### [**IBM HR Analytics Employee Attrition & Performance**](http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html)

Problem Definition

The key to success in an organisation is the ability to attract and retain top talents. It is vital for the Human Resource (HR) Department to identify the factors that keep employees and those which prompt them to leave. Organisations could do more to prevent the loss of good people.

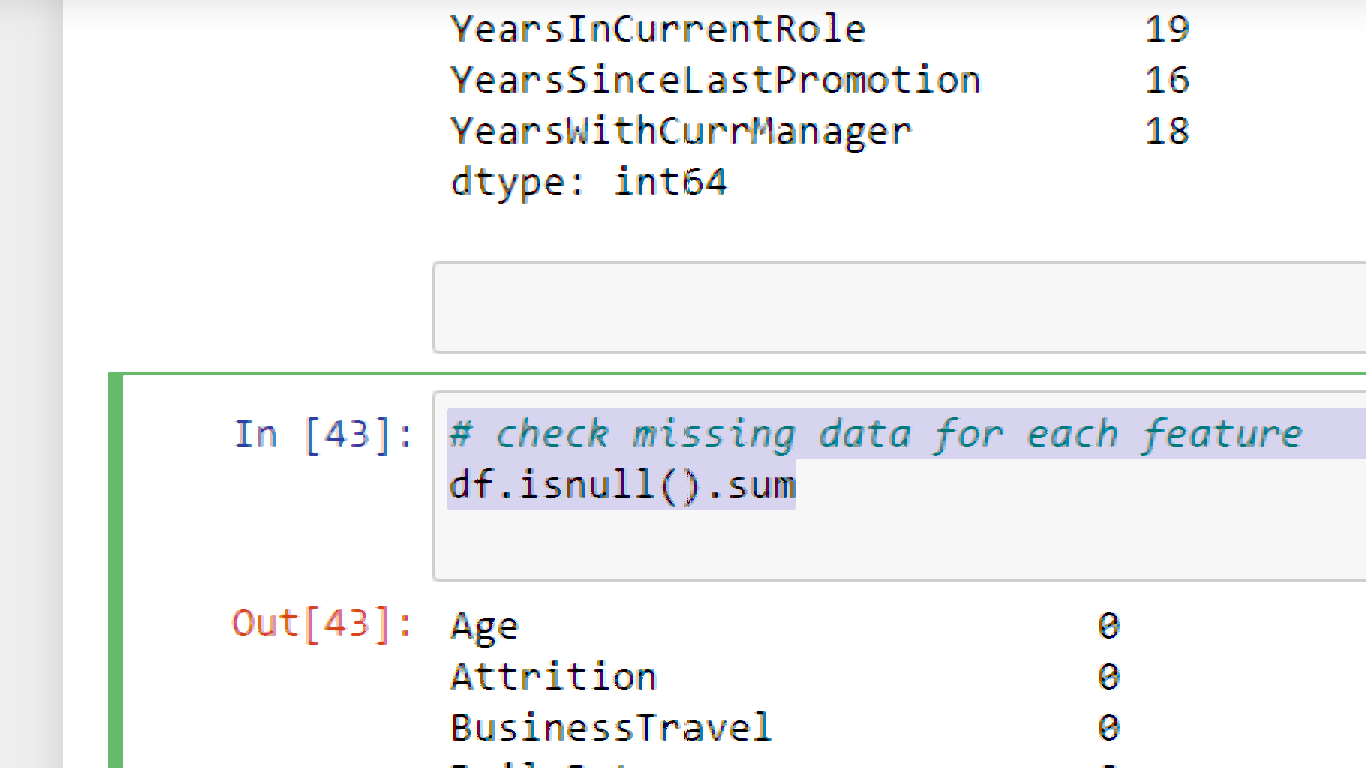
This project is based on a hypothetical dataset downloaded from [IBM HR Analytics Employee Attrition & Performance](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset). It has 1,470 data points (rows) and 35 features (columns) describing each employee’s background and characteristics; and labelled (supervised learning) with whether they are still in the company or whether they have gone to work somewhere else. Machine Learning models can help to understand and determine how these factors relate to workforce attrition.

