against categorical variables :

Inferences made from above countplote:

* The attrition rate is more for employees under lab technician, sales executive, research scientist, sales representative. So these job roles needs to be focused.
* The attrition rate is more for employees who are single which can be explained in a way that single employees are more focused on their career path.
* Employees who does overtime tends to leave the organization. It seems that employees are not rewarded for their extra effort
* The attrition rate is more for employees who have less business travel chance

**Hypothesis Testing:**

A statistical hypothesis is a hypothesis that is testable on the basis of observed data modeled as the realised values taken by a collection of random variables.

The main aim of this model is to answer the following questions: What are the common characteristics of employees lost in attrition?

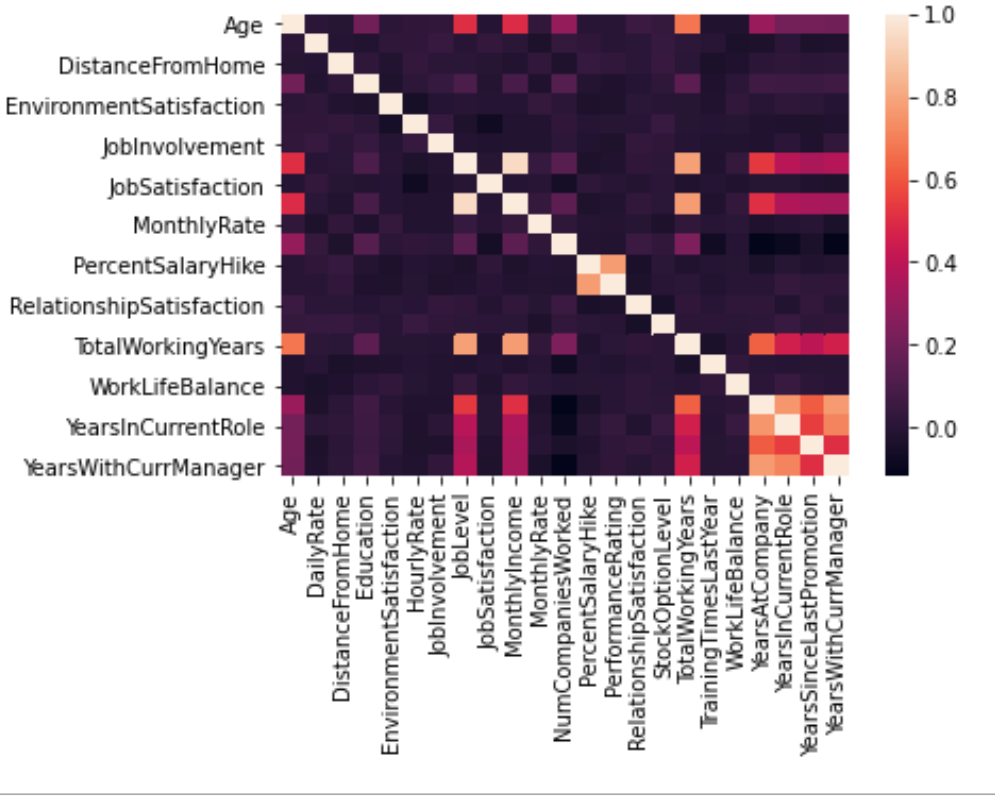
These models includes statistical tests and make inferences about the population.

Blow mentioned are the conclusions drawn from these statistical tests:

* The employees lost in attrition have 3-4 years average age less than those who stayed.
* The daily rate of employees who left had average daily rates less than those who stayed.
* The employees who left had average commute distance more than those who stayed.
* The employees who left were less satisfied with their work environment on average.
* The job involvement level for employees who left is less on average than those who stayed.
* The employees lost in attrition belonged to lower job level on average than those who stayed.
* The employees who left were less satisfied with their job.
* The average monthly income of employees who left the company is less than those who stayed.
* The employees who left worked for less years in the company on average than those who stayed.
* The employees who left were given less stock option levels on average
* The employees lost in attrition had less training opportunities than those who stayed.
* The employees who left had poor work life balance on average.
* The employees who stayed worked for more years in their current role on average than those who left.
* The employees who left worked under their current manager for less years on average than those who stayed.
* The employees who left had a less chance for business travel.
* The employees from sales and R&D department are more who left the organization.
* The employees under job role Laboratory Technician, Sales Executive, Sales Representative are more who left the organization.
* The employees who does overtime tends to leave the organization
* Single employees tends to leave the organization
* The average monthly income of sales and HR department is more from R&D department
* The average monthly income for employees under job role laboratory technician, Research Scientist and Sales Representative is comparatively less.

**Correlation Analysis** :

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Below is the correlation matrix of the variables.



Inference from correlation matrix:

* Job level and monthly income are positively correlated which means that the monthly income for higher job level employees is comparatively more
* Performance rating and percent salary hike are positively correlated i.e the employees with high performance rating gets more salary hike.
* Total working years and monthly income are positively correlated.
* Total working years and job level are correlated i.e job level increases with increase in working years.

