

HR Analytics

Aaron Dsouza

Data Science

18-04-2021

# Problem Definition:

* Hiring and retaining employees are extremely complex tasks that require capital, time, and skills.
* Small business owners spend 40% of their working hours on tasks that do not generate any income such as hiring.
* An average company loses anywhere between 1% and 2.5% of their total revenue on the time it takes to bring a new hire up to speed.
* Companies spend 15% - 20% of the employee's salary to recruit a new candidate.
* It takes 52 days on an average to fill a position.
* As an aspiring Data Scientist, I have been allocated extensive data which has been collected by an HR team. My job is to develop a model that could predict which employees are more likely to quit.

## Data Analysis:

* Data set link: <https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics>
* The Human Resources dataset has 1470 entries. Each entry contains the following information about an individual:
* Age: Age of an employee.
* Attrition: Attrition implies weather an employee will stay or quit the organization. (Target variable)
* Business Travel: It implies whether an employee travels rarely, frequently or he /she does not travel.
* Daily Rate: The daily rate per employee.
* Department: The various departments in which the employees work.
* Distance from Home: Distance an employee takes to reach the organization (in Kms).
* Education: Describes the education levels of an employee.
* Education Field: Describes the Qualification field of an employee
* Employee Number: Unique employee number.
* Environment Satisfaction: Environment satisfaction levels of an employee
* Gender: gender of an employee
* Hourly Rate: Hourly rate per employee
* Job Involvement: job involvement level of an employee
* Job Level: Job level of an employee.
* Job Role: Job role per employee.
* Job Satisfaction: levels of job satisfactions.
* Marital Status: Employee Marital Status
* Monthly Income: monthly income per employee.
* Monthly Rate: monthly rate per employee.
* Number of Companies Worked: The total number of companies an individual employee has worked previously.
* Over Time: Whether an employee is getting over time or not.
* Percent Salary Hike: Salary Hike in percentage per employee.
* Performance Rating: performance rating per employee.
* Relationship Satisfaction: relationship satisfaction per employee.
* Standard Hours: standard hours per customer.
* Stock Option Level: Whether an employee has bought company stocks or not.
* Total Working Years: total years the employee has been with the company.
* Training Times Last Year: count of training times per employee.
* Work Life Balance: levels of work life balance per employee.
* Years at Company: number of years in company.
* Years in Current Role: number of years in current role
* Years Since Last Promotion: number of years past since last promotion.
* Years with Current Manager: number of years with current manager.