**Data Analysis.**

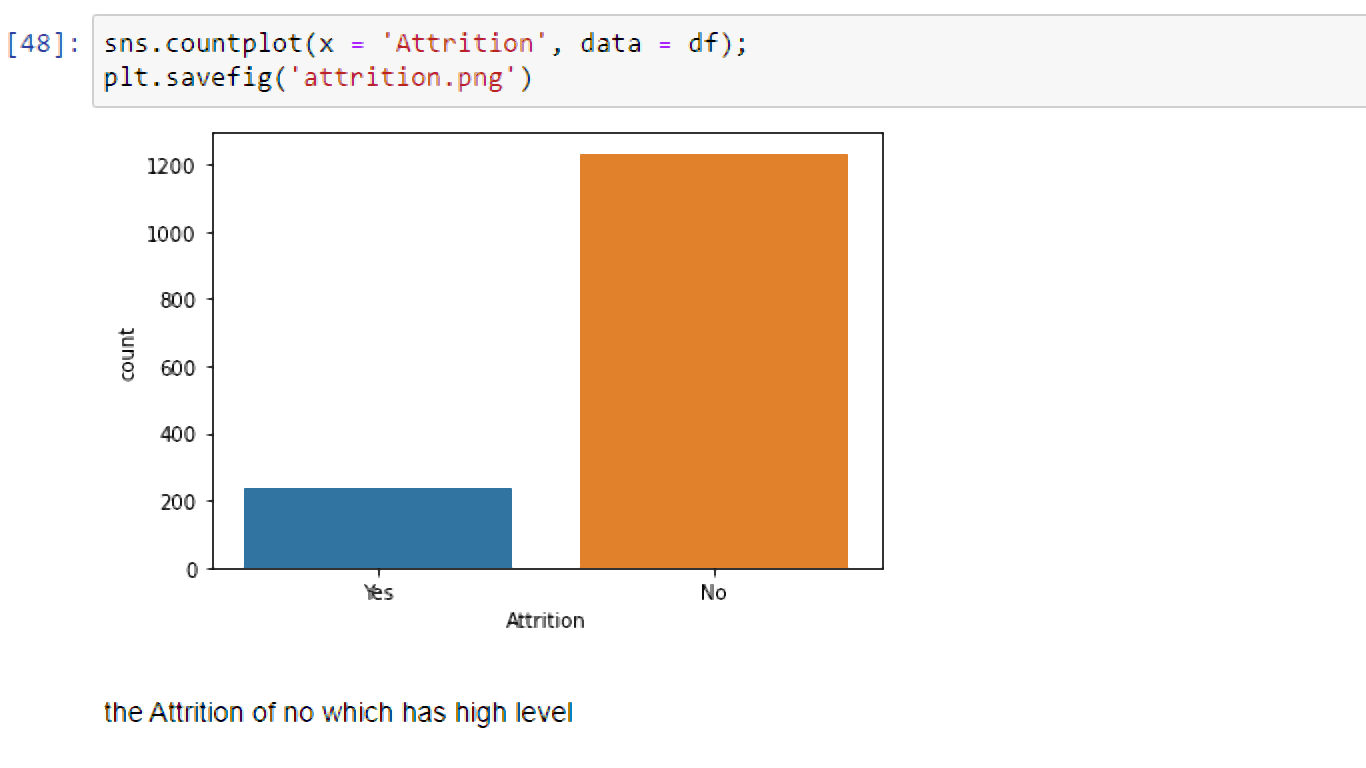
## Exploratory Data Analysis (EDA)

# check missing data for each feature

df.isnull().sum



In this dataset we have checked the null values The dataset is well organised with no missing values. Target class is imbalance, with attrition rate of 16%.