**Predicting Attrition** :

Since our main aim is to predict whether the employee leaves the company or stays. We have performed a handful classification based models: Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest.

Before digging deeper into every model and selecting best model, we will understand the basic steps used before applying any classification model to the data.

* Converting categorical variables to dummy variables: Predicting models require all input and output variables to be numeric. The categorical variable is therefore converted to dummy variables. A dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.
* Splitting data to train and test: In a dataset, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. ... So, we use the training data to fit the model and testing data to test it. The models generated are to predict the results unknown which is named as the test set.
* Balancing target data: As we have already seen that the distribution of data is highly imbalanced. When you have imbalanced classes, your model might just learn to choose the majority class instead of reacting to the data. So we have used SMOTE technique to balance the attrition distribution.

Now, Let’s dive deeper into classification models one by one.

**Logistic Regression :**

The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function. Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The best predictors are found using forward selection technique.

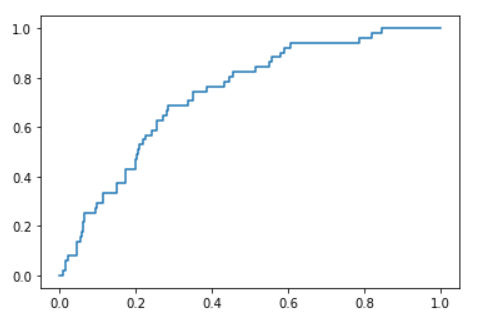
Before seeing the results for every model, we need to understand few metric terms:

* **Accuracy Score:** Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. It is the number of correct predictions made divided by the total number of predictions made, multiplied by 100 to turn it into a percentage.
* **Precision:** precision gives answer for what proportion of positive identifications made by model was actually correct.
* **ROC curve:** A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
* **ROC AUC:** The area under the ROC curve (AUC ) is a measure of how well a parameter can distinguish between two diagnostic groups.

Results of Logistic Regression model:

|  |  |
| --- | --- |
| Parameter | result |
| Accuracy score | 0.765 |
| precision | 0.327 |
| roc\_auc | 0.736 |

ROC\_AUC curve:



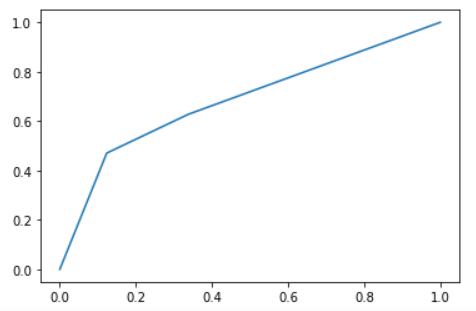
* The accuracy score is 0.765 which means out of total predictions, the model is able to predict 76% of the times correctly.
* The precision is 0.327 which means when it predicts the employee will leave, it is correct the employee will leave, it is correct 32% of the time which is quiet low.
* ROC\_AUC score is 0.736 which signifies how good the model is aggregating all the threshold values.

**K-Nearest Neighbors**

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. All such distance based algorithms are affected by the scale of the variables. all the features needs to be in the same scale so that there is no biasing.

Results of KNN model:

|  |  |
| --- | --- |
| Parameter | result |
| Accuracy score | 0.806 |
| precision | 0.45 |
| roc\_auc | 0.69 |
|  |  |



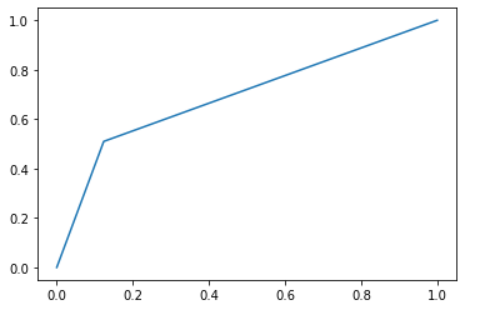
* The accuracy score is 0.806 which means out of total predictions, the model is able to predict 80% of the times correctly.
* The precision is 0.45 which means when it predicts the employee will leave, it is correct the employee will leave, it is correct 45% of the time which is quiet low.
* ROC\_AUC score is 0.69 which signifies how good the model is aggregating all the threshold values.

**Decision Tree:**

Decision Tree algorithm belongs to the family of supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

Result of Decision Tree model:

|  |  |
| --- | --- |
| Parameter | result |
| Accuracy score | 0.813 |
| precision | 0.46 |
| roc\_auc | 0.69 |



* The accuracy score is 0.813 which means out of total predictions, the model is able to predict 81% of the times correctly.
* The precision is 0.46 which means when it predicts the employee will leave, it is correct the employee will leave, it is correct 45% of the time which is quiet low.
* ROC\_AUC score is 0.69 which signifies how good the model is aggregating all the threshold values.

