IMPLEMENTING MACHINE LEARNING ON HR ANALYTICS PROJECT

**HR Analytics Project**

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the *effectiveness of their* employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

*Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead*, ***it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.***

* **Problem Definition**

*Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead*, ***it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.***

*A study by the*[***Center for American Progress***](https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf)*found that companies typically pay about one-fifth of an employee’s salary to replace that employee, and the cost can significantly increase if executives or highest-paid employees are to be replaced.*

In other words, the cost of replacing employees for most employers remains significant. This is due to the time spent to interview, find a replacement, bonuses and loss of productivity for several months while the new employee gets accustomed to new role.

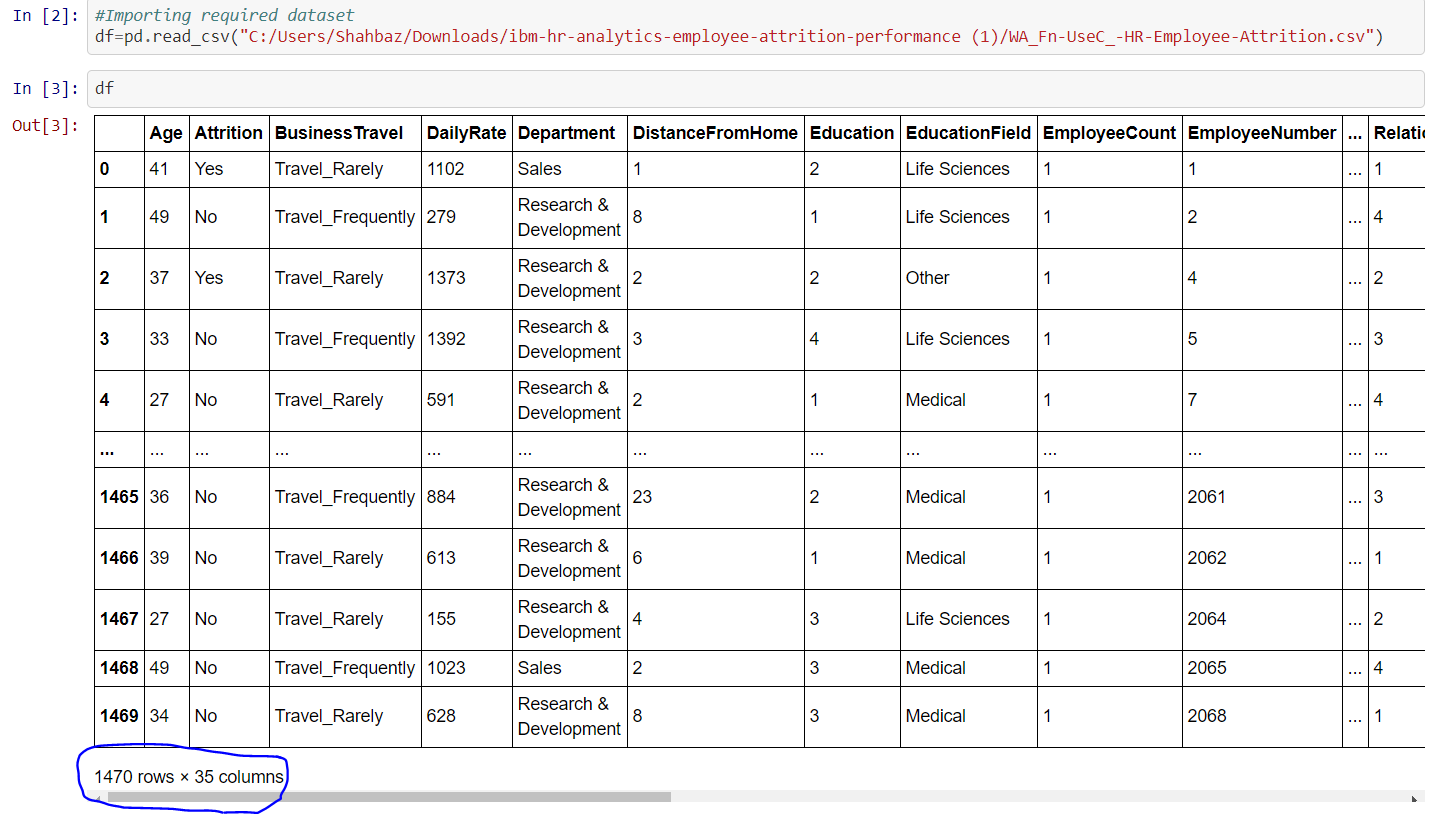
In this study, we will attempt to solve the following problem statement i.e.

* What is the likelihood of an active employee leaving the company?
* What are the key indicators of an employee leaving the company?
* What strategies can be adopted based on the results to improve employee retention?
* **Data Analysis**

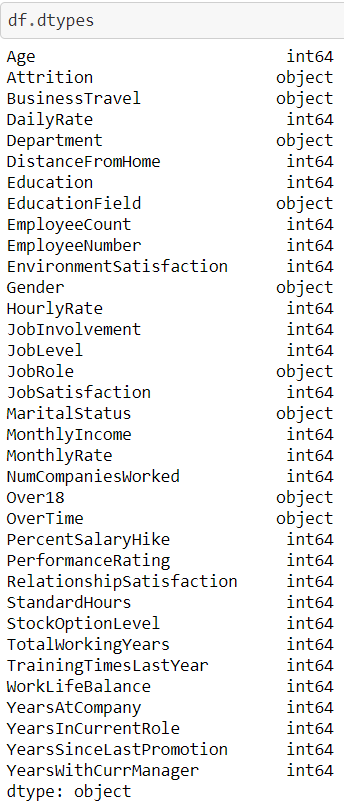
In this case study, a HR dataset was sourced from **IBM HR Analytics Employee Attrition & Performance** which contains data for 1470 employees with various information about the employees. I will use this dataset to predict when employees are going to quit by understanding the main drivers of employee churn.