Are employees leaving because they are poorly paid? Employees are paid an hourly rate of $30 to $100, and attrition seems to happen at every level regardless of employee hourly rate. This can be confirmed later at feature importance.

df.columns.to\_series().groupby(df.dtypes).groups

We have divided the dataset into the numercial value and categorical value so the we can convert the categorical value into the numerical value so that we can predict the modelling the value.

Column 'Employee Count' is all 1s which indicate every observation is linked with 1 employee only

Column 'Standard Hours' is all 80s which means everyone in this dataset works as a fulltime employee and we could definitely drop it as well.Column 'Over18' is another interesting column which tells us every employee in this dataset is over 18 and we will drop it as well.

df = df.drop(columns = ['EmployeeNumber', 'EmployeeCount', 'StandardHours', 'Over18'])

df.head()

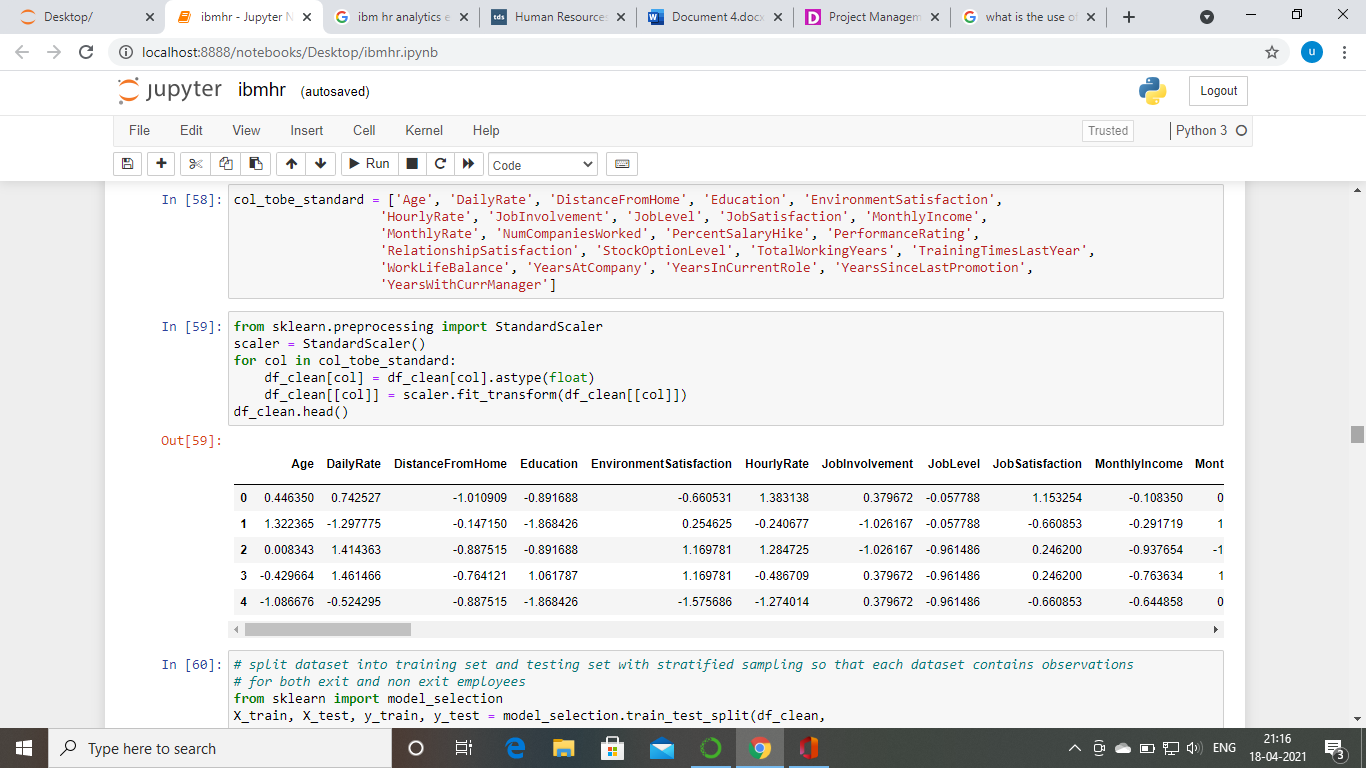
By using the dataset, we are going the drop the 'EmployeeNumber', 'EmployeeCount', 'StandardHours', we don’t want these columns for our data set so we can use it for best modelling