### Problem Type:

### The target variable is Attrition.

### 

### The target variable is categorical in nature.

### 

### Solving it as a Classification Problem.

### 

### On the circumstances given we need to predict whether an employee will leave the company or not.

### Data frame information:

### The dataset has two datatypes Object and Int.

### 

### There are no duplicated values in dataset.

### 

### No null values in the dataset.

### 

### A total of 1470 employees in the company.

### 237 employees left the company.

### 1233 employees stayed in the company.

### A total of 16% employees left the company and 84% employees stayed in the company.

# EDA (Exploratory Data Analysis):

# 

# Fig. 1 Job Role with hue as Attrition

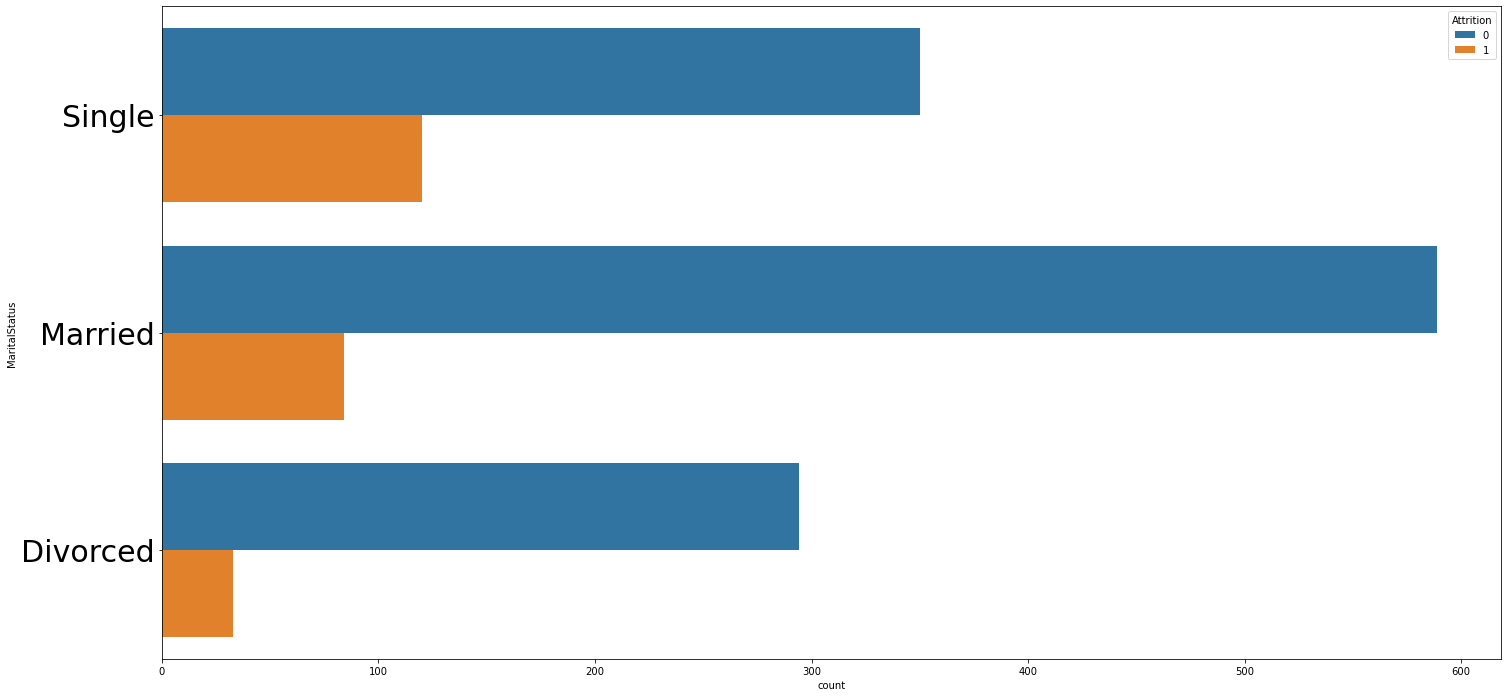
### The Job role feature describes the individual roles per employee. Fig. 1

### Shows the distribution of job roles among the entries in the dataset. We can clearly observe that most of the employees are working in the Sales Executive role. With the help of count plot from the seaborn library and taking the hue attribute as Attrition we can conclude that in terms of job role, the highest number of employees who quit the company were working as Laboratory Technician. We can say that the better the position of an employee in a company higher are the chances of he/she not leaving the company.

