IBM HR Analytics Employee Attrition and Performance

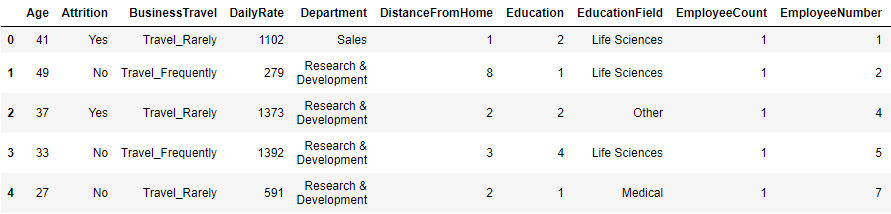
****

* **Problem Statement:**
* The purpose of this project is to predict the attrition of each employee, to find out which employees are more likely to leave the organization. This type of predictions will help organizations to prevent attrition and plan in advance to hire the new candidates.
* Here, we have dataset of IBM HR analytics employee attrition. This dataset contains records of 1470 employees with 35 variables. All variables are related to employees working life and personal characteristics.
* Out of 35 variables, one variable is dependent (Target) variable and 34 variables are independent variables.
* Independent variables contain employee’s data like age, education, years at company, total working years etc. Target variable is attrition with possible outcome of yes or no. ’Yes’ depicts an employee that left the company and ‘No’ depicts an employee that did not leave the company.
* In this dataset we have to find out, which employees leave the organization or not.
* **EDA(Exploratory Data Analysis):**

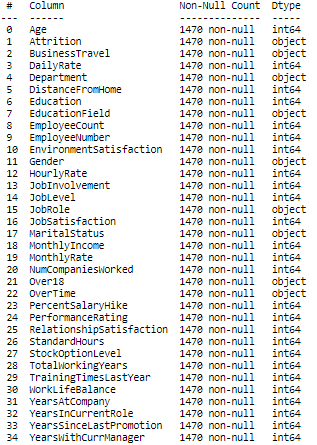
**Data Preparation:**

Data preparation is one of the most important aspects of machine learning. It takes more time if our data is more complex or uncleaned. Now, let’s look at our dataset.





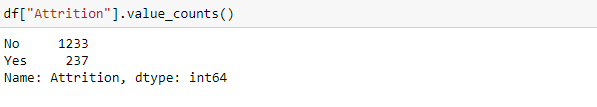




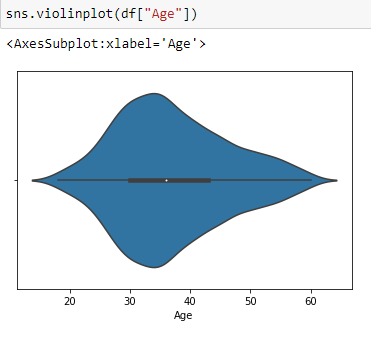
We don’t find any missing values in our data and data is also cleaned, so there is no need to format data. We also check data types of each attributes and there is no need to change data types of any variables.

**Data Analysis:**

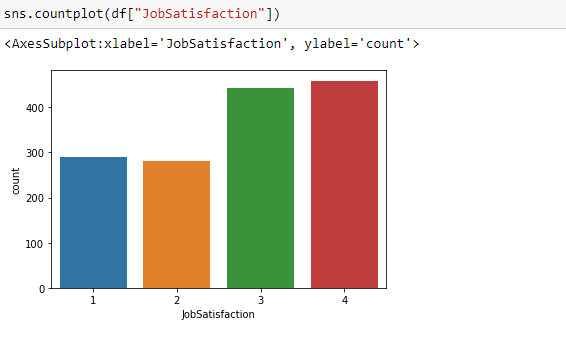
Now, let’s look at the data and extract some useful information from the data:



* Out of 1470 observations, we found 1233 employees with no and 237 employees with yes in attrition column. So, this depicts imbalanced classification in data and we will solve this during splitting train and test data.