We use get dummies for the three columns 'Business Travel', 'Gender', 'Marital Status'

**StandardScaler**

standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. Standard scalar does not meet the strict definition of scale I introduced earlier.

Minmaxscalar(feature range = (0, 1)) will transform each value in the column proportionally within the range [0,1]. use this as the first scaler choice to transform a feature, as it will preserve the shape of the dataset (no distortion).

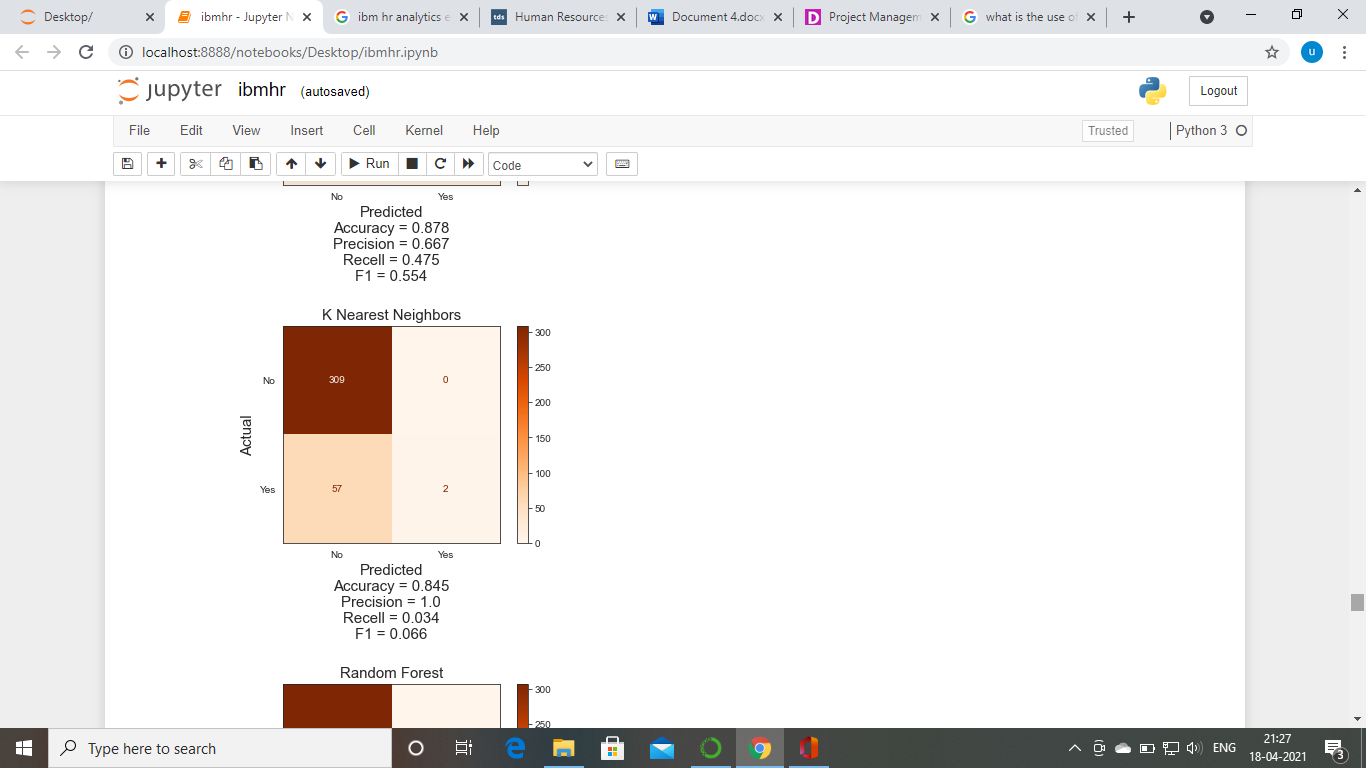


**Building Machine Learning Models**

Now we will train several Machine Learning models and compare their results. Note that because the dataset does not provide labels for their testing-set, we need to use the predictions on the training set to compare the algorithms with each other. Later on, we will use cross validation.

kNeighbourClassifier

By default, the KNeighborsClassifier looks for the 5 nearest neighbors. We must explicitly tell the classifier to use Euclidean distance for determining the proximity between neighboring points