### 

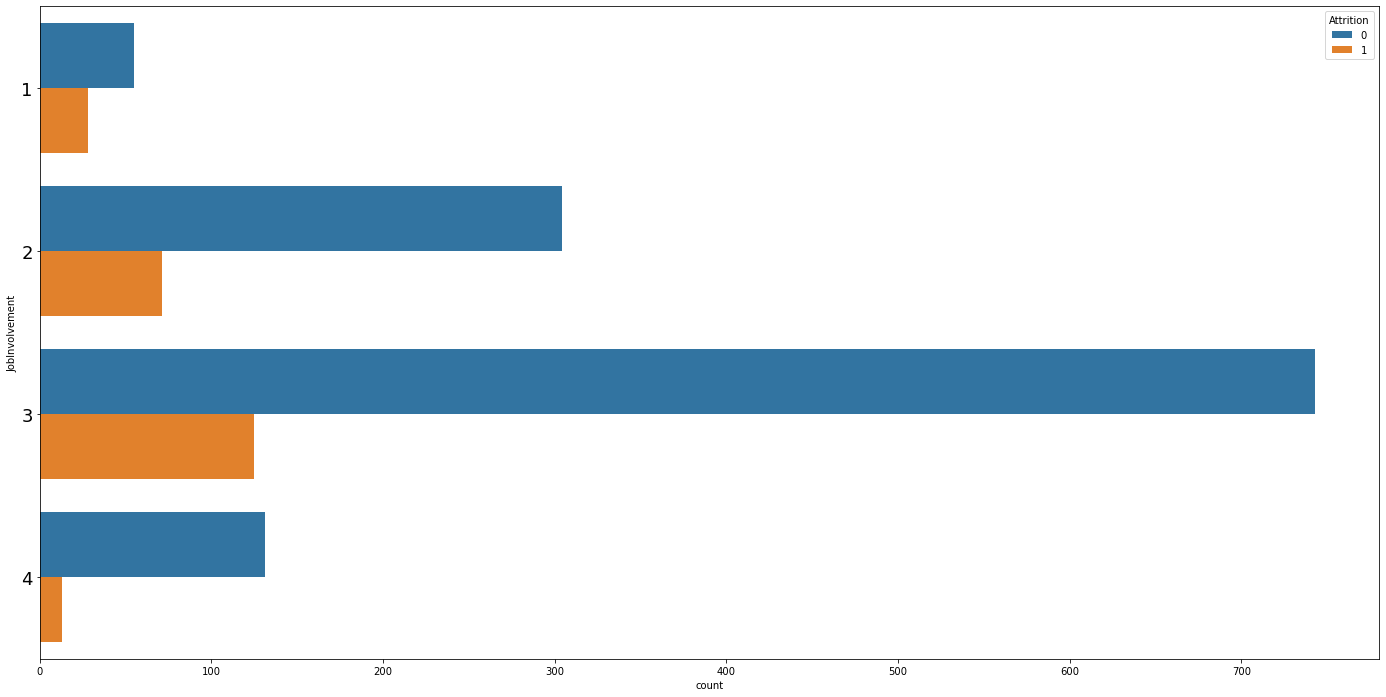
# Fig. 2 Age with hue as Attrition

The Age feature describes the age of an individual. Fig. 2 shows the age distributions among the entries in our dataset. The ages range from 18 to 60 years old with the majority of ages between 27 to 45 years. We used count plot from the seaborn library and using hue attribute as Attrition we can conclude that younger people especially between the age of 18 to 24 are having a higher tendency of leaving the job. Maybe they are looking for better opportunities. Most of the employees beyond the age of 40 have a lower tendency to leave the job. Maybe because of the experience they have and are more important to the company.



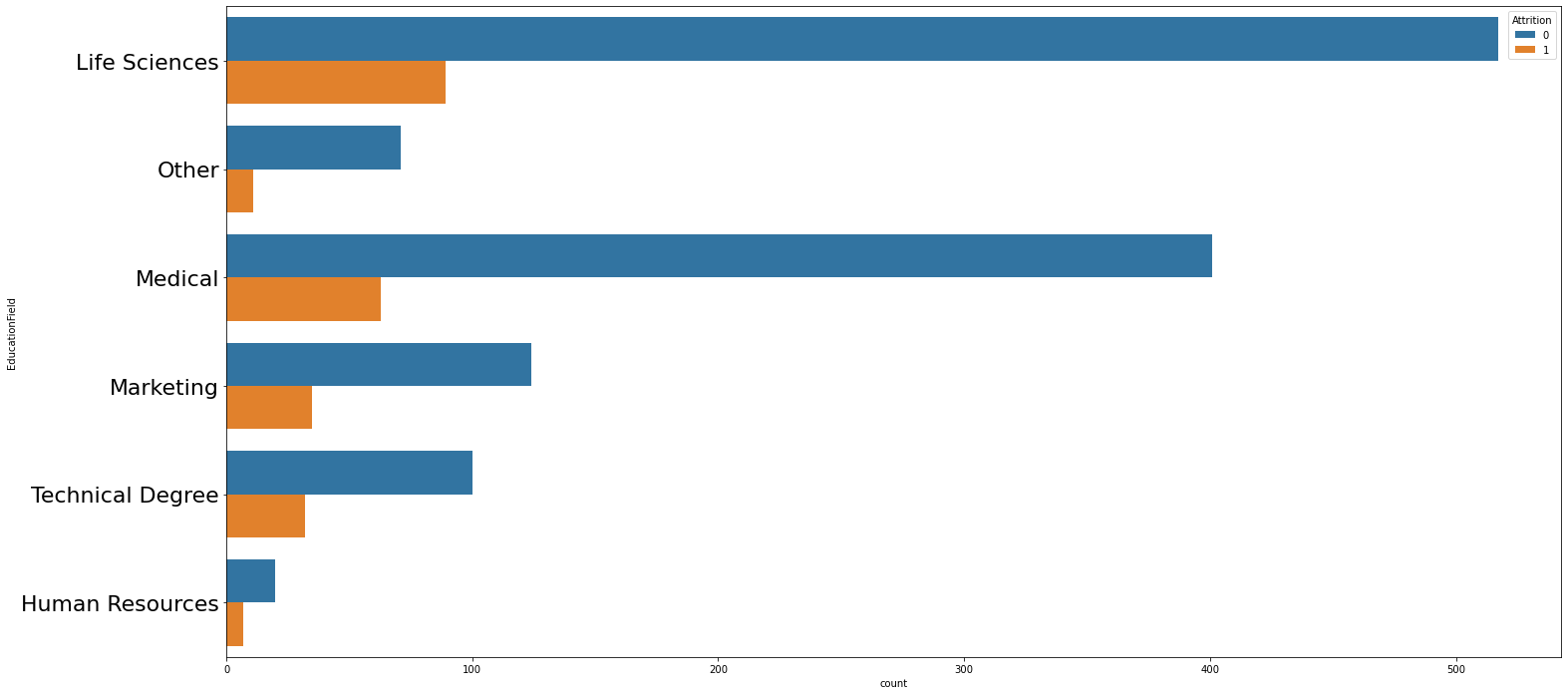
# Fig. 3 Marital status with hue as Attrition