**Decision Tree using max depth**:

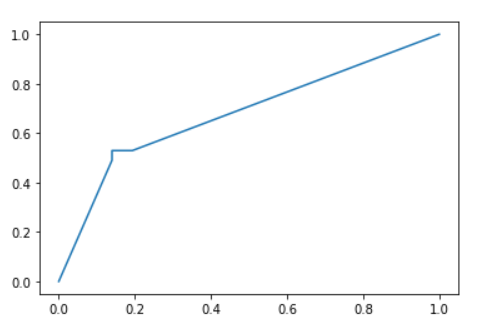
One of the problems with the decision tree is it gets overfit with training sample and becomes too large and complex. A complex and large tree poorly generalizes the new samples data whereas a small tree fails to capture the information of training sample data. Pruning may be defined as shortening the branches of the tree. The process of reducing the size of the tree by turning some branch node into the leaf node and removing the leaf node under the original branch. Pruning I very useful in the decision tree because sometimes what happens is that decision tree fits well on training data very well but performs very poorly on testing data.

To generate decision trees that will generalize to new problems well, we can tune different aspects of tree:

The maximum depth of the tree is simply the largest distance between the root to a leaf. It is also bad to have a very low depth because your model will underfit sohow to find the best value, experiment because overfitting and underfitting are very subjective to a dataset, there is no one value fits all solution.

Results after max depth:

|  |  |
| --- | --- |
| Parameter | result |
| Accuracy score | 0.802 |
| precision | 0.44 |
| roc\_auc | 0.679 |



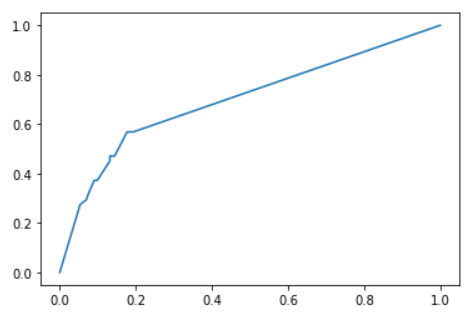
* The accuracy score is 0.803 which means out of total predictions, the model is able to predict 80% of the times correctly.
* The precision is 0.44 which means when it predicts the employee will leave, it is correct the employee will leave, it is correct 45% of the time which is quiet low.
* ROC\_AUC score is 0.67 which signifies how good the model is aggregating all the threshold values.

**Decision tree using Min Sample leaf :**

While splitting a node, one could run into the problem of having more no of samples in one of them, and less on the other. This will not take us too far in our process and would be waste of resources and time. If we want to avoid this, we can set a minimum for the number of samples we allow on each leaf.

Result:

|  |  |
| --- | --- |
| Parameter | result |
| Accuracy score | 0.809 |
| precision | 0.44 |
| roc\_auc | 0.70 |



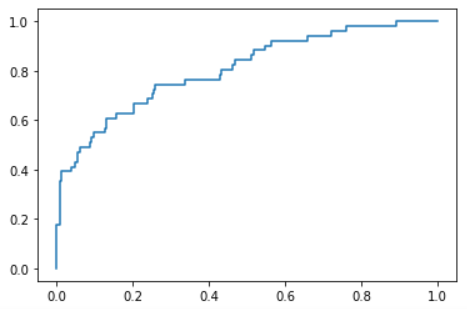
* The accuracy score is 0.809 which means out of total predictions, the model is able to predict 80% of the times correctly.
* The precision is 0.44 which means when it predicts the employee will leave, it is correct the employee will leave, it is correct 45% of the time which is quiet low.
* ROC\_AUC score is 0.70 which signifies how good the model is aggregating all the threshold values.

**Random Forest:**

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Results:

|  |  |
| --- | --- |
| Parameter | result |
| Accuracy score | 0.881 |
| precision | 0.83 |
| roc\_auc | 0.806 |



* The accuracy score is 0.88 which means out of total predictions, the model is able to predict 88% of the times correctly.
* The precision is 0.83 which means when it predicts the employee will leave, it is correct the employee will leave 83% of the time.
* ROC\_AUC score is 0.80 which signifies how good the model is aggregating all the threshold values.

CONCLUSION : Since the accuracy score, precision and area under the curve of ROC curve for random forest is max, we can consider this as best model.

SUMMARIASATION:

The main aim of this project was to find the similar characteristics followed by employees who are most likely to leave. Using data exploration, we did observe some of the common patterns. To control the attrition rate the organization needs to focus on their young employees giving them training opportunities, reward for their overtime. Sales and R&D department needs more focus as most employees who left worked for this department. Most employees who left are not satisfied with their job. Also monthly income plays a major role here. As a suggestion, the organization can introduce a forum for the employees where they can share their views about their issues, why they are not fully involved in the job, why they are not satisfied and company and employees can work on it by mutual understanding. Since employees who left had poor work life so organisation can offer flexible work schedules to employees having poor work life balance to retain their employees.