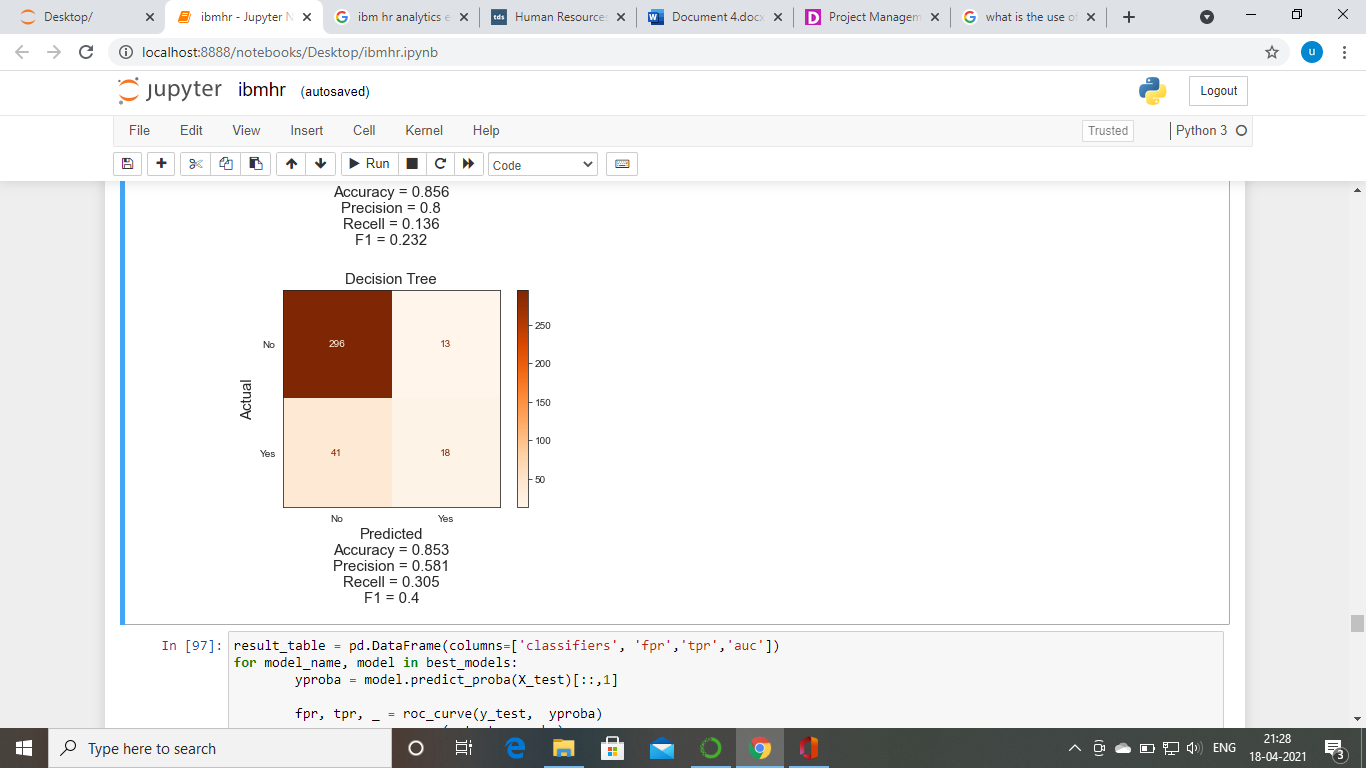


Gaussian Naive Bayes

A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e, normal distribution

