

# Human Resources Analytics

Understanding the Attrition in Human resources

Human Resource is a department of an organization that deals with the hiring, training, administration, and benefits of the employees. An organization hires several employees every year and invests lots of time, money, and resources to increase the performance and effectiveness of an employee. An organization would benefit more if it can attract and retain the right employees and the Human Resource department needs to know what are the factors responsible for hiring and retaining important and good employees.

# **Introduction**

## HR Analytics

HR Analytics is an analytical study that is applied to the human resource department of an organization to understand the employees and increase their performance to get a better return on the investment made on them. With the help of HR Analytics, one can get an overall insight as per the gathered data and can make the right decisions for the employees which in result would benefit the organization.

## Attrition

Attrition in human resources can be referred to as the voluntary or involuntary reduction in the workforce of an organization due to any internal or external factors. HR professionals are the ones who take an initiative in designing an organization’s work program, benefits, and other different factors that help in retaining the employees. A major problem due to the higher attrition rate is the cost of losing an employee and again hiring a new one by spending more money on all the initial hiring and training processes, due to which it becomes important for an organization to understand the factors leading to high attrition rates.

## Understanding attrition in HR

The main aim of this project is to understand the attrition in HR analytically and then coming up with a solution to the problem with the help of machine learning. Nevertheless, it will also explain the major factors responsible for attrition so that the HR department can make the right decisions based on the analysis.

What is Machine Learning?....................

Machine learning is the science of getting computers to act without being explicitly programmed with the use of Data. It involves computers learning from past data to make predictions and decisions on their own.

## Problem statement :

Due to Gradual loss in the workforce, the organizations need some methodologies to understand the factors responsible for higher attrition rates so they work on designing and planning the right work programs and benefits so that they can retain good employees.

To solve this problem, we have been provided with Attrition of an employee as yes or no for several employees and multiple factors and information associated with that employee using which we aim to build a model which predicts the attrition of an employee using several input features.

## Dataset :

The dataset given can be found from the following link :

<https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics>

The dataset contains different features and information on the employee which are listed below :

1. Age - The age of the employee ranging between 18 to 60 years
2. Attrition: Attrition is the target feature using which we will predict the attrition of the employees.
3. BusinessTravel: Business travel gives an insight on the traveling done by the employee, which has been categorized as Travel\_Rarely, Travel\_Frequently Non-Travel.
4. DailyRate: The rate of the amount paid daily
5. Department: Department of the employee
6. DistanceFromHome: Employees’ distance from home
7. Education: Education of the employee
8. EducationField: Field of education the employee comes from
9. EmployeeCount: Count of employees
10. EmployeeNumber: EMPLOYEE ID
11. EmployeeSatisfaction: Satisfaction of the employee with the environment
12. Gender: Gender of employee male or female
13. HourlyRate: Hourly salary of an employee
14. JobInvolvement: Job involvement of the employee
15. JobLevel: Level of the job of an employee
16. JobRole: Job role of the employee
17. JobSatisfaction: Satisfaction level with the job
18. MaritalStatus: Marital status of employee
19. MonthlyIncome: Monthly salary
20. MonthlyRate: Monthly rate
21. NumCompaniesWorked: Number of companies worked in
22. Over 18: Overage of 18 yes or no
23. Over Time: Overtime categorized as yes or no
24. Percent Salary Hike: Percentage of salary hike
25. PerformanceRating: Performance rating of employee
26. RelationshipSatisfaction: Relationship with the employer
27. StandardHours: Standard hours worked
28. StockOptionLevel: Stock options given to the employee
29. TotalWorkingYears: Total years worked
30. TrainingTimesLastYear: Hours spent in training
31. WorkLifeBalance: Time spent between work and outside
32. YearsAtCompany: Total number of years at the company
33. YearsInCurrentRole: Total years in the current job role
34. YearsSinceLastPromotion: Years since last promotion
35. YearsWithCurrManager: Years spent with current manager

More info on the following columns:

Education :- 1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor'

EnvironmentSatisfaction :- 1 'Low', 2 'Medium', 3 'High', 4 'Very High'

JobInvolvement :- 1 'Low', 2 'Medium', 3 'High', 4 'Very High'

JobSatisfaction :- 1 'Low', 2 'Medium', 3 'High', 4 'Very High'

PerformanceRating :- 1 'Low', 2 'Good', 3 'Excellent', 4 'Outstanding'

RelationshipSatisfaction :- 1 'Low', 2 'Medium', 3 'High', 4 'Very High'

WorkLifeBalance :- 1 'Bad', 2 'Good', 3 'Better', 4 'Best'

## Contents of the Article:

The article explains all the steps end to end to understand the attrition through exploratory data analysis and then build a machine learning model to predict the attrition of an employee.

Following are the steps that are taken to build this project which we will be covering in this article:

1. Exploratory Data Analysis
2. Feature Engineering and preprocessing
3. Analyzing Feature Importance
4. Model Building and Evaluation
5. Hyperparameter Tuning
6. Saving the model for predictions and deployment
7. Conclusion

So let’s start with the exploration of the dataset and build a good prediction model.

# Exploratory Data Analysis

Exploratory data analysis is a statistical approach to analyze data and summarise the main characteristics of the data which gives us insight into the business or any use case or problem statement.

In EDA we generally explore the data for the below-mentioned point:

- General info about the dataset

- Missing values

- Analyzing types of variables

-> Numerical variables

-> Categorical variables

- Visualization

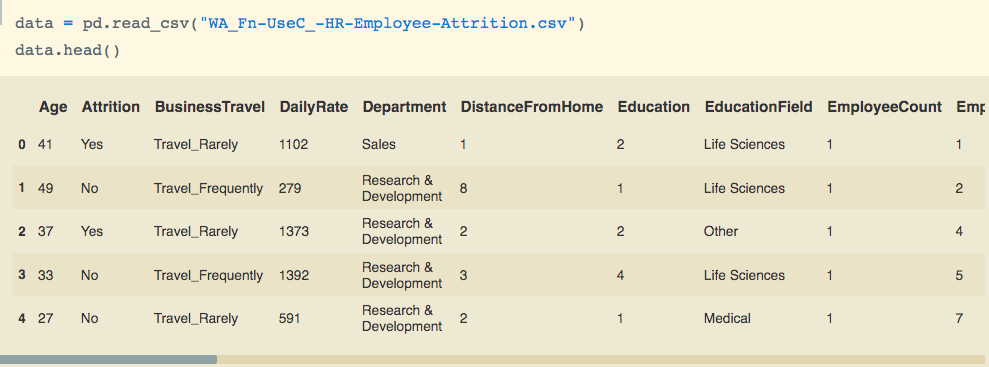
- Detecting outliers

- Distributions of variables

- Relationship between independent and dependent features

To begin with first will be importing libraries important for EDA and other important libraries required in this project.

After importing the libraries will load the dataset and have a look at the head of the data -

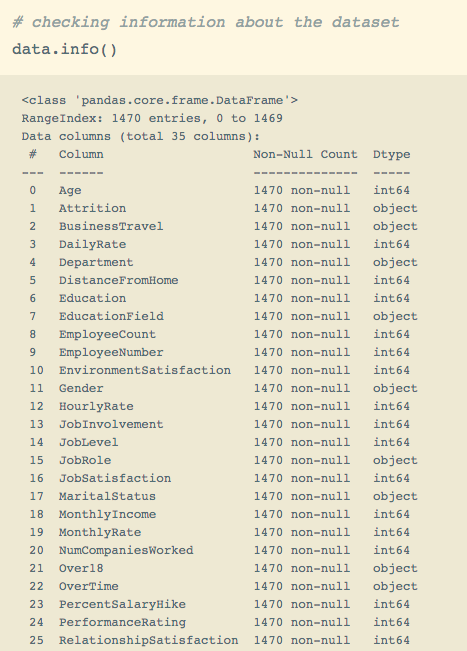


After looking at the first 5 records of data we can identify the following points:

1. The data has two types of features numerical and categorical
2. Also, there are majorly discrete numerical features

Let’s move further into the exploration to get more information about the data.

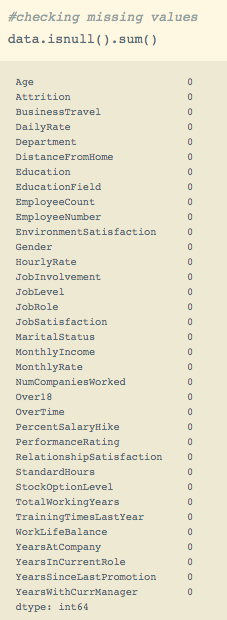
1. Some General info -



We can use the command data.info() to get the information on the data which includes, the total number of rows and columns, data types of variables, number of non-null values in the dataset, and the memory usage as well.

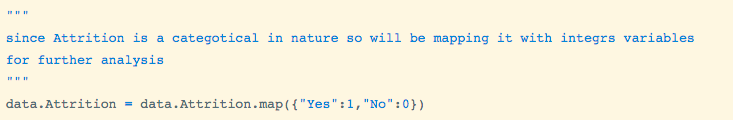
From the info of the dataset, we can see that there are a total of 35 features, Attrition being the target feature and 26 variables are integer type variables, and 9 are object type variables.

One of the major EDA processes is to detect and treat the missing values as they reduce the fit of a model or can lead to biased models. Moving further will check if the dataset has missing values.



As the output is zero for all the columns there are no missing values in the dataset. We can also check the missing values by visualizing them on heatmaps.

Moving ahead since the target feature “Attrition” is a categorical variable and we have to analyze relationships with it so will encode its categories by mapping them with integer values, due to which we will also be able to check its correlation with numerical features.