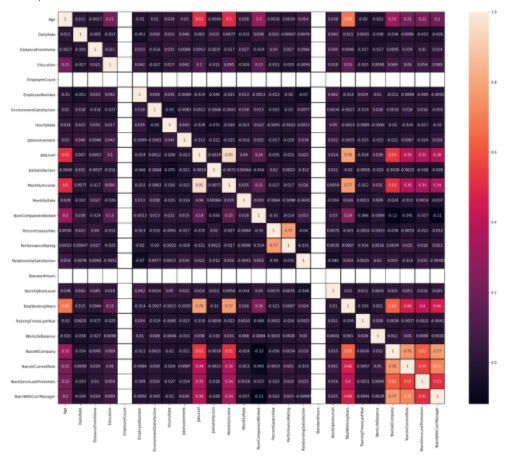


Most of the employee’s age is between 28 to 38 years in our dataset.

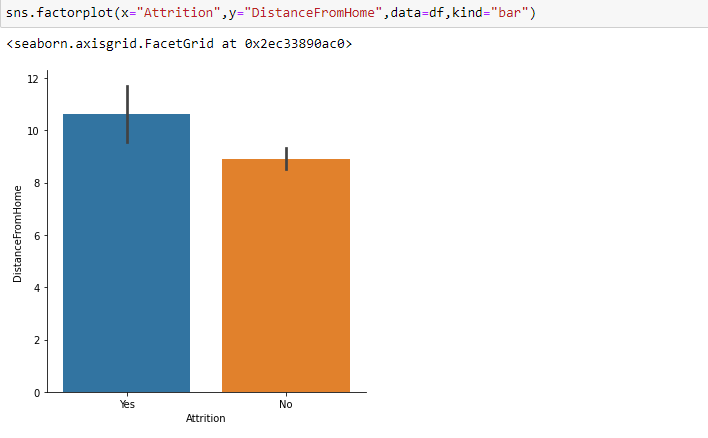


Job satisfaction is divided into four levels. From level 1 represents low job satisfaction to level 4 high job satisfaction. Most of the employees have high job satisfaction which is level 3 and 4, while nearby 600 employees have low job satisfaction which is level 1 and 2.

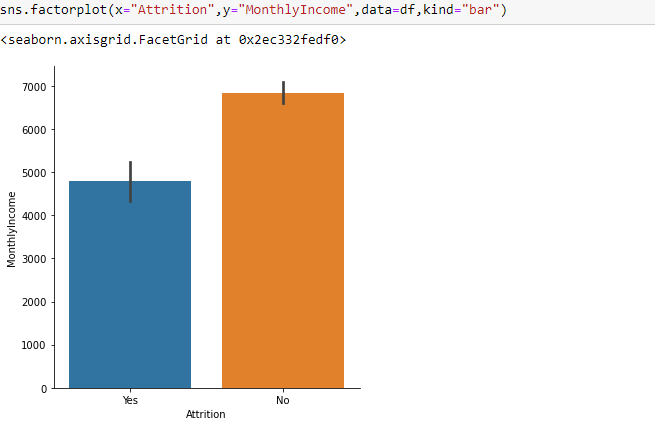
* Now, check the correlation of every numerical attributes using heatmap.
* Heatmap is nothing but a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors.



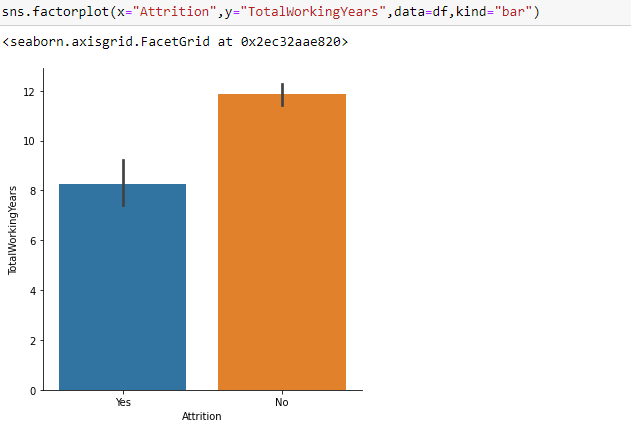
* 1 represents the high positive correlation, while 0 represents weak negative correlation.
* From the above heatmap we can see which variables are strongly correlated and which variables are poorly correlated.
* Now, do analysis of attrition with other variables.