**Data description, Importing Dataset & Exploratory Visualisations**

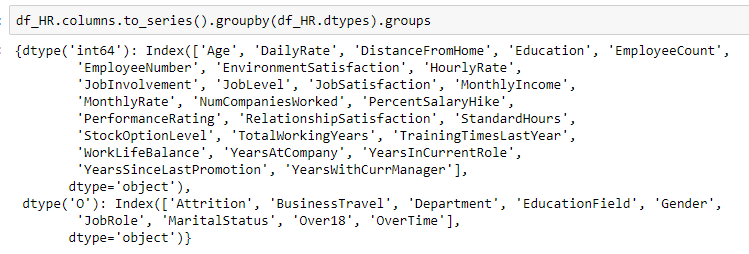
First of all, let’s import the dataset and make a copy of source file for this analysis. As seen , dataset has 1470 rows & 35 columns.



The dataset contains several numerical and categorical columns providing various information on employee’s personal and employment details.

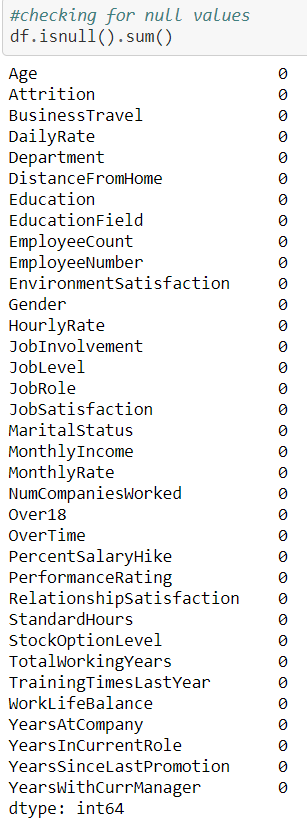


Breaking down the columns by their types i.e. (int64, object):



**Data Source**

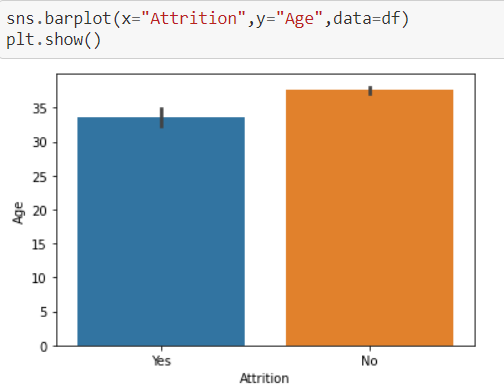
The data provided has no missing/null values.



**Feature Distribution by Target attribute:**

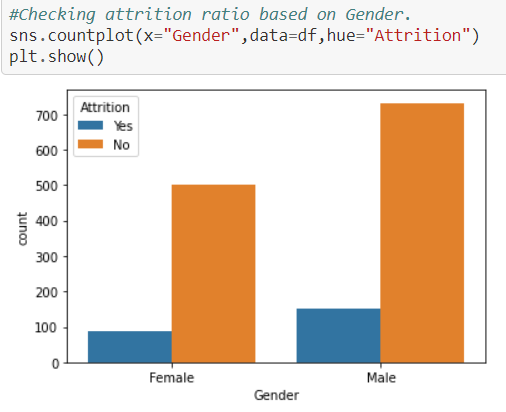
**Age:**

It is seen that attrition as highest in age group 32-34 years while above 35 years, no attrition observed. The age distribution for active and ex-employees only differs by approximately one year, with an average age of ex-employees at around 34 years and for current employees as 37 years approximately.



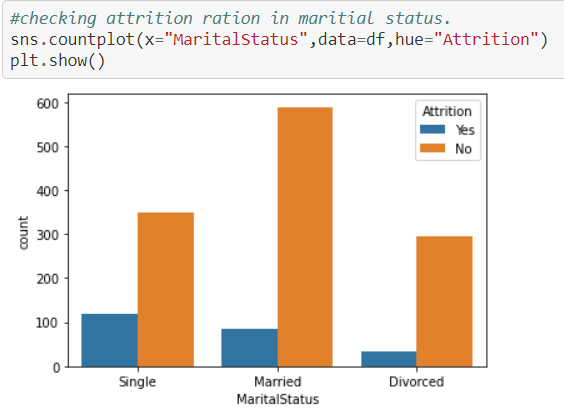
**Gender**

Gender distribution shows that the dataset features a higher relative proportion of male ex-employees than female ex-employees, with normalized gender distribution in dataset as around 125% for males and 90% for females.



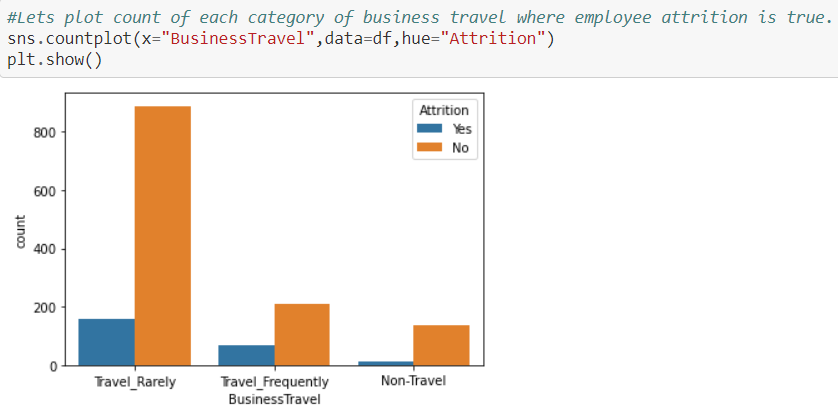
**Marital Status**

The dataset features three marital status: Married (600%), Single (350%), Divorced (300%). Married employees show the largest portion of leavers at 80%.