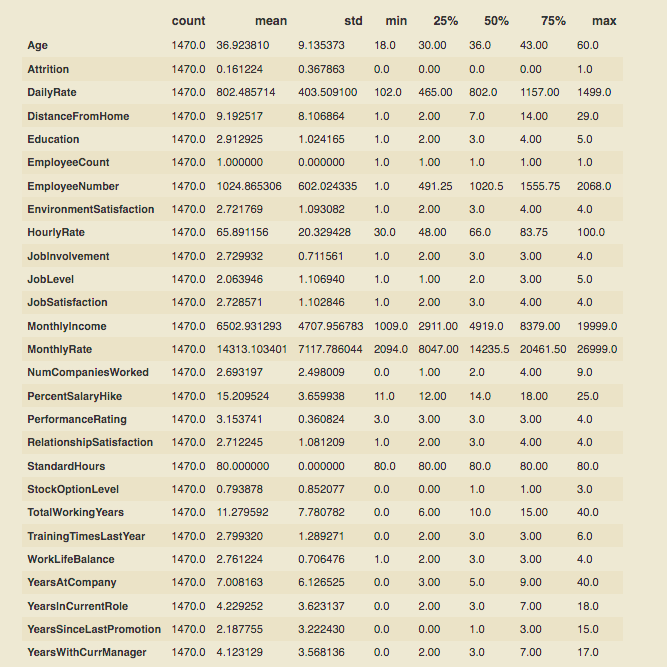


Next, will be checking the statistical summary of the numerical features:

We can check the statistical summary with the help of command data.describe()-



Here I am using “.T” with the command to get the transpose of the data frame for a better analysis.

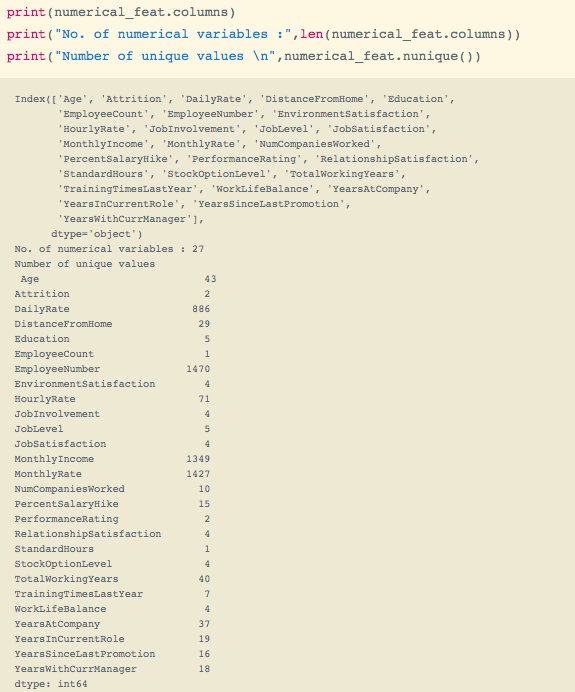


From the statistical summary, we get the statistical significance of the data, that is we get to know the central tendencies - mean, median, mode, and spread of the data standard deviation and percentiles including the min and maximum values through which we can get to know the range of the data.

Let’s explore the categorical and numerical variables in detail with visualizations to understand them better and how they affect the attrition of employees.

We will first start with exploring numerical variables

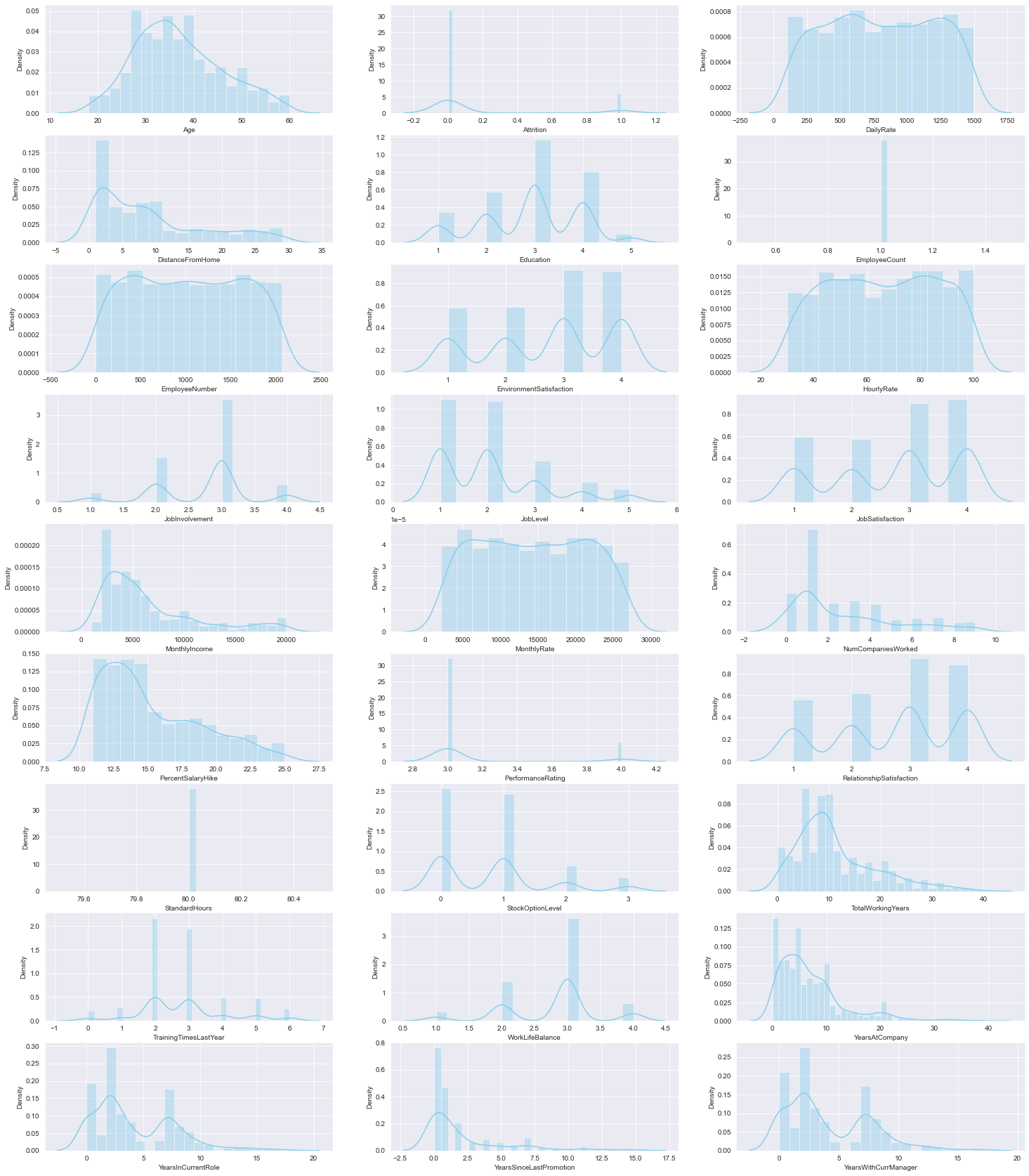




We can observe that:-

1. There are a total of 27 numerical variables
2. Monthly income, daily rate, employee number, and monthly rate are continuous
3. The rest of the variables are discrete since they are countable and are the rankings of some features as given in the ‘Data’ section of this article

Now let’s see how the numerical variables are distributed by using a distribution plot method (sns.distplot()) from the seaborn package.