The Marital status describes the whether an employee is Married, Single or Divorced. Fig. 3 shows the distribution of marital status of employees in the company. In Fig. 3 the count plot shows the distribution of marital status with hue as attrition. From the plot we can conclude that Single employees tend to leave compared to married and divorced employees.



# Fig. 4 Job involvement with hue as Attrition

The Job involvement describes the level of participation an employee has in his/her department of work. Fig. 4 shows the distribution of job involvement levels of employees there are 4 levels. In Fig. 4 the count plot shows the distribution of job involvement levels with hue as attrition. From the plot we can conclude that the lower the job involvement level for an employee the higher are the chances that the employee will quit the organization.



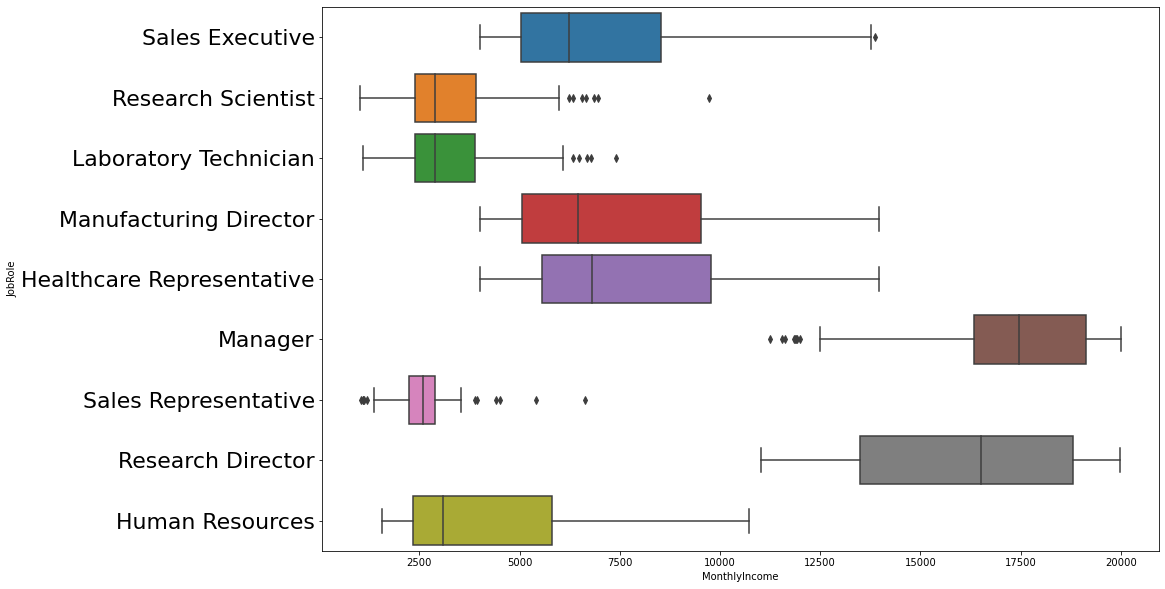
# Fig. 5 Education field with hue as Attrition

### The Education field describes the education domains the employees belong to. Fig. 5 shows the distribution of education fields which include Life Sciences, Medical, Marketing, Technical Degree, Human resource and other fields. In Fig. 5 the count plot shows the distribution of Education fields with hue as attrition. From the plot we can conclude that the greatest number of employees are working in the life Science field and the life science field also has the highest rate in terms of employees leaving the company. Human recourses field has the best rate in terms of employees not leaving the organization.