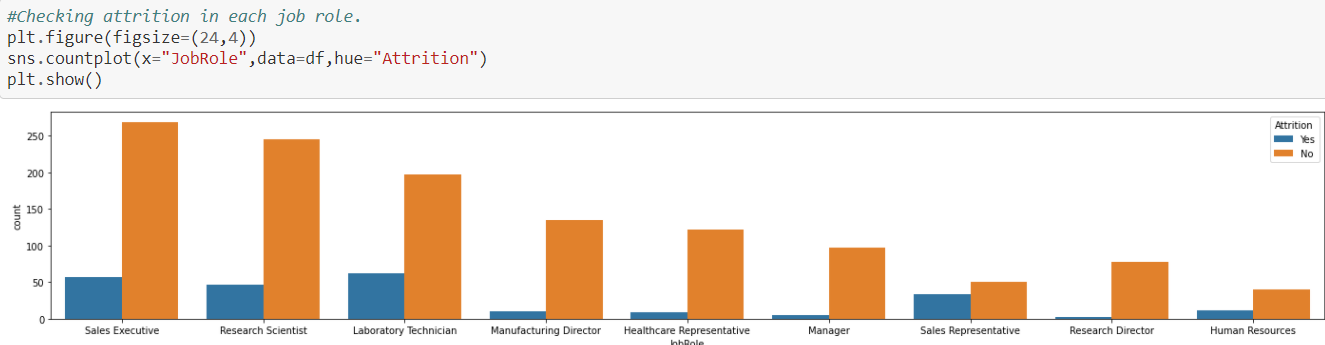
**Business Travel frequency and Attrition Status**

The relationship shows that there is a large normalized proportion of Leavers for employees that travel “Rarely”. Travel metrics associated with Business Travel status were not disclosed.



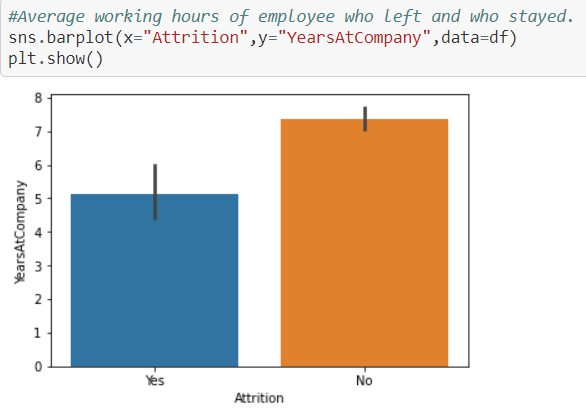
**Leavers considering Job Role**

Several Job Roles are listed in the dataset: Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources.

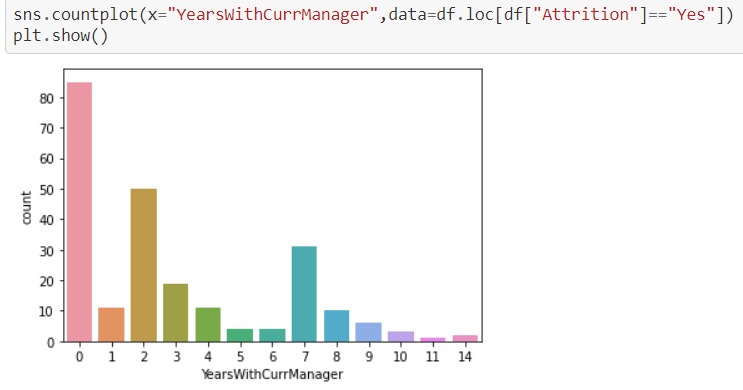


**Years at the Company**

The average number of years at company for currently active employees is 7.30 years approximately and ex-employees is around 5 years.

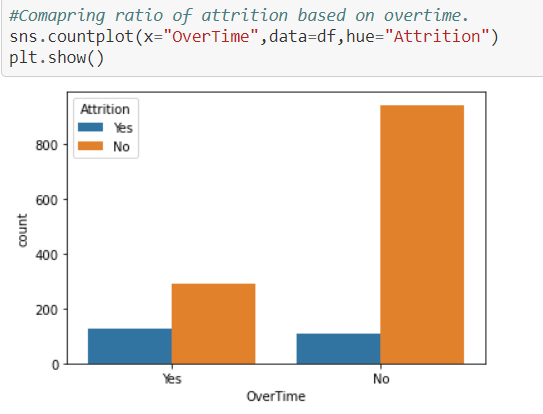


**Years with Current Manager**



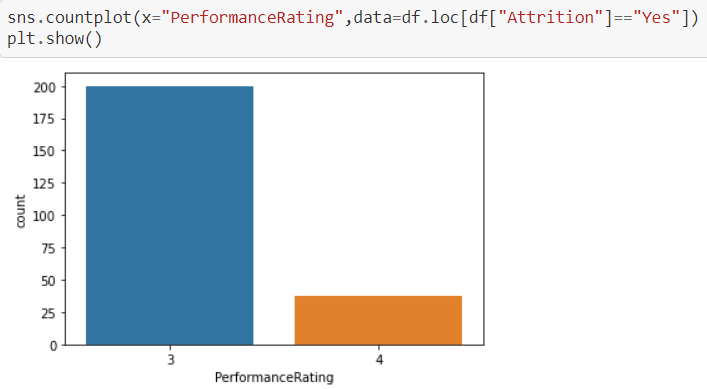
**Overtime**

Some employees have overtime commitments. The data clearly shows that there is significant larger portion of employees that left the company.



**Performance Rating**

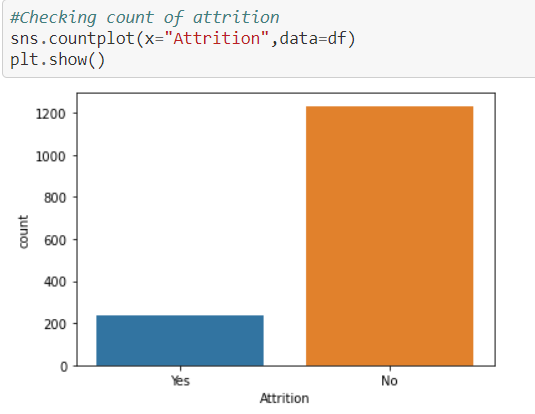
Here graph shows performance rating of people who left company. Maybe people didn't get quiet the rating they deserved and most of them falls under category 3 hence this could be the reason why people left company.