We can make the following observations from the numerical data –

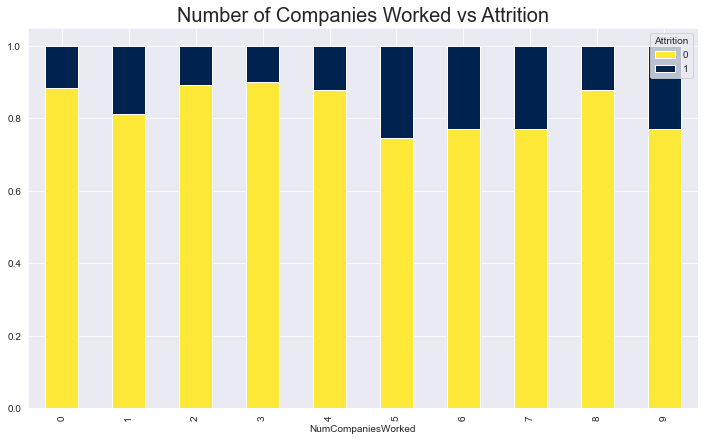
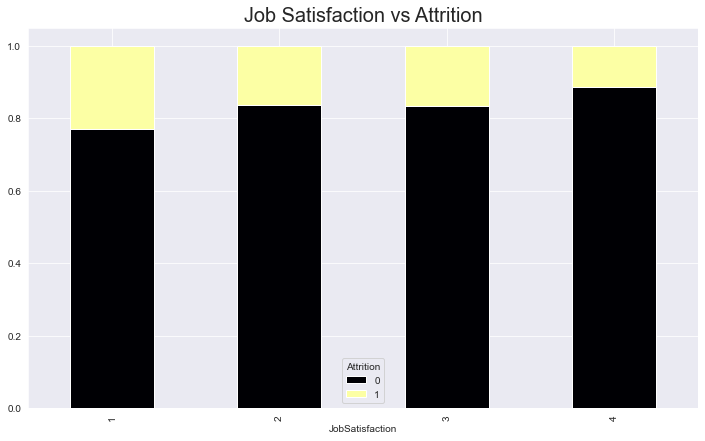
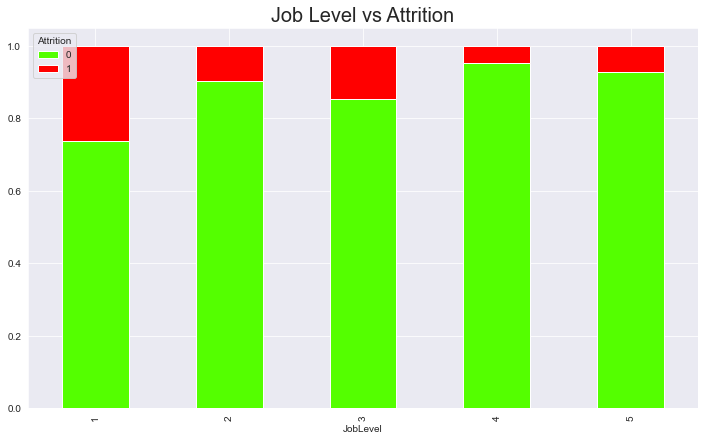
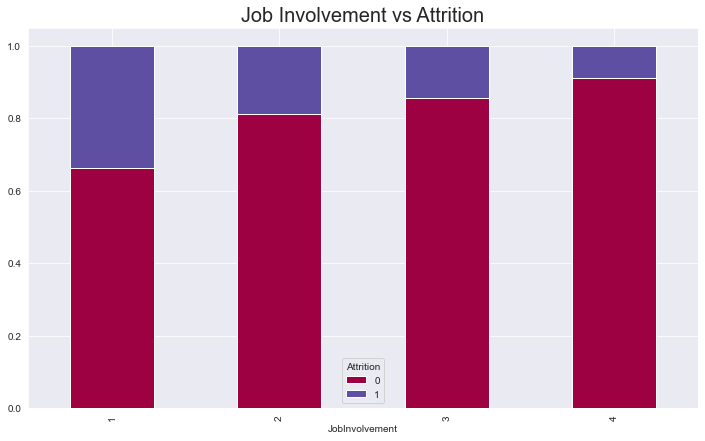
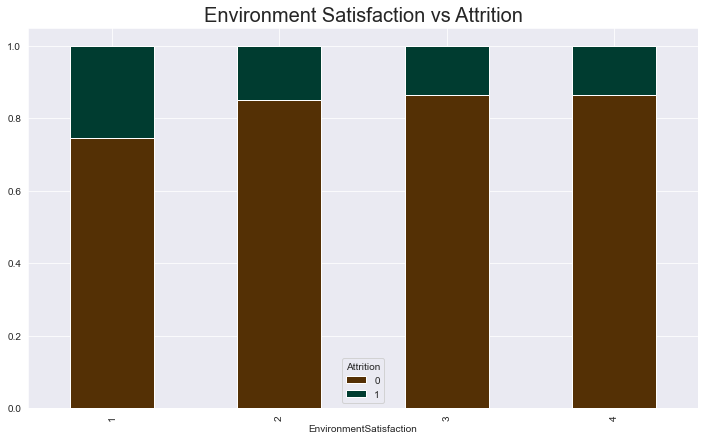
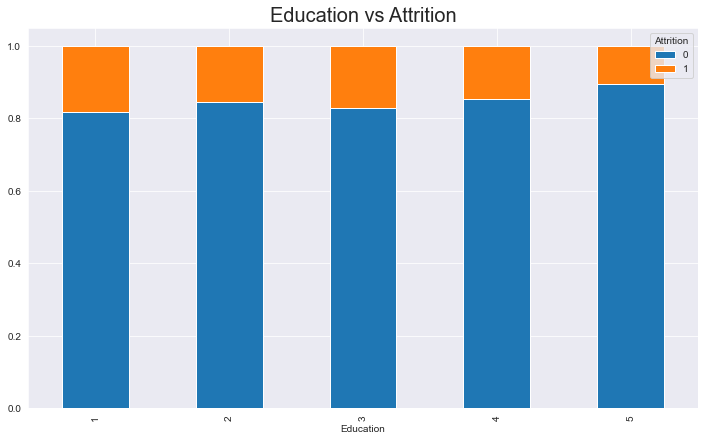
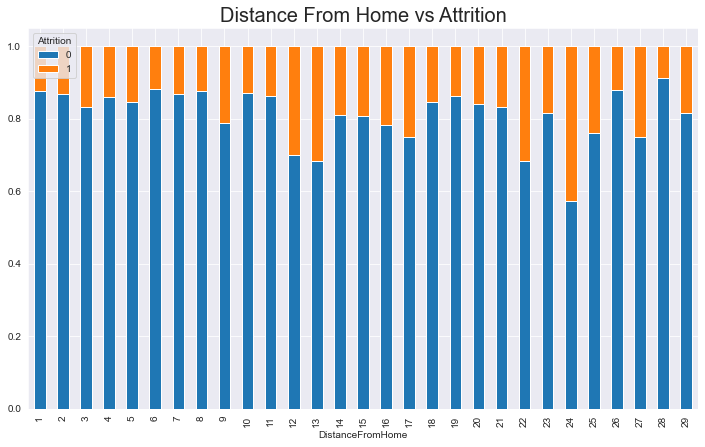
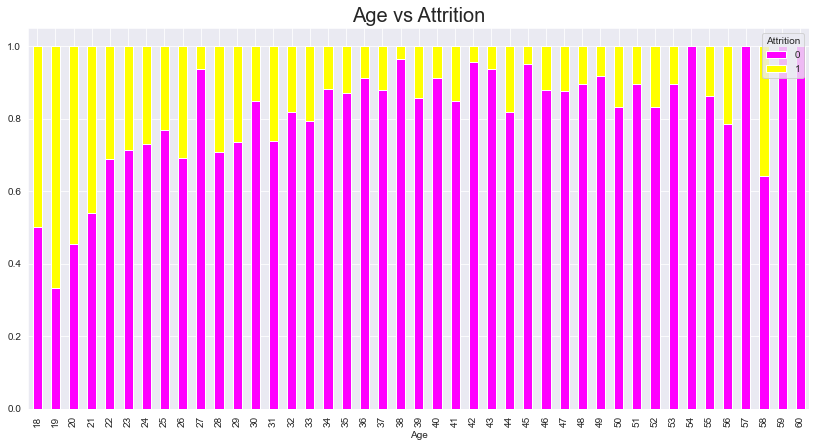
1. Most of the employees are either in their ’30s or 40’s age
2. Most of the employees have an education of level 3
3. Majorly customers rated the Environment satisfaction as 3 and 4
4. Most of the customers have job involvement of level 3
5. Most of the employees are from job levels 1 and 2
6. Mostly the employees rated job satisfaction as 3 and 4
7. Mostly the employees have worked in 1 organization followed by worked in 0 organizations
8. Most of the employees have a performance rating of 3
9. Majorly employees either have no stock options or level 1 stock option
10. The majority of employees have work experience of 5-10 years
11. Most of the employees have a work-life balance rating of 3