### 

# Fig. 6 Monthly income vs Gender

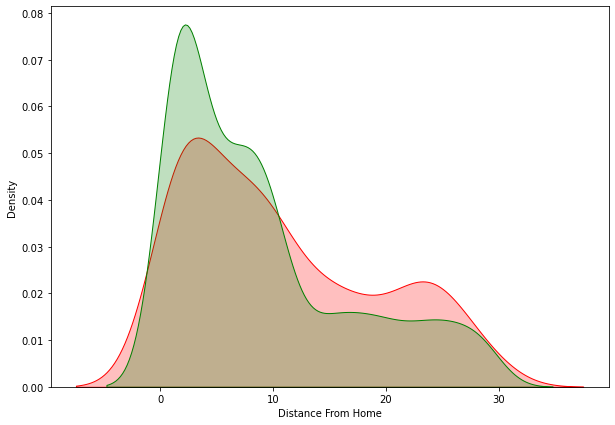
Fig. 6 Describes the bivariate relationship between the Monthly income and Gender. From Fig. 6 we can clearly observe that the Female employees are having higher Monthly income as compared to the Male employees in the organization.



# Fig. 7 Monthly income vs Job role

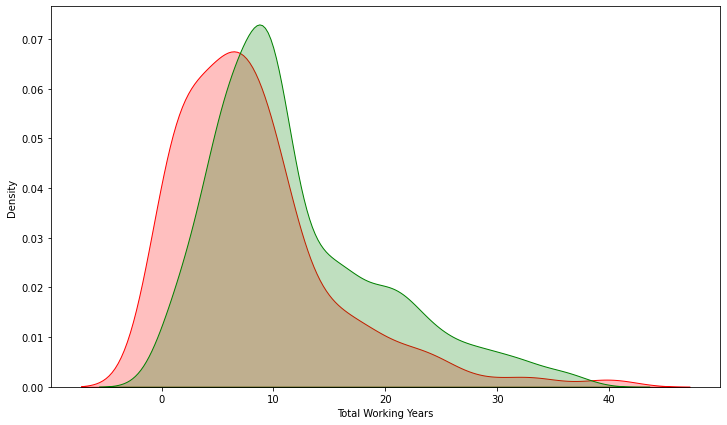
### 

Fig. 7 Describes the bivariate relationship between the Monthly income and Job role. From Fig. 7 we can clearly observe that the individuals belonging to the Job role Manager has the highest monthly income as compared to the other job roles. The Sales representatives is the lowest paid job role in the organization.



# Fig. 8 Distance from home (KDE plot)

KDE (Kernel Density Estimate) is used for visualizing the Probability Density of a continuous variable. KDE describes the probability density at different values in a continuous variable. In Fig .8 the red curve represents the employees who left the organization and the green curve represents the employees who stayed in the organization. There is a higher chance that an employee might leave the organization if the distance from home is more than 10 km.



# Fig. 9 Total working years (KDE plot)

# In Fig .9 the red curve represents the employees who left the organization and the green curve represents the employees who stayed in the organization. From the above plot it is evident that there is a very low chance that any employee will leave the organization after completing more than 20 years of service.

### 

# Fig. 10 Years with current manager (KDE plot)

# In Fig .10 the red curve represents the employees who left the organization and the green curve represents the employees who stayed in the organization. From the above plot it is evident that the more the years an employee has with the current manager, higher are the chances that the employee will stay with the organization.

# Pre-processing Pipeline:

### No null values in the data set.

### 

### The columns: "**EmployeeCount","StandardHours","EmployeeNumber","Over18"** have been dropped from the data frame. These features were not useful for our analysis or had too much bad data.

### There where OUTLIERS detected in the columns “Monthly income”, “Number of companies worked” and “Years in current role”.

### The outliers were not removed as it is entirely possible that the outlier values could be genuine. Instead for Machine Learning those algorithms were used which are not affected by outliers.

### Data Skewness was considered within the range (-0.5 to +0.5) to obtain Normally Distributed curves.

### Skewness was only considered in continuous variables and not in categorical variables.

### One Hot Encoding technique from Sklearn was applied on all the categorical variables.

### Train Test split from Sklearn was used in order to split the data for training and testing the model taking the test size as 0.25

### Min Max Scalar from Sklearn was used to transform the features in a given range of [0,1].

# Building Machine Learning Models:

### A total of 8 Machine learning (Classification) models were used in order to predict whether the employee will stay or the employee will leave the organization.