**Target Variable: Attrition**

The feature “Attrition” is what this machine learning problem is all about. We are trying to predict the value of the feature ‘Attrition’ by using other related features associated with the employees personal and professional history.

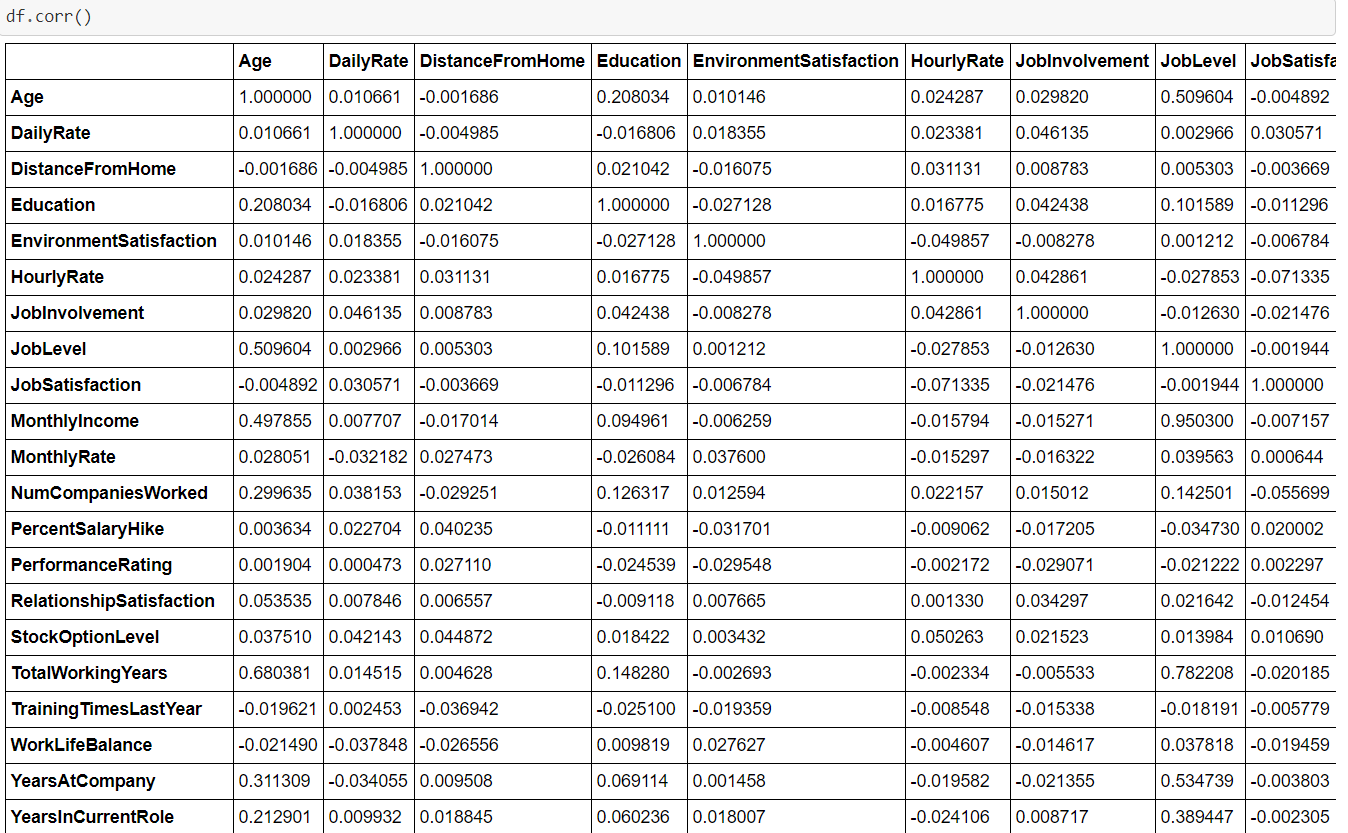
Machine Learning algorithms typically works best when the number of instances of each classes are roughly equal. We will have to address this target feature imbalance prior to implementing our Machine Learning algorithms.

. 

The bar shows it is highly imbalanced, out of 7 employees only 1 is facing attrition.

**Correlation**

Let’s take a look at some of most significant correlations. It is worth remembering that correlation coefficients only measure linear correlations.