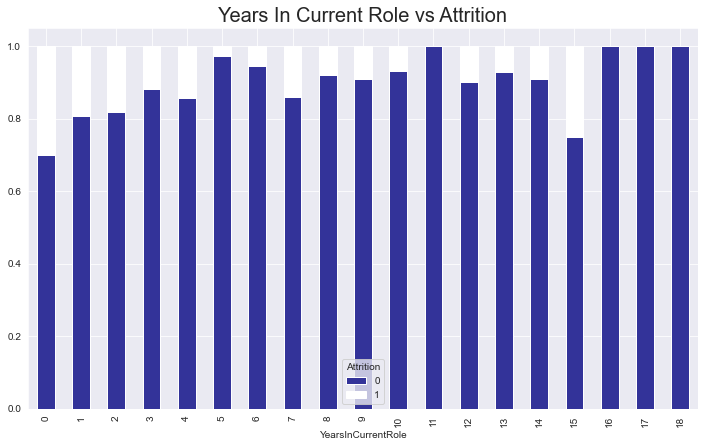
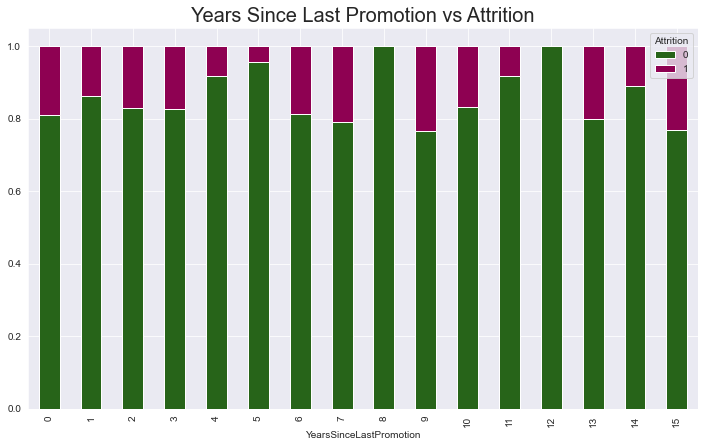
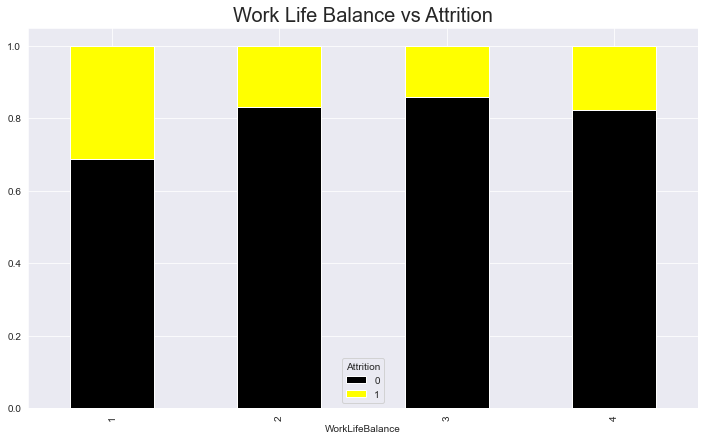
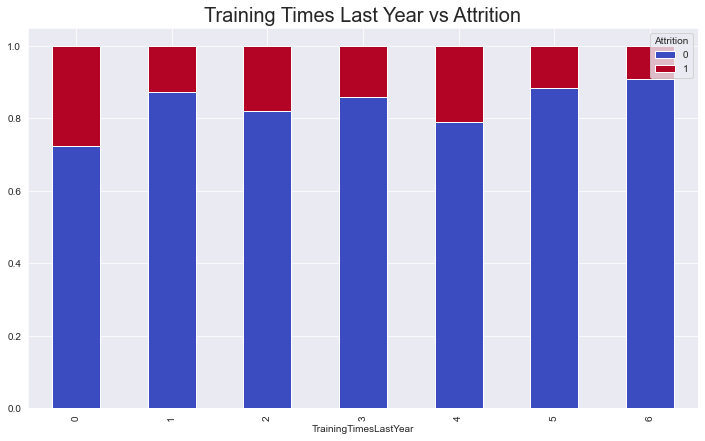
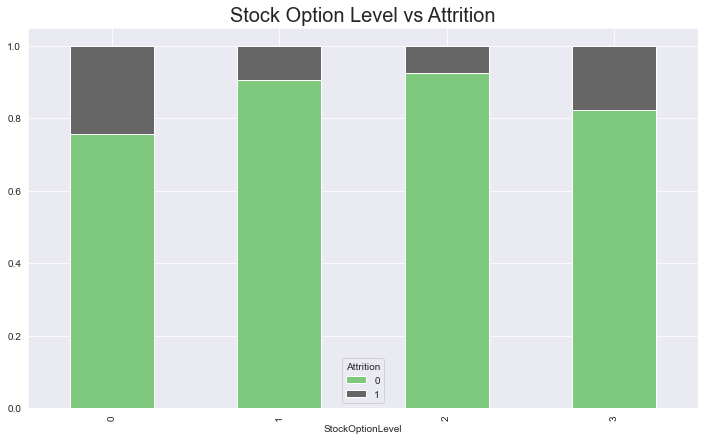
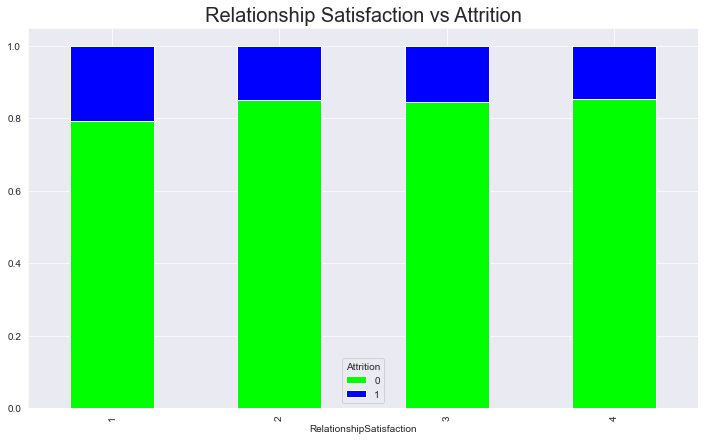
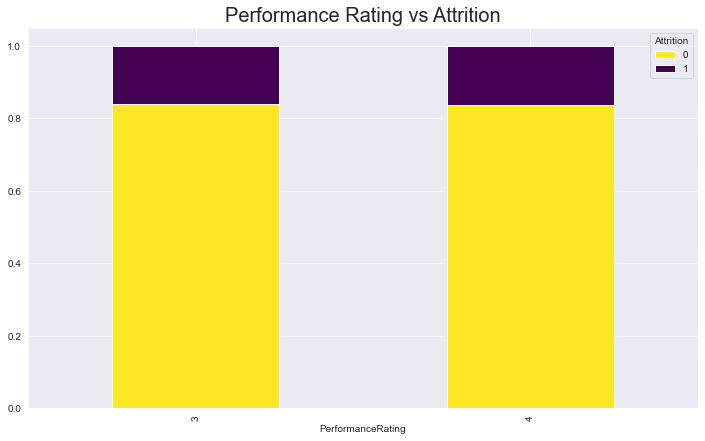
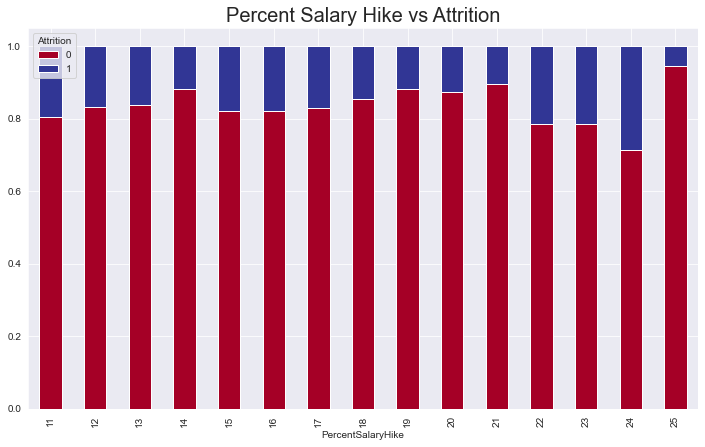
Also, let’ visualize the target variable :



We can observe there is a huge difference in the distribution of classes in the Target variable so this implies that there is an imbalance in the dataset. Due to an imbalanced dataset, it is biased to the majority class and gives a misleading accuracy score so will handle the imbalance of the dataset further in the feature engineering process.

Proceeding with analyzing the attrition with the numerical features will help in understanding the factors responsible for attrition.

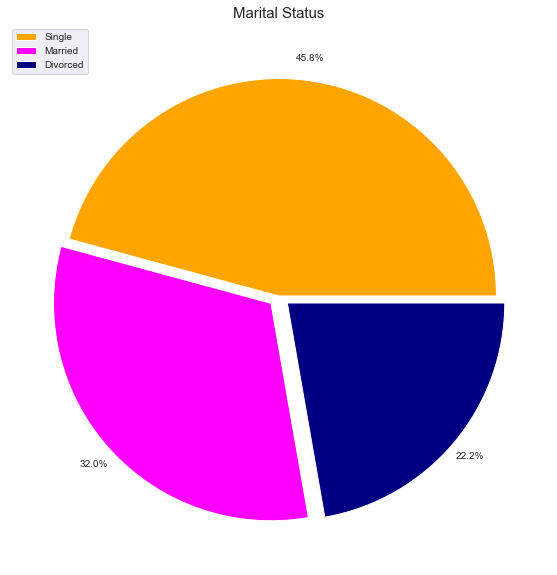
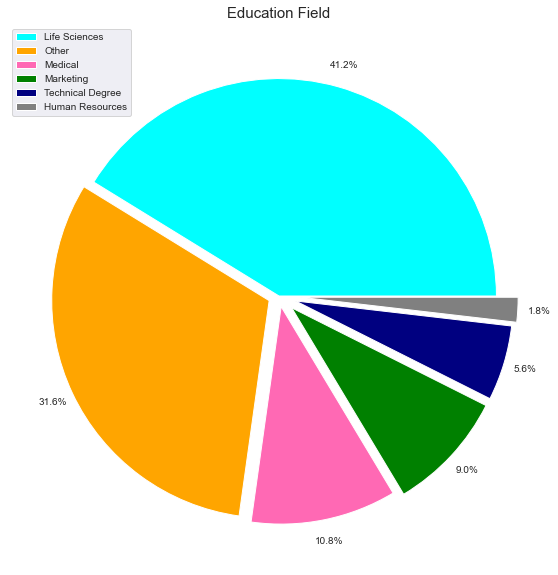
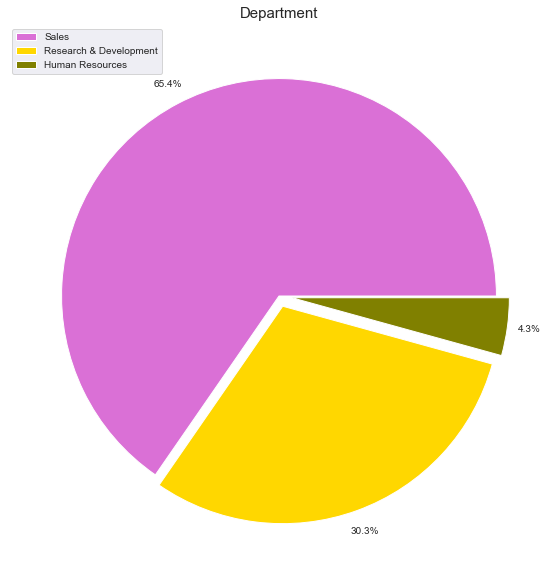
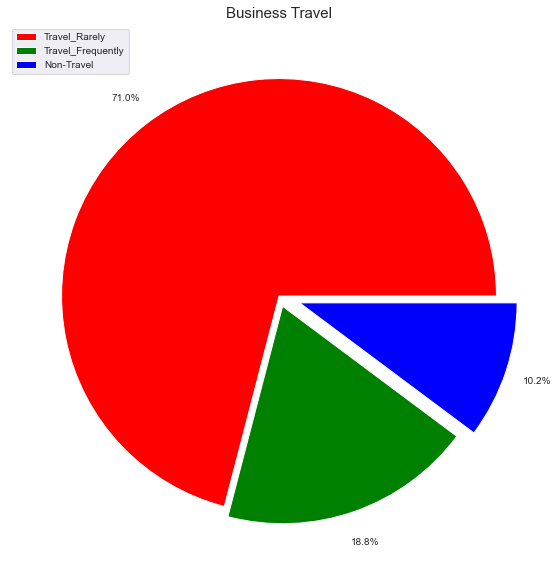