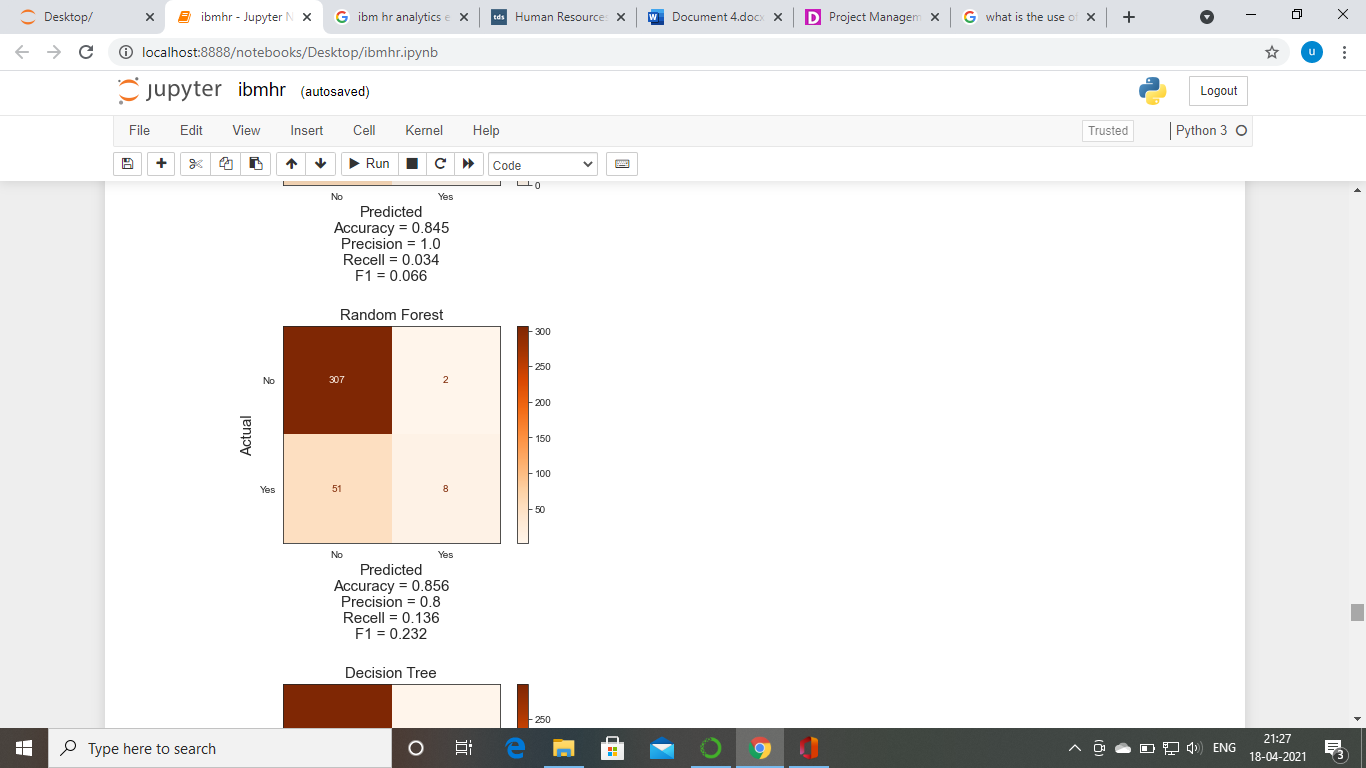
A Decision Tree

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.