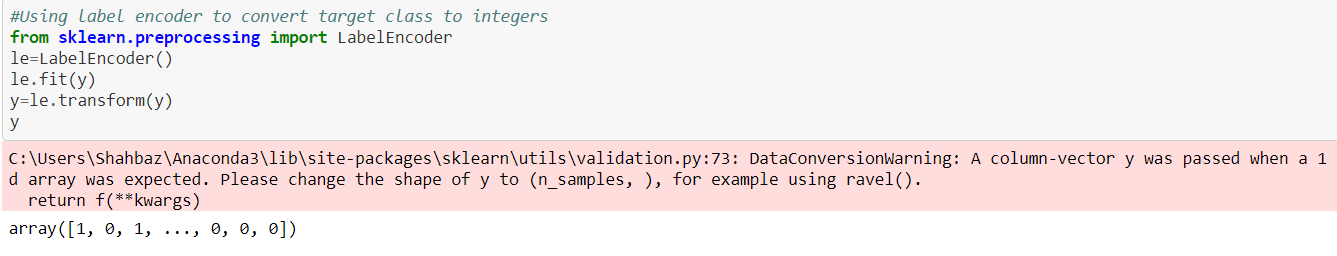


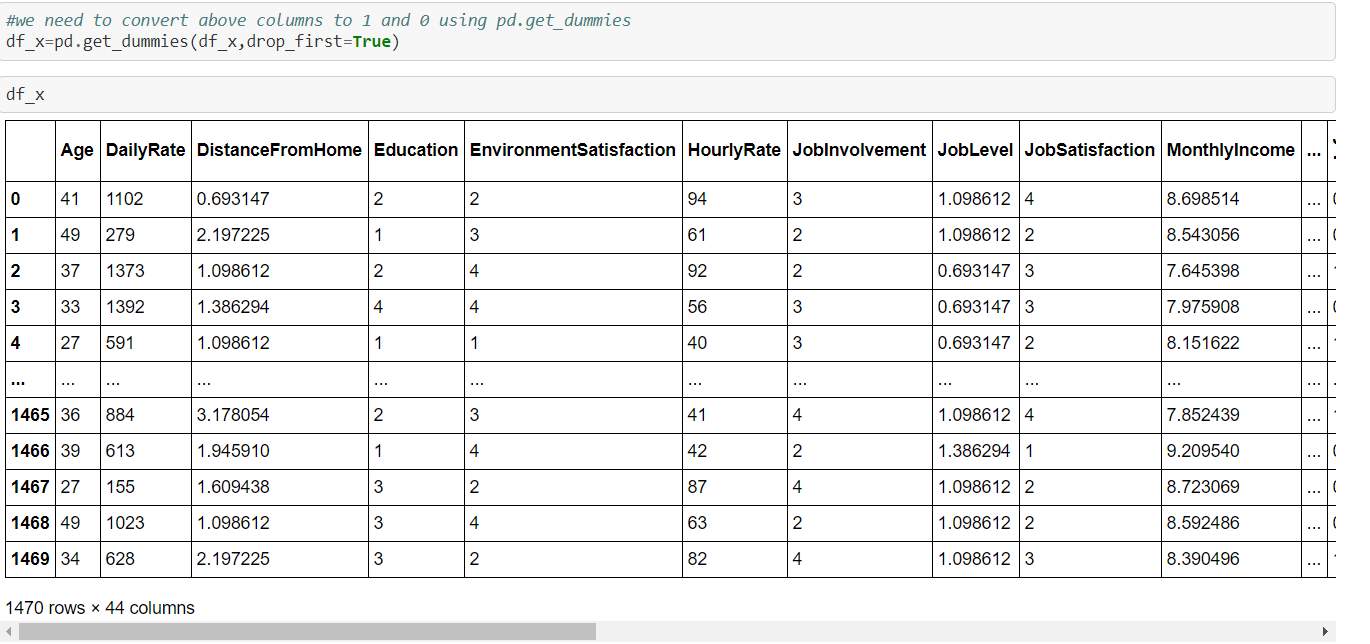
As shown, we see “Distance from Home”, “Job Satisfaction”, “Training times last years” are negatively correlated while “Age”, “Job Level”,” Total working years” are some positive aspects of correlation.

* **EDA Concluding Remarks**
* The dataset does not feature any missing data values and all features are of correct data type.
* The strongest positive correlations are: Age, Job Level, Relationship Satisfaction, Total Working Years.
* The strongest negative correlations are: Distance from Home, Job Satisfaction, Training times last years, Percent Salary Hike, Performance Rating & Work life Balance.
* The dataset is imbalanced with majority of observations describing Current Active Employees.
* Married Employees shows the largest proportion of leavers, compared to Single and Divorced counterparts.
* People who live away from their work shows higher proportion of leavers.
* People who travel rarely shows the highest portion of leavers.
* People who have to work “Overtime” clearly shows higher portion of leavers compared to their counterparts.
* Performance rating is one of the major factors of people leaving companies because people didn’t get quiet good rating even after working for long in any particular organization and hence the leavers proportion is highest compared to their counterparts.
* **Pre-Processing Pipeline.**

**Encoding**

Machine learning algorithms can typically only have numerical values as their predictor variables. Hence **Label Encoding** becomes necessary here as it encode categorical labels with numerical values.