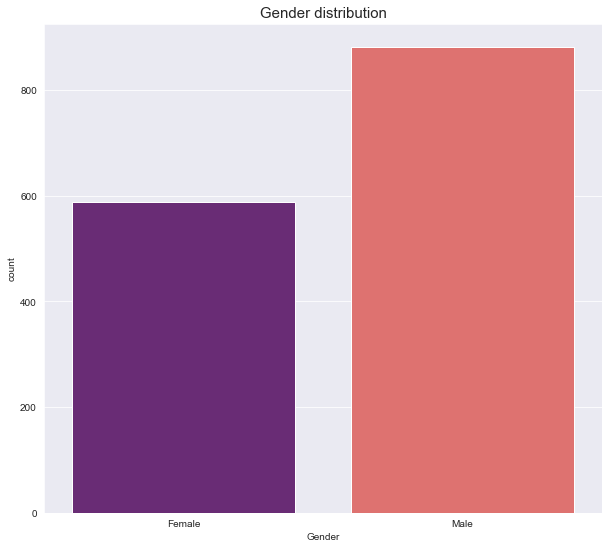


These plots are stacked bar charts where the upper color code represents the attrition of employees. The legend shows the color codes which have 0 as no attrition and 1 as attrition. We can make below conclusions from the plots above :

* Usually, employees who are younger or specifically below the age of 30 have higher attrition than the older employees.
* Attrition is majorly higher for employees who have a larger distance from home, i.e. if an employee is living nearby there is less chance for him or her to leave the organization.
* Attrition is higher in the case of education levels 3 and 1 and is the least in the case of level 5 education. Hiring should be done more for higher education level candidates.
* Attrition is higher for Environment Satisfaction level 1 and least for level 4. Low environment satisfaction can lead to higher attrition.
* Attrition is higher for Job Involvement level 1 and least for level 4 which means employees who are more engaged in their jobs have low chances of leaving.
* Attrition is higher for level 1 jobs and level 3 jobs and least for job level 5 i.e. employees with higher job levels are less likely to leave the organization.
* Attrition is higher for job satisfaction level 1 and least for level 4, which implies employees with a higher level of job satisfaction are less likely to leave the organization.
* Employees who have worked in more than four companies have a higher attrition rate i.e. if an employee has worked in more than 4 organizations he/she would likely leave the company.
* The attrition rate seems similar for performance ratings of 3 and 4.
* Attrition is higher for level 1 relationship satisfaction and least for level 4, which implies that the individual with a very good relationship with the employer has low chances of leaving the organization.
* Attrition is higher in the case of 0 level of stock options and lesser in the case of level 1 and 2, which implies the employee stays in the organization if they are given the options to buy stocks
* Attrition is high for a work-life balance of level 1, or we can say that employees with low work-life balance are more likely to leave the organization.
* Attrition is higher if years in the current role are more, but it can vary for 15 yrs there is a good amount of attrition and for 0 years the attrition is high. So we can imply that if the employee is in the current role for less than five years attrition will be higher and the same implies if the employee is in the current role for more than 10 yrs.
* Employees who have not been promoted since last 8 yrs have a significant amount of attrition whereas if the last promotion is more than 8 yrs then the attrition can be higher but collectively attrition can vary
* As time passes with a manager the attrition goes down, attrition is high for employees who have fewer years with the analyzing current manager.

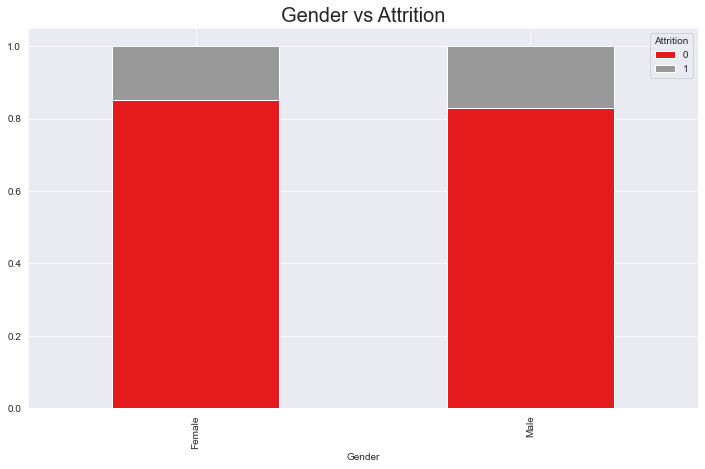
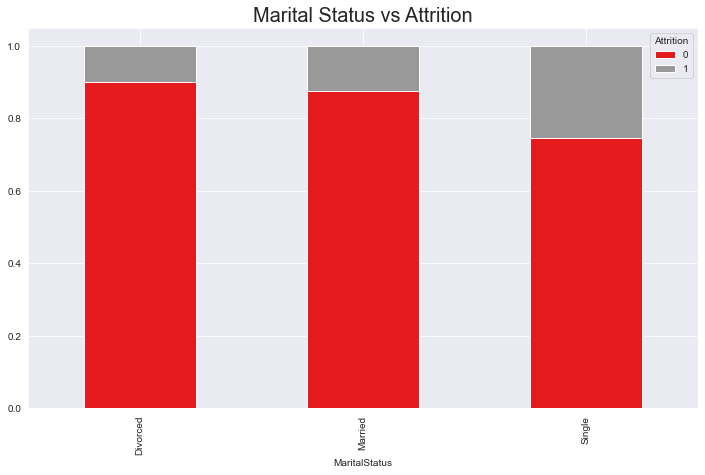
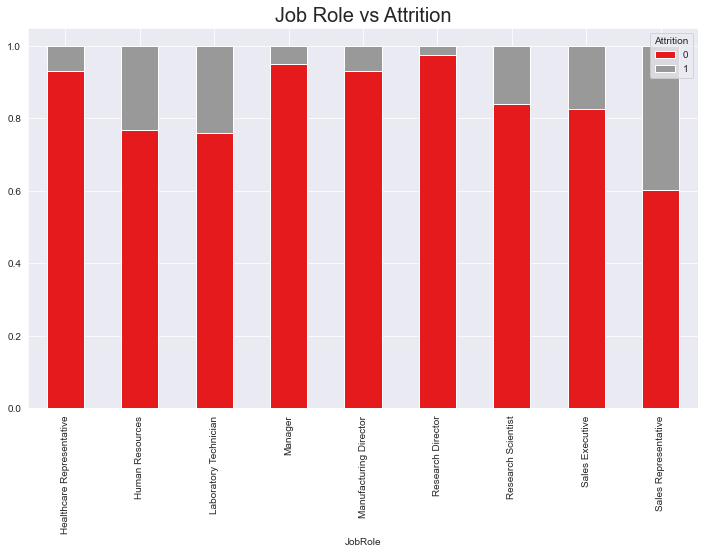
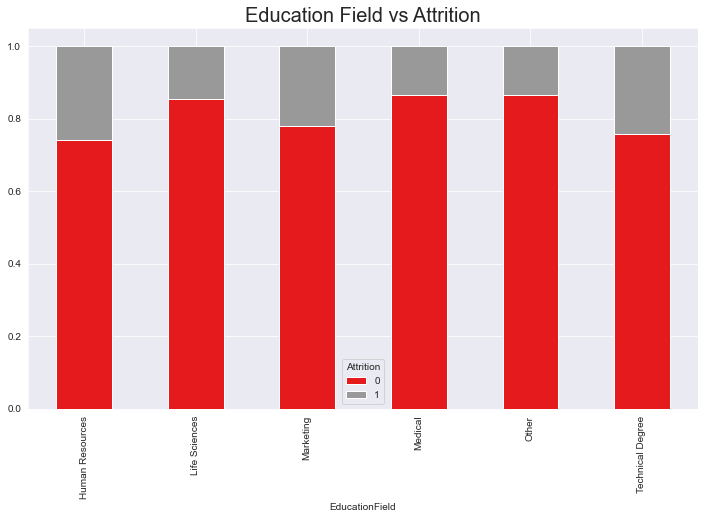
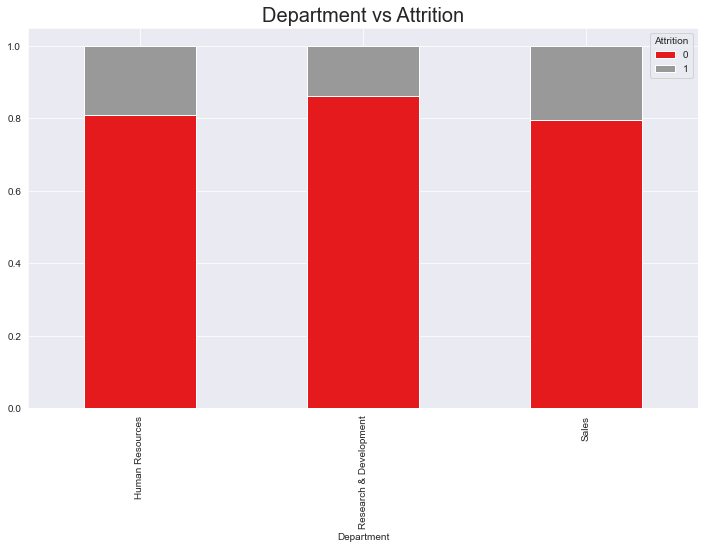
Will now analyze the distribution of categorical variables and their impact on attrition using pie charts and stacked bar charts.





From the above plots, we can make the following conclusion about the class distribution -

* Major no of employees travel rarely, employees who do not travel are very less in number.
* Sales department makes the major part of an organization, research and development make about 30 percent of an organization, Human resource being the least contributor to the strength.
* Life sciences have the highest distribution and Technical degree and Human resource being the least which means more number of employees are from life sciences education background, whereas from technical and human resources education background employees are lesser in number.
* The major number of employees are from sales executive job roles and the least are from human resources.
* The majority of employees are unmarried.
* As per the above plot, the count of males is more than females in an organization.