### 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 dichotomous, which means there would be only two possible classes.

### For loop was used to decide the best random state for the algorithm

### Stratified Cross validation technique was used.

### number of splits = 10 to decide the cross-validation score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Training Score** | **Accuracy Score** | **Cross validation Score** |
| Logistic Regression | 0.839 % | 0.842 % | 0.834 % |

### Naive Bayes:

### Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms

### For loop was used to decide the best random state for the algorithm

### Stratified Cross validation technique was used.

### number of splits = 10 to decide the cross-validation score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Training Score** | **Accuracy Score** | **Cross validation Score** |
| Naïve Bayes | 0.79 % | 0.81 % | 0.74 % |

### Random Forest Classification:

### Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

### For loop was used to decide the best random state for the algorithm

### Stratified Cross validation technique was used.

### number of splits = 10 to decide the cross-validation score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Training Score** | **Accuracy Score** | **Cross validation Score** |
| Random Forest Classifier | 1 % | 0.87 % | 0.86 % |

### Decision Tree Classification:

### Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**.

### For loop was used to decide the best random state for the algorithm

### Stratified Cross validation technique was used.

### number of splits = 10 to decide the cross-validation score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Training Score** | **Accuracy Score** | **Cross validation Score** |
| Decision Tree Classification | 1 % | 0.80 % | 0.80 % |

### Support Vector Classification:

### Support vector machines (SVMs) are particular linear classifiers which are based on the margin maximization principle. They perform structural risk minimization, which improves the complexity of the classifier with the aim of achieving excellent generalization performance.

### For loop was used to decide the best random state for the algorithm

### Stratified Cross validation technique was used.

### number of splits = 10 to decide the cross-validation score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Training Score** | **Accuracy Score** | **Cross validation Score** |
| Support Vector Classification | 0.839 % | 0.836 % | 0.838 % |

### ADA Boost Classification:

### AdaBoost is best used to boost the performance of decision trees on binary classification problems.

### For loop was used to decide the best random state for the algorithm

### Stratified Cross validation technique was used.

### number of splits = 10 to decide the cross-validation score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Training Score** | **Accuracy Score** | **Cross validation Score** |
| ADA Boost Classification | 0.895 % | 0.896 % | 0.876 % |

### K-NN Classification:

### K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

### For loop was used to decide the best random state for the algorithm

### Stratified Cross validation technique was used.

### number of splits = 10 to decide the cross-validation score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Training Score** | **Accuracy Score** | **Cross validation Score** |
| K-NN Classification | 0.85 % | 0.85 % | 0.81 % |

### Gradient Boosting Classification:

### Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.

### For loop was used to decide the best random state for the algorithm

### Stratified Cross validation technique was used.

### number of splits = 10 to decide the cross-validation score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Training Score** | **Accuracy Score** | **Cross validation Score** |
| Gradient Boosting Classification | 0.96 % | 0.90 % | 0.87 % |

# 

# Fig. 11 Algorithm performance table

# **Algorithm performance:**

# 

# Fig. 12 Algorithm performance

# Based on the cross-validation score ADA Boost classifier has the best performance.

# **Hyperparametric Tuning of ADA Boost Classifier:**

# Grid Search CV from Sklearn was used for hyperparametric tuning.

# Best Estimators (**algorithm='SAMME', learning\_rate=1, n\_estimators=100**)

# Best Parameters ('algorithm': 'SAMME', 'learning\_rate': 1, 'n\_estimators': 100)

# Best Score was 0.8602 %

# Model Accuracy was 0.90 %

### Concluding Remarks:

### 

# Fig. 12 Feature Importance

# The first step I took, was to visualize the distribution of each feature and its effect on the employee **Attrition** (whether an employee will stay or quit the organization). From the analysis, I conclude that the most useful features for predictions were “Monthly Income”, “Monthly Rate”, “Hourly rate”. **The ADA Boost Classifier** proved to be the best model for classification based on the Cross-Validation scores. An accuracy of 0.90 % was achieved by hyperparametric tuning of the model.

# ROC AUC CURVE:

# 

# Fig. 13 Roc Auc Curve

### In Fig 13 based on the True positive and False positive rate and the Area under the curve we can conclude that the model is performing good.

### 