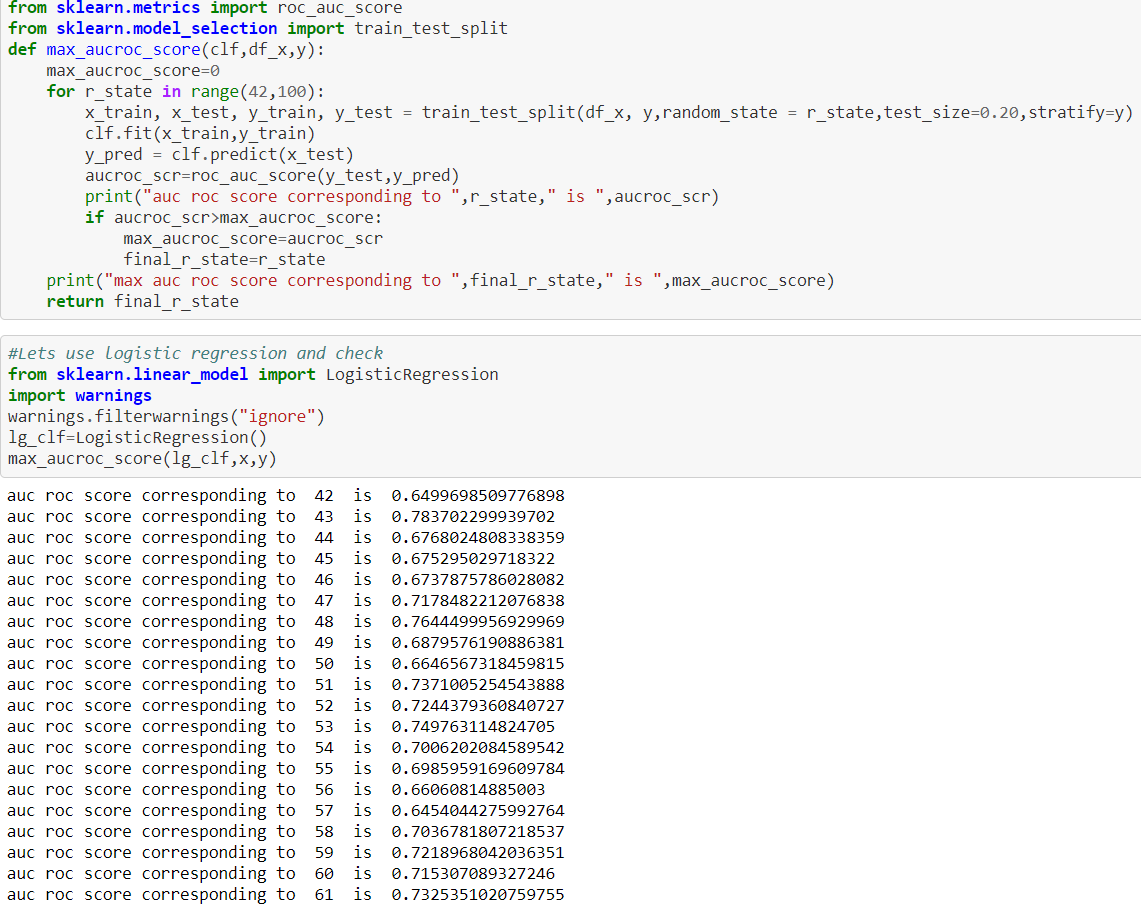




get\_dummies () here is used for data manipulation. It converts categorical data into dummy or indicator variables.

**AUC-ROC Curve**

Since this is imbalanced dataset hence, we will focus on AUC-ROC curve.

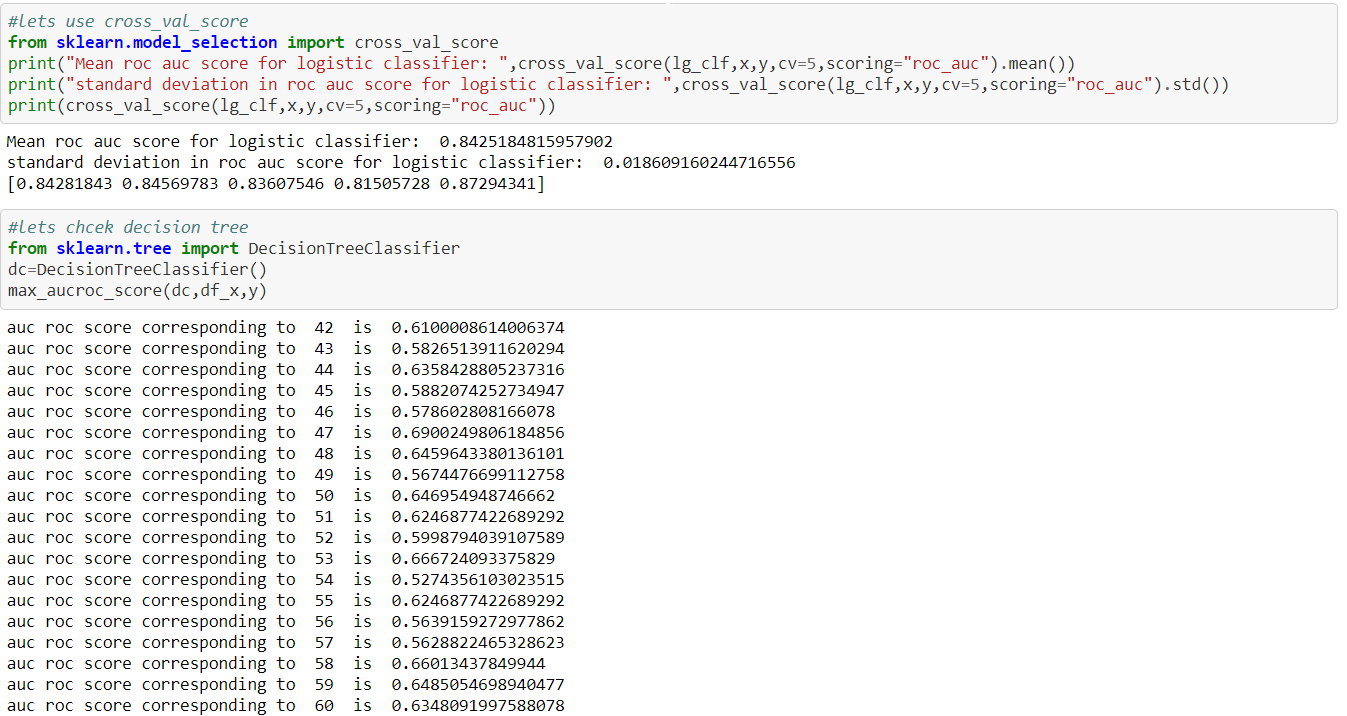


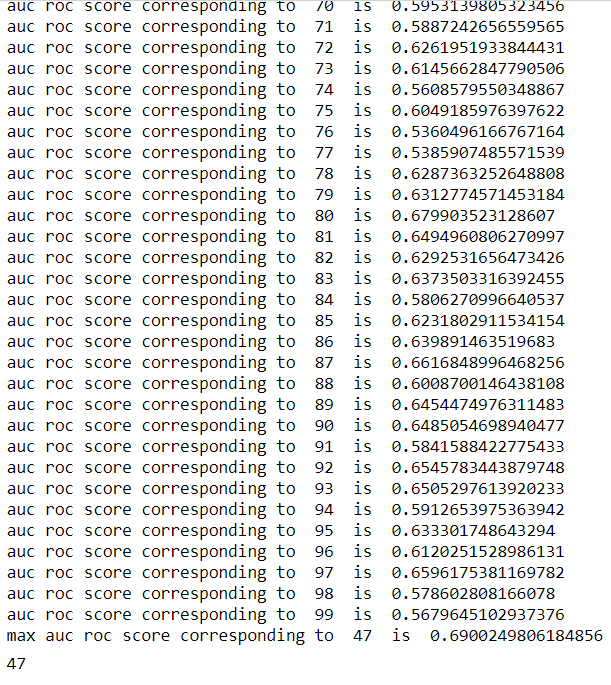
Here, AUC-ROC curve is the model selection metric for bi-multi class classification problem. ROC is a probability curve for different classes. ROC tells us how good the model is for distinguishing the given classes, in terms of predicted probability.