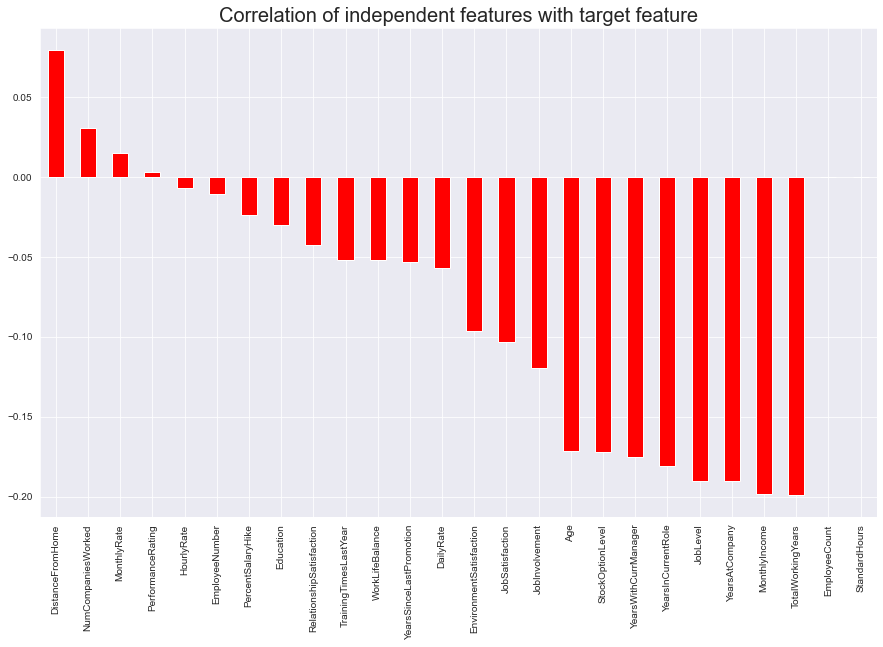
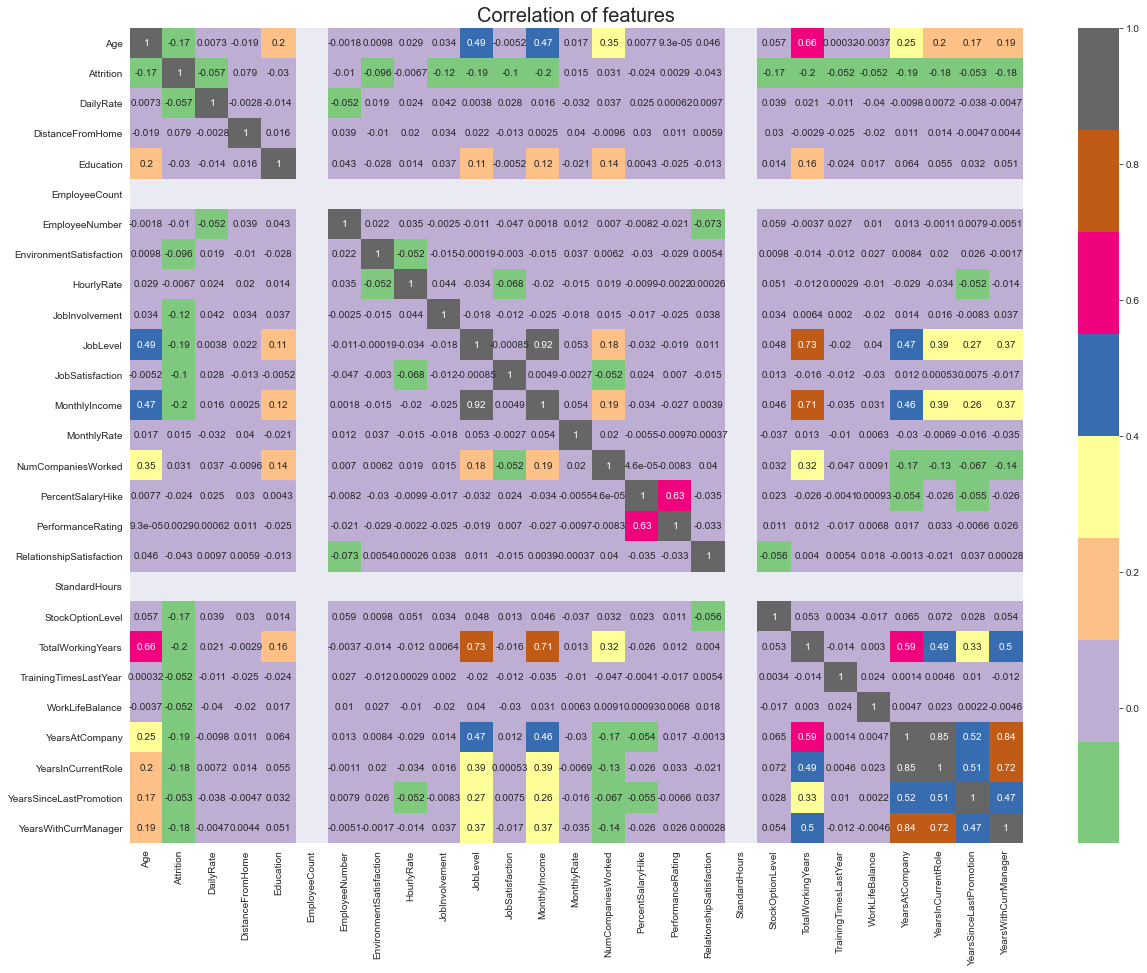
From the above plots, we can conclude the following points -

* Attrition is higher for frequently traveling candidates and less for non-travelers, or we can say that employees who travel frequently are more likely to leave the organization.
* Sales and Human resources have a higher attrition rate whereas R&D has a lesser rate, which means employees from the R&D department are less likely to leave the organization.
* Attrition is higher from HR educational background whereas Medical and other educational background being the least.

Employees from Medical and other educational backgrounds are less likely to leave.

* Attrition is highest for Sales representatives, human resources, and laboratory Technicians, which means employees from these job roles have a higher chance of leaving the organization.
* Attrition is higher for unmarried employees and least for divorced, which means unmarried employees are more likely to leave the organization.
* Attrition is more for male employees, which means males are more likely to leave the organization.

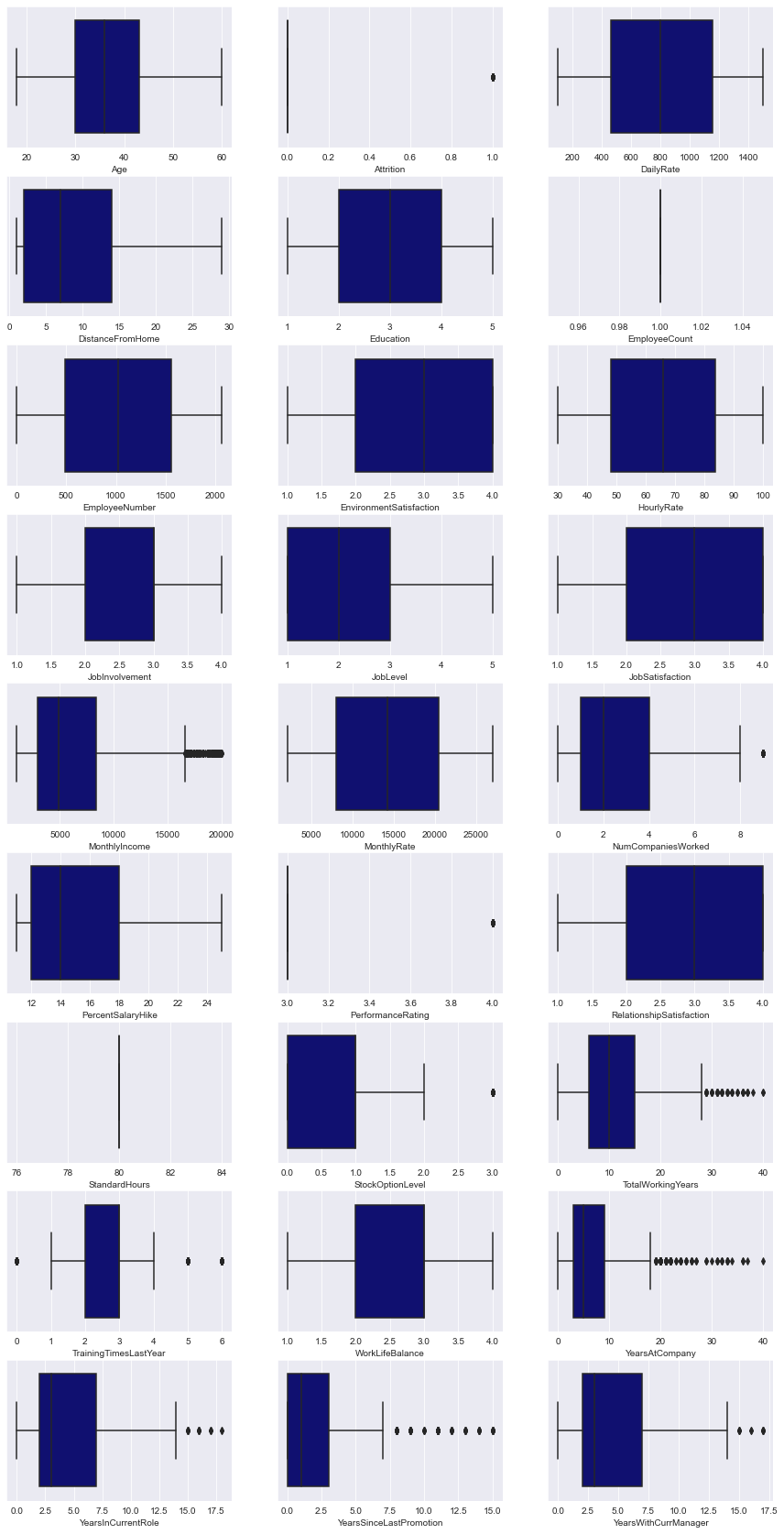
As we have done some analysis on the distribution of features and how the features affect the attrition rate, let’s check the strength of the relationship between the variables using correlation which will give an insight on which features are related to each other, and also we will visualize the strength of the relationship between the target and predictor variables. Correlation is a statistical measure of the strength of association between two variables. For determining the relationship we will use the Spearman correlation technique and will plot it on a heatmap to understand the strength visually in a better way. Spearman correlation considers the ordinal variables.