Random forest

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks).

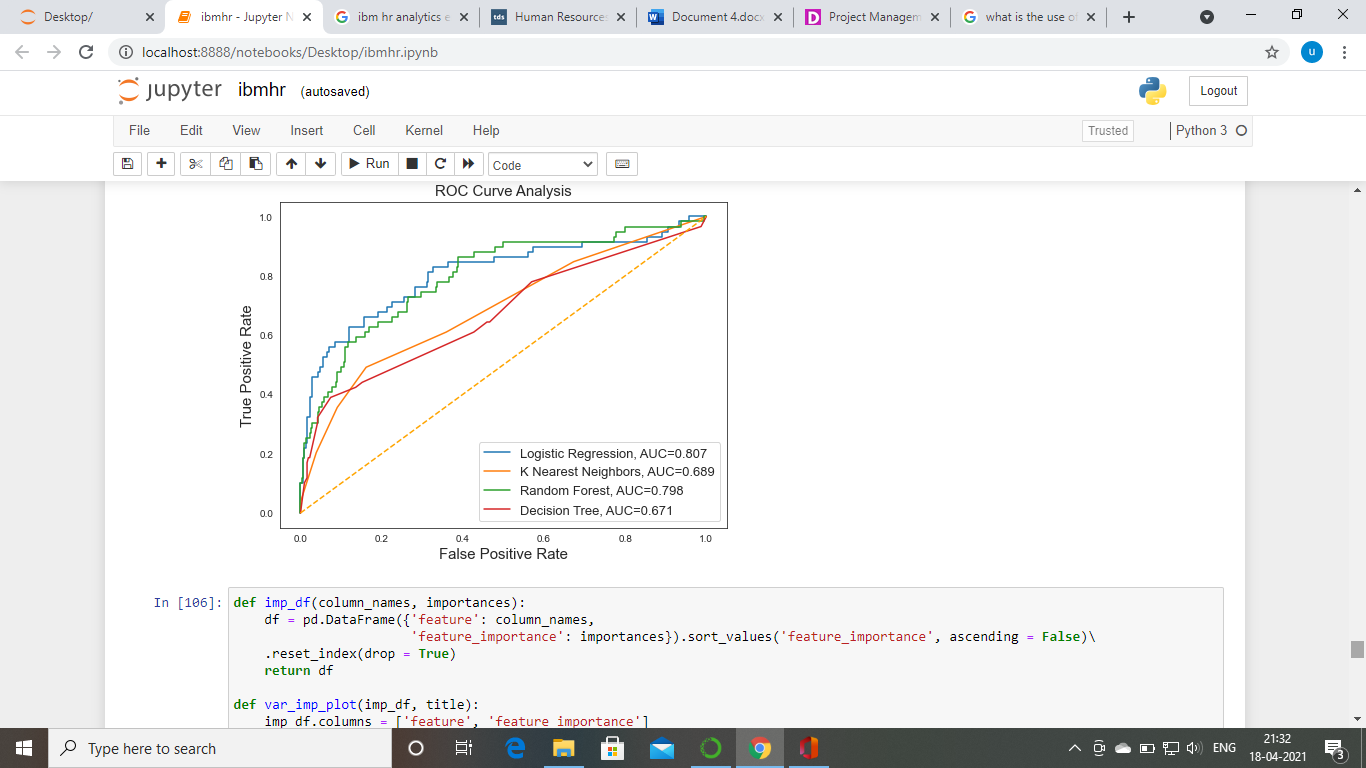


AdaBoost

AdaBoost is best used to boost the performance of decision trees on binary classification problems. Adaboost was originally called Adaboost. M1