**Cross Validation / Decision Tree Classifier**

Cross validation or Decision Tree methods using here for checking/ testing the model on the dataset. Cross validation is a technique involves reserving a particular sample of dataset on which we can’t train the model. Later on, testing the model on the given sample before finalizing it, Train the model using the remaining part of the dataset.

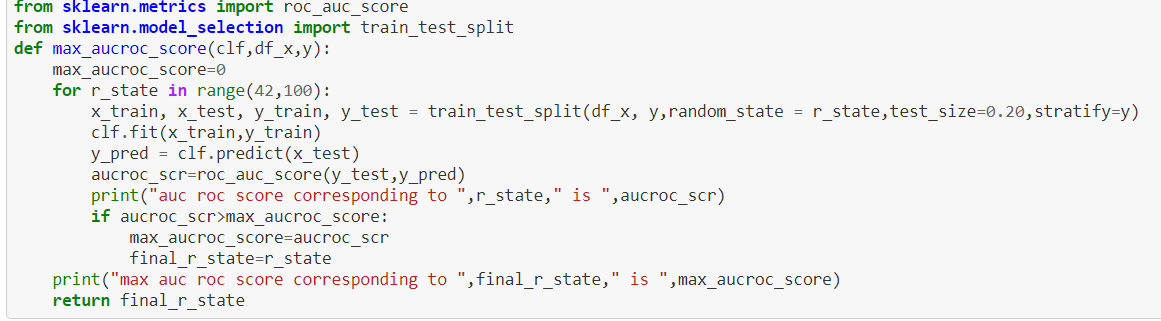
Decision Tree on other hand are non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.





**Splitting the data into Training and Testing sets**

Before implementing or applying Machine Learning algorithms, we must decouple training and testing dataframe from our master dataset.



* **Building Machine Learning Models**

**Baseline Algorithms**

The algorithms considered/ used in this section are: Logistic Regression, Random Forest, SVM, Decision Tree Classifier, Gaussian NB, KNN, Gradient Booster, Ada Booster.

Let’s evaluate each model and provide accuracy & standard deviation scores.

**Classification Accuracy** is number of correct predictions divided by all predictions or a ratio of correct predictions to total predictions. It is used when number of observations in each class is roughly equivalent.

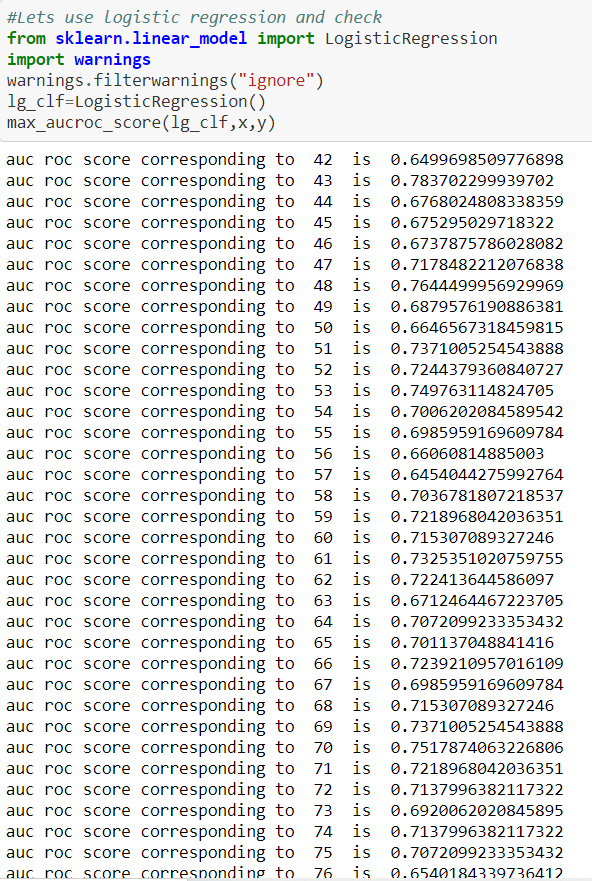
Area Under ROC Curve is a performance metric for binary classification problems. ROC is probability curve for different classes. ROC tells us how good the model is for distinguishing the given classes, in terms of predicted probability.

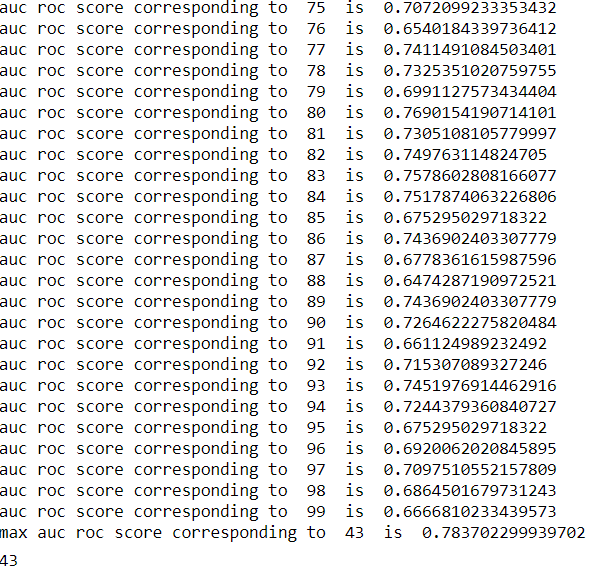
The AUC represents model’s ability to discriminate between positive and negative classes.

Based on ROC AUC comparison analysis, **Logistic Regression & SVC** shows the highest mean AUC scores. We will further look into these.