From the above visualizations we can conclude the following point:

* The above plot gives visual information on the correlation of variables, also some of the variables are correlated to each other which also depicts multicollinearity. The white spaces depict that there is no correlation between the variables.
* Distance from home is the highly positively correlated feature and Total working years is a highly negatively correlated feature concerning attrition.

Further, we will proceed with detecting outliers in the dataset. Outliers are the abnormal points that lie far away from normal observations. Treating an outlier depends on the use case and there are different ways to handle and detect the outliers. Here the method I am using to detect outliers is by visualizing them with boxplots.



As we can see that there are many outliers in the dataset, however, the outliers here majorly are because of the discrete variables which imply that these are natural outliers and we will keep them as they are important for model building and prediction, whereas, Monthlyincome and column might have artificial outliers which might be due to error in data collection.

Till now we have explored the dataset from all aspects and have analyzed their distribution and relation with target variables, so we will now move to feature engineering where we will do some pre-processing and engineering steps on the feature to send it further to model building phase for prediction.

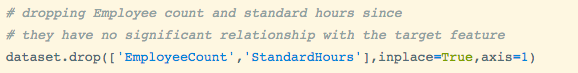
# Feature engineering

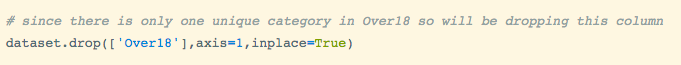
Feature engineering is the phase where more information is extracted from the dataset and pre-processing such as feature transformation is done on the distribution of the variables which improves the predictions.

Moving further in the feature engineering process :

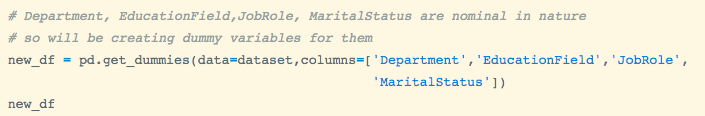
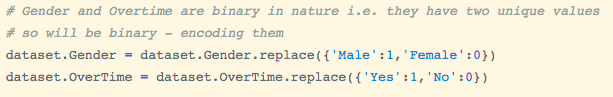
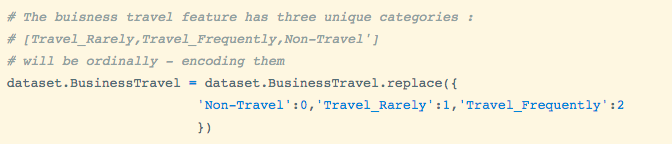
* Will be dropping some of the features with multicollinearity issue and unnecessary columns
* Imputation of missing values: Since we don't have missing values we won't have to perform this step
* Handling categorical variables
* Handling imbalanced dataset
* Scaling : Normalisation or Standardisation

As in the exploratory data analysis, we identified that MonthlyIncome, JobLevel and YearsAtCompany, and YearsInCurrentRoleare highly correlated we would have to drop one of the variables each among which the features having weak strength of relationship with the target variable since they have multicollinearity due to which there will be higher variance in the predictions of the model if some minor changes are made in the model. Also will be dropping some features which do not correlate with the variables which are EmployeeCount, StandardHours. Also will drop some features which are not important for prediction.

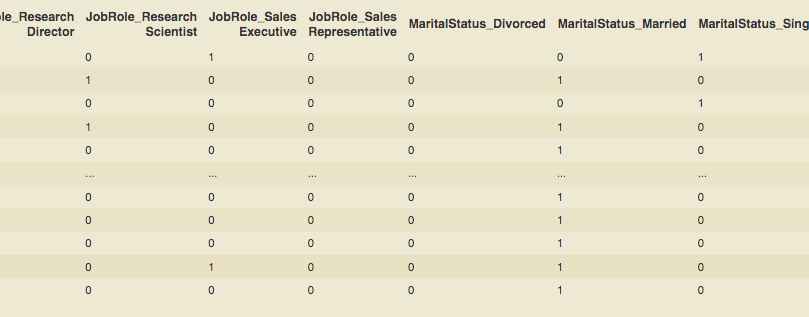
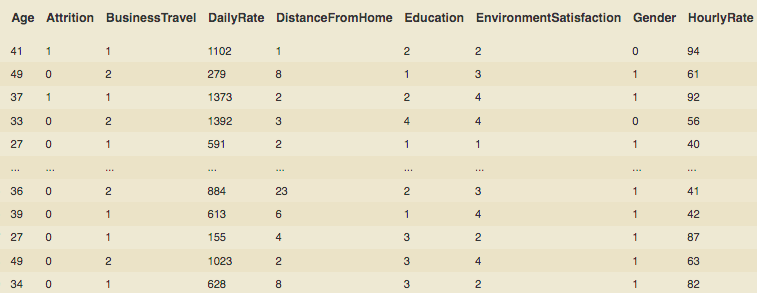




Now we will handle the categorical variable, where we will convert the categorical features into numerical features since the models do not work on object type datatype, however, we have a few models which do accept categorical variables but even they have the algorithm to convert the categorical features to numerical inbuilt. There are several ways to convert the categorical features to numerical and this process of conversion is known as ‘Encoding Categorical Data’.

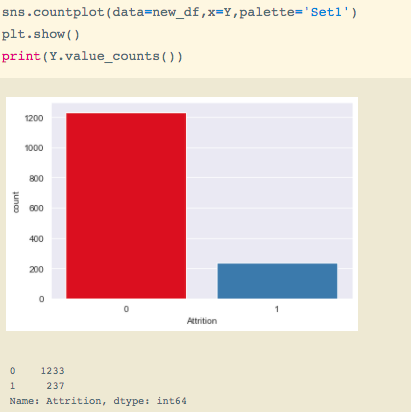
In our data we have Business Travel as an ordinal type variable so will encode it in ordinal categories and features with the binary class will be encoded with 0 and 1. The rest of the categorical variables will create dummy variables.

Our transformed data looks something like this:

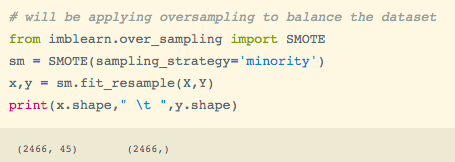


While exploring the Target variable we found that the data is imbalanced due to which there can be misleading predictions because the model will get biased towards the majority class. So will be handling the imbalance in the dataset by applying an oversampling technique SMOTE (Synthetic Minority Oversampling Technique), which creates synthetic data points of the minority class to match the distribution of the majority class.

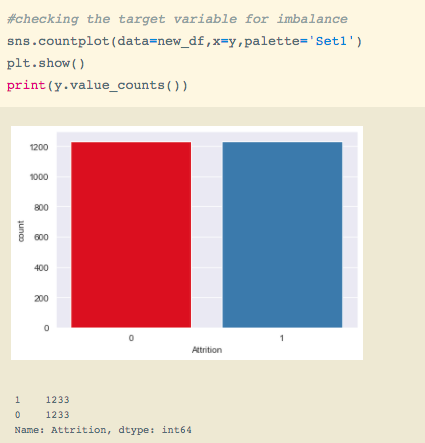
Before Oversampling the dataset -



Applying oversampling technique SMOTE:-



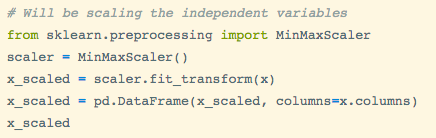
After oversampling -



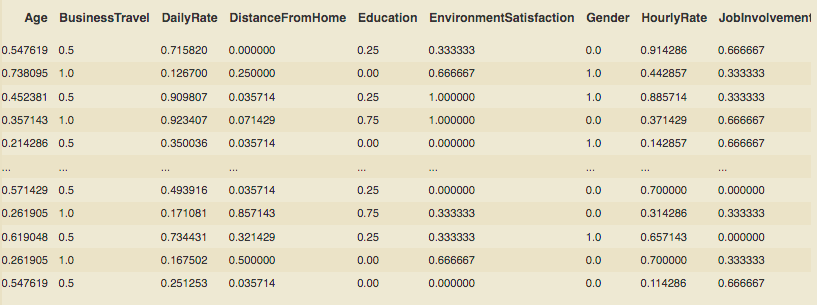
Now our dataset is balanced, so we will move further to normalizing the independent features.

Normalization or scaling is done to bring down the larger values and smaller values in the same range so that the machine learning model can learn better from the data for good predictions.

Here we will normalize the dataset. In normalization, we bring down the values in the range of 0 and 1 for which I will be using the MinMaxScaler method from sklearn.preprocessing class.



After normalizing our dataset looks something like this -



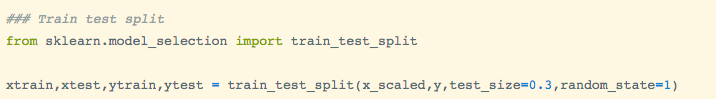
All the values have been normalized and brought down in the range of 0 -1.

# Analyzing Feature Importance

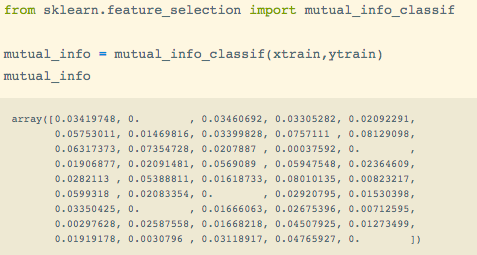
In this process as suggested by the name, we analyze the feature importance for the machine learning model to take into account the features that have valuable importance or are relevant for prediction. This step is usually done to retrain the model when it is not performing well with low accuracy scores.