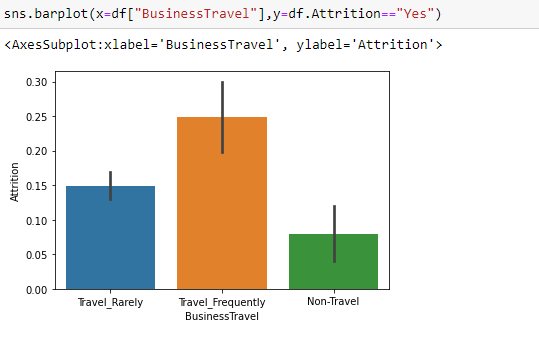
Major of the employees who leave their jobs are travel far from home, which is almost above 10 km.



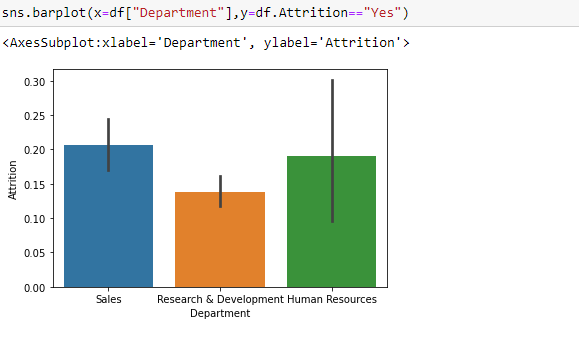
From the above count plot we can see that employees with low salary are more likely to leave their job. Most of the employees who are likely to leave their job have average salary below 5000 Rs. per month.



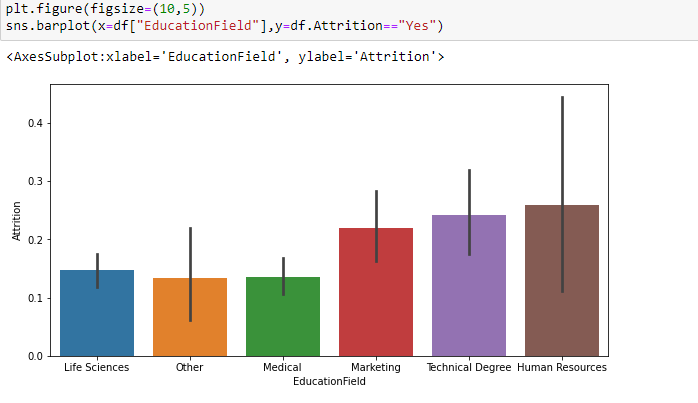
Employees with low or moderate work experience left their job, while employees with high work experience which is above 11 years are not likely to leave their job.



From the above bar plot we can see that employees who travel frequently for work are more likely to leave their job, while employees who do not travel are less likely to leave their job.



There are three departments where employees work. Most of the employees who work in sales and HR department are left their job, while who work in R&D department is less likely to leave their job compared to sales and HR department.



Employees are from different education backgrounds like HR, marketing, medical, etc. but, employees whose education fields are HR, technical and marketing are more likely to leave their job compared to other education field. Employees from medical and other educational background are less likely to leave their job.

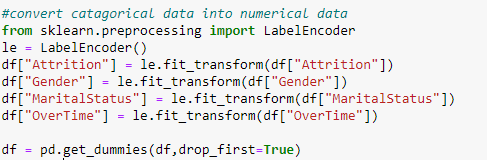
* Now, we drop unnecessary attributes or columns. which is described below:

1. Employee number: This is just continues count of employees and don’t affect our dataset so, we drop this column.
2. Employee count: This is just count of the employee in each row and has constant value throughout the data, which is 1.
3. StanderdHours: This is employee’s standard working hours per week and it has constant value throughout data, which is 80.
4. Over18: This attribute describe employees age over 18 or not and it has also constant value ‘Yes’ in allover dataset because all of the employees who work in this organization are above 18 years old.