**Logistic Regression**

GridSearchCV allows use to fine tune hyper-parameters by searching over specified parameters values for an estimator.



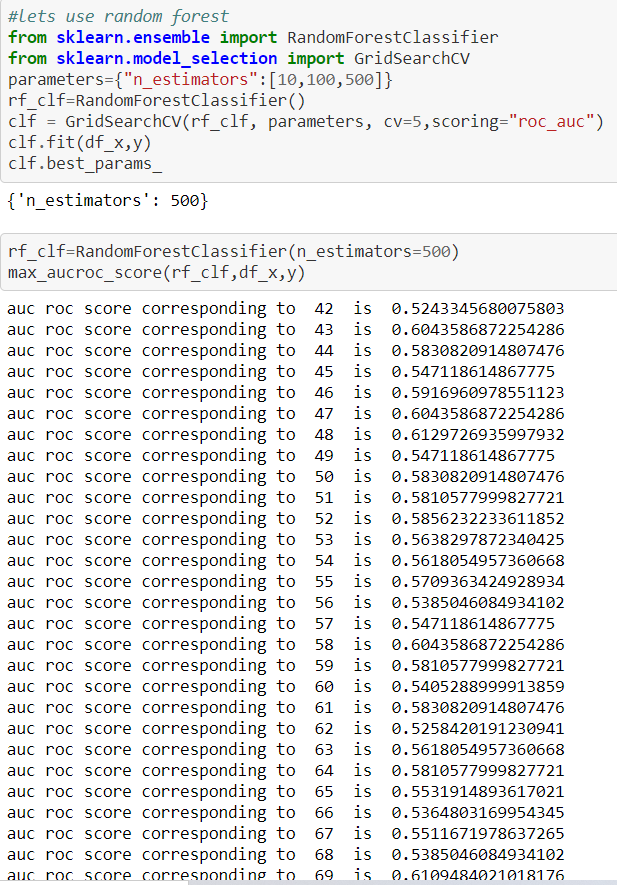


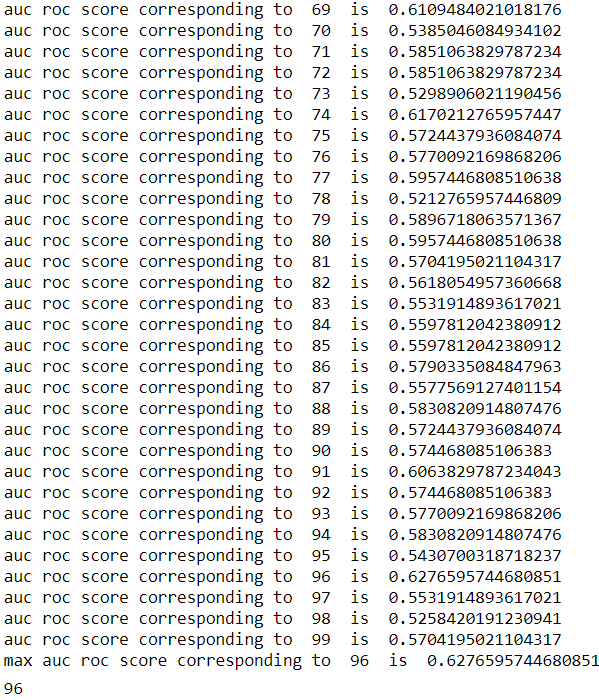
**Label Probability**

Instead of getting binary estimated target features (0 or 1), a probability can be associated with the predicted target. The output provides a first index referring to the probability that the data belongs to class 0 (employee not leaving), and the second refers to the probability that the data belongs to class 1 (employee leaving). Predicting probabilities of a particluar label provides us with the measure of how likely an employee is to leave the company.

**Random Forest Classifier**

Taking a look at random Forest Algorithm. Will fine-tune the Random Forest algorithm’s hyper parameters by cross validation against AUC score.





Random Forest allows us to know which features are of most importance in predicting the target feature (“Attrition” in this project).

* **Concluding Remarks**

**Risk Score**

Risk score is general practice refers to creating an easily calculated number that reflects the level of risk in the presence of some risk factors.

Employees can be assigning a “Risk Score” based on the predicted label such as:

* Low Risk for employees with label < 0.6.
* Medium Risk for employees with label between 0.6 & 0.8.
* High Risk for employees with label > 0.8.

**Indicators of People Leaving an Organization.**

The strongest and most common indicators for people leaving an organization are:

* **Monthly Income**: People whose monthly income is higher are less likely to leave the company due to financial stability.
* **Over Time**: People who worked overtime in response to business continuity also , at some point people do overtime in response to they get appreciated for their work. Failing to get so, made people leave the company.
* **Age**: Employees aging from 25-35 are more likely to leave the company due to more opportunities at other firms with good packages( Until and unless they been provided with good packages in their current company).
* **Distance from Home**: People who live far away from company are more likely to leave the company due to over travelling and also time consuming. Companies should provide cabs or transportation allowance in an aspect to keep an employee.
* **Total Working Years**: The more years an employee serves a company the less likely is his/her chance to leave the organisation (conditions apply).
* **Years at Company**: Employees who serve their 2-3 years to an organisation should be identified as potentially higher risk of leaving (due to experience and higher demand of payroll).
* **Years with Current Manager**: A large number of people are either likely to leave the organisation in a span of 6 months to 1 year after appointing with new manager. It totally depends upon the several metrics to be looked upon: patterns in the employees who have resigned: this may indicate recurring patterns in employees leaving in which case action may be taken accordingly.

**Final Outcome:**

A strategic retention plan can be drawn for each **Risk Score**group. In addition to the suggested steps for each feature listed above, face-to-face meetings between a HR representative and employees can be initiated for **medium-** and **high-risk employees** to discuss work conditions. Also, a meeting with those employee’s Line Manager would allow to discuss the work environment within the team and whether steps can be taken to improve it.