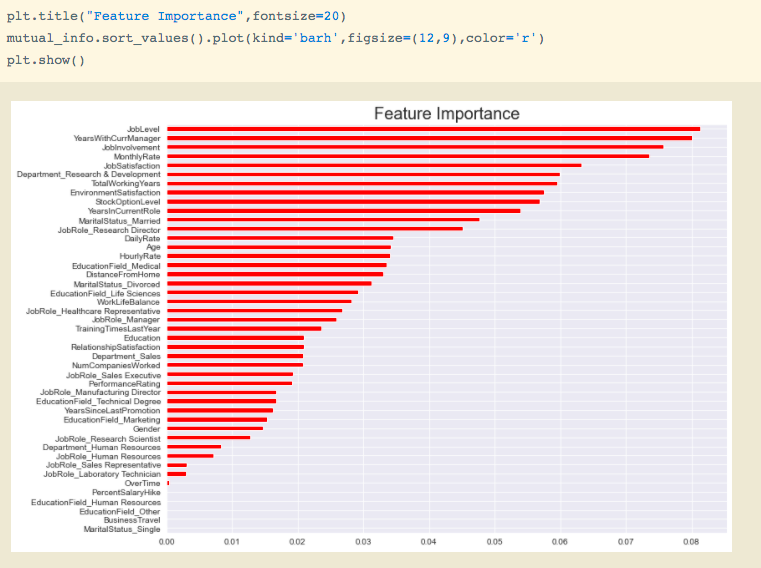
So I will be just plotting the Feature importance for a check and if my model does not perform good I will come back to this pipeline to select the best and relevant features for prediction.

For this, will I will be will first split the dataset into train and test data using train test split-



Now we will use the mutual\_info\_classif method from sklearn.feature\_selection class to get the best features. Mutual \_info\_classif works on the principle of information gain or we can say it's similar to information gain.



Visually plotting Feature importance - 

From the above visual we can see the importance of the features, from which we can select the best features to improve model performance.

# 

# Model Building and Evaluation

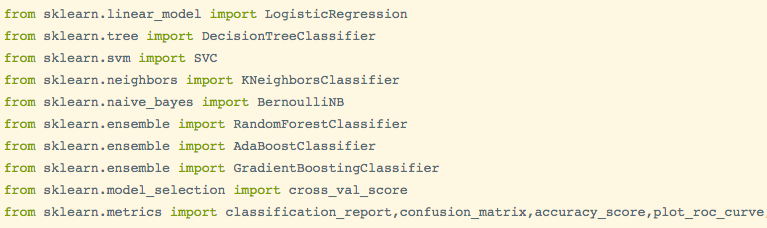
We will now proceed to the main step of our machine learning where we will fit the model and will get predictions as output.

Also, we will evaluate the model on different evaluation metrics -

* Accuracy score
* Confusion matrix
* Precision
* Recall
* F1 score
* Auc-Roc curve

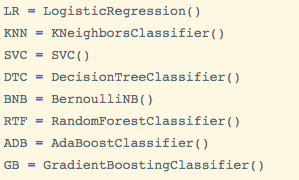
After this, we will validate the performance of the model using the cross-validation technique based on accuracy score to get unbiased evaluation results.

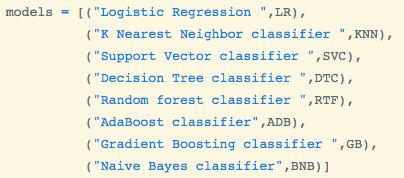
First will be importing important libraries for this process -



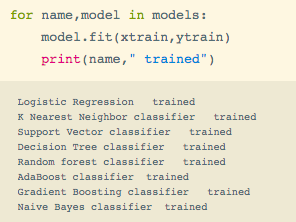
Here I will create instances of multiple models and will fit the models using a loop so that I can get the evaluation report of all different classifiers and further I will pick the model which will be performing best out of these and will further cross-validate that model.

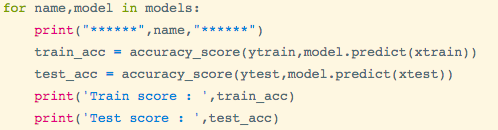
Instances of the model -





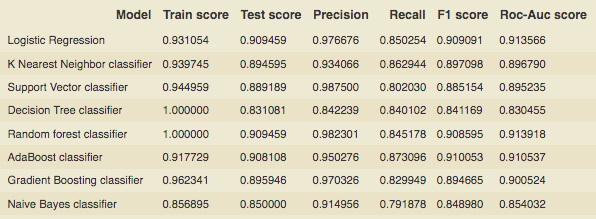
Further, I will train and evaluate these models using for loop -







Let’ get the report of all the models -

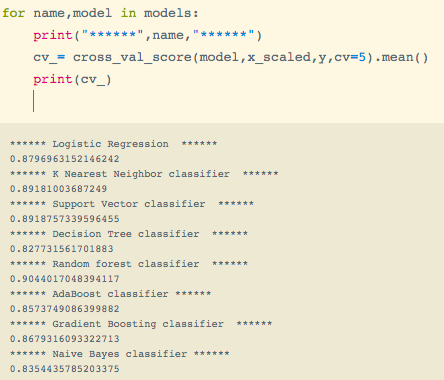


From the above report we can conclude the following point :

* Logistics regression and Random forest are giving the best performance among all models with the highest train and test accuracy scores and it is generalized with low variance.
* Precision being 0.976 is also quite excellent for logistic regression and 0.982 for Random forest. Precision is an evaluation metric that refers to the percentage of results that are relevant or we can represent it as (True Positives) divided by (Actual Results)
* Recall being 0.85 is quite excellent for logistic regression and 0.982 for Random forest 0.845. The recall is an evaluation metric that refers to the percentage of relevant results that are correctly identified by the model or we can represent it as (True Positives) divided by (Predicted Results)
* F1 score is 0.909 for logistic regression and 0.908 for Random forest where F1 score is an evaluation metrics that tell us how much the model is robust in predicting the precise results it can be also defined as the harmonic mean of precision and recall
* The roc-Auc score is 0.91 is highest for Logistic regression and 0.91 for Random forest, where the roc-AUC score tells us that at how much percentage the model can identify the different classes correctly which mean Logistic Regression is 91% correct in identifying the different classes from the target variable

Therefore we conclude that Logistic regression and Random forest are the best models among all the models model with excellent precision and recall, but we need to perform validation on the models so will move further to cross-validate the models to check if the model that we are choosing is not overfitting and is not selectively biased. Cross-validation takes the data that was not taken into consideration during the training process since the train test split separates the data into a random number of sample sizes.

Here I am using for loop to find cross-validation for all the models at once -



From the above report, we can conclude that Random forest is the best model because it has the highest cross-validation score among all the models since cross-validation resolves the overfitting issue so we can also say that it is the most generalized model with the highest accuracy of 90%. So will be classifying the random forest as the best model and will proceed further with tuning it for higher accuracy.

# Hyperparameter Tuning

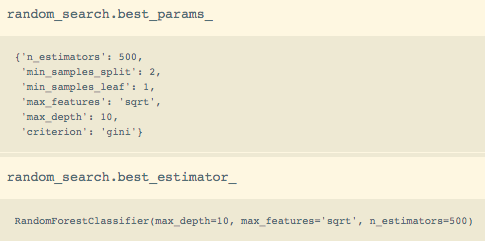
Hyperparameter tuning is a problem of choosing optimal parameters for a machine learning algorithm. The idea of this is to find out the best parameters for the model so that it can minimize the error and produce good results. There are two different methods to perform hyperparameter tuning:

* Grid Search: In this approach, the model is evaluated for a grid of hyperparameters with some range of values and it tries every combination of the parameter values to find the best hyperparameters. Since it searches through every combination throughout the grid it becomes computationally expensive.
* Random Search: In this approach, the model is evaluated for random combinations of values of hyperparameters to find the best parameters. It resolves the issue of Grid Search since it does not go through all the combinations of the parameter settings it just searches for random combinations of parameters.

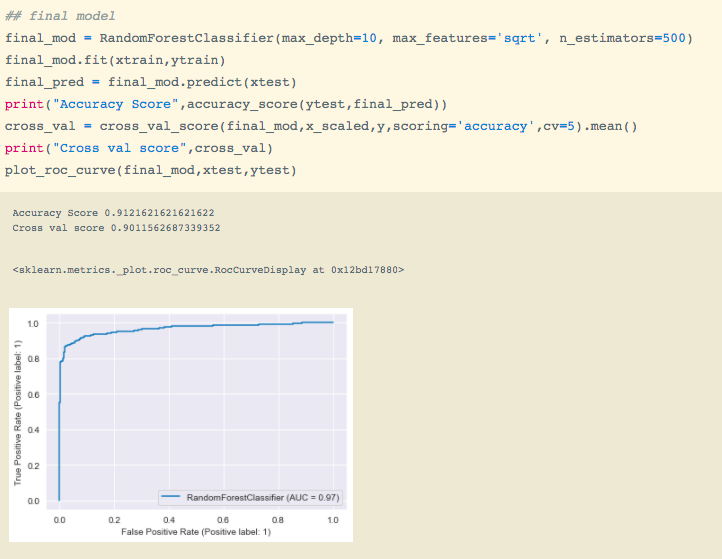
I will be using Random search for my model just because it is a little bit faster than grid search.



After fitting a random search I will check the best parameters and further I will check the best estimators and will pass the best estimators further to our final model for best accuracy and performance and then we will evaluate it for its performance.



Passing the best estimators to the model and evaluating it -



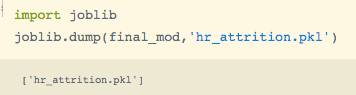
From the above report we can conclude:-

* The overall performance has increased of the model
* Accuracy and cross Val scores have increased and
* Auc score is 97% which implies our model is now 97% correctly identifying different classes.

Since we have designed our model and tuned it now I will save it for further deployment and prediction.

# Saving the model

This is the very last stage of our project. Here I will be saving the model by serializing it using the joblib package. Joblib is a utility package for saving and loading Python objects.



We can see above that I have saved the final model using joblib.

# 

# Conclusion

This marks the end of our project. We moved step by step through every process starting from EDA to model building and achieved a good model with excellent accuracy of 90%.

## 