**Encoding:**

* After dropped unnecessary attributes, we convert categorical data into numerical data because machine learning model takes input data only in a numerical form. Here, we use two methods for encoding data.

1. Label Encoder: This method converting categorical data into numerical data in numerical sequence like 1,2,3,4 etc. This method used on ordinal data and also used when we have large number of unique values in single attribute.
2. Dummies: This method produce new dummy column for each unique categorical value. The dummy column is one which has a value of one when a categorical event occurs and a zero when it doesn’t occur.

Our dataset code for encoding data is as below:

**Handling Outliers:**

* Now, we check and remove outliers from data.

Outliers are extreme values that fall a long way outside of the other observations. Here, we checked our data with two methods for removing outliers.

* 1) Zscore and 2) Inter Quartile Range

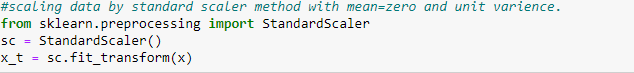
By using both methods, we conclude that percentage loss of data is very high, which is approximately above 40%.So, removing outliers is not a good idea because it impact on our result.

* We split our independent variables as x and target variable as y for further process.



**Scaling:**

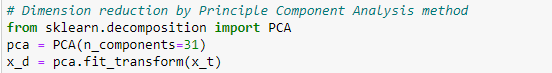
After splitting the data, we are scaling our data by using standard scaler method. Scaling is a method to normalize our independent variables. Scaling is essential for machine learning algorithms that calculate distance between data. For instance, most of the classifiers calculate the distance between two points by the distance. If one of the features has large value then distance consider this particular feature. Scaling is necessary, where large and small values present in variables.



Here, we use standard scaling method, which transform data with a mean value of zero and standard deviation of 1.

**Principal Component Analysis (PCA) :**

Principle component analysis is a technique of reduce the dimensionality of the variables. PCA is reducing the dimensionality of the large dataset, by transforming large set of variables into small set of variables. Reducing the number of variables affect our accuracy, But the trick is trade a little accuracy for simplicity. We use this method because smaller datasets are easier to explore and visualize, also make process faster for machine learning algorithms. So, idea behind the PCA is reduce the number of variables of a dataset, while preserving as much information as possible. Below is the code of PCA method that we use in this dataset:



We already have 44 independent variables but by using principle component analysis we convert those 44 variables into 31 variables.

**EDA concluding remarks:**

Here, EDA process is complete for our dataset. In conclusion, we plot different types of graphs for better data understanding and then, by using different types of algorithms and methods, we clean and prepare data for further process or machine learning algorithms.

* **Pre-processing Pipeline:**

Here, we revise topics which we are covered above:

1. Data preparation: In this section, we load dataset, check null values present in data and check data types of each variable.
2. Data Analysis: In this section, we analyze data with attrition attribute in graphical format and do some other analysis which help in batter understanding of data.
3. After that, we drop unnecessary columns because those columns not affect our target variable.
4. Encoding: In this part, we convert categorical data into numerical data by using label encoder and dummy columns, because machine learning algorithms only understand numerical data.
5. Handling Outliers: We use two methods for remove outliers but both methods remove large percentage of data. So, we do not remove outliers.
6. Scaling: We done scaling for normalize our data. In this we use standard scaler method, which convert data with mean = 0 and standard deviation = 1.
7. Principle component analysis (PCA): We use this method for reduce number of variables and make process faster for machine learning algorithms.

* **Building Machine Learning Models:**

The model building process consists of selecting model that is based on various machine learning methods used for experimentation. In this process we trained machine learning algorithm to predict the labels from the features. The goal is to find the best classifier model for analyzed problem and classifier with best classification results is used for prediction. The classification algorithm that we used is:

1. Logistic Regression
2. GaussianNB
3. Decision tree classifier
4. SVC (Support Vector Classifier)
5. K neighbors classifier
6. Random forest classifier
7. Ada boost classifier
8. Gradient boosting classifier
9. Xg boost classifier

We need to train the model to classify new observations. For training and testing we divided our original dataset into two parts training set and testing set. The accuracy of the machine learning algorithms increases with the amount of data using during training. The original data set divided into two parts with 80:20 ratio. 80% data is used for training and 20% data is used for testing. Code for splitting data is as below:

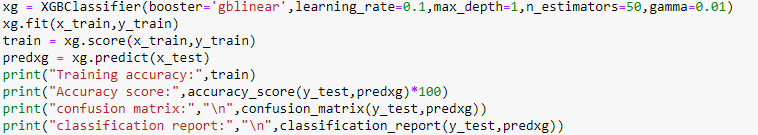


Model allow training set to find relation between variables and based on that model predict test result. Here, we use stratify sampling method during splitting train and test data because our target variable is imbalanced classified. Stratify sampling method take same percentage of samples of each target class as the complete set.

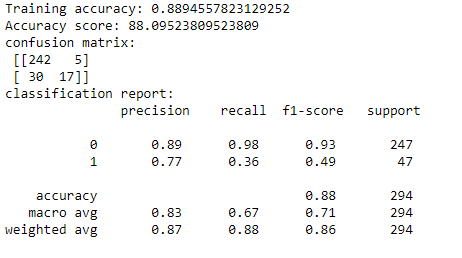
* We use score technique, which provide training set accuracy and also use accuracy score, confusion matrix and classification report which provide testing accuracy. We use lots of algorithms but here, we describe only best model:
* We find our best model is Xgboost classifier. Xgboost belongs to a family of boosting algorithms that transform weak learners into strong learners. We use Xgboost classifier with gridsearchCV. GridsearchCV is used to find best parameters from algorithms and then we apply those best parameters in particular algorithms. Below is the code for finding best parameters by using gridsearchCV:

 Output:

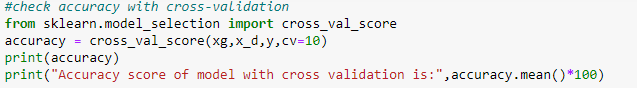
* By using gridsearchCV with xgboost, we get output with best parameters and apply these parameters in xgboost classifier.



Output:



* After trying lots of algorithms we get good accuracy using Xgboost.
* Here, we check xgboost classifier accuracy using cross validation to check model is underfitted or overfitted.
* We use cv=10, that means our train set divided into 10 parts and then provide mean accuracy of those 10 parts.



Output:



* We see that model is not underfitted or overfitted with training data.
* Here, we are interested in who left the company, So let’s check accuracy of each algorithms who left the organization:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithms | Train accuracy | Test accuracy | Precision | Recall | F1-score |
| Logistic Regression | 0.89 | 0.87 | 0.71 | 0.36 | 0.48 |
| GaussianNB | 0.86 | 0.79 | 0.31 | 0.23 | 0.27 |
| Decision tree classifier | 0.100 | 0.72 | 0.22 | 0.30 | 0.25 |
| SVC | 0.91 | 0.86 | 0.80 | 0.17 | 0.28 |
| K neighbors classifier | 0.87 | 0.85 | 0.75 | 0.13 | 0.22 |
| Random forest classifier | 0.100 | 0.85 | 0.86 | 0.13 | 0.22 |
| Ada boost classifier | 0.88 | 0.86 | 0.65 | 0.32 | 0.43 |
| Gradient boosting classifier | 0.97 | 0.83 | 0.44 | 0.17 | 0.25 |
| Xg boost classifier | 0.88 | 0.88 | 0.77 | 0.36 | 0.49 |

* **Conclusion :**
* To conclude, first we assess the data statistically and then we classified them. The dataset is processed, dividing into train set and test set. Then, we selected various classification algorithms, which provide different results and able to evaluate best classification result by using xgboost. By using xgboost we predict that whether an employee was likely to leave the company.
* Result predicted by the best algorithm depicts that the main attrition variables are Job satisfaction, monthly income, age and distance from home.
* An employee’s attrition value prediction help organization to know when they are going to wrong and which type of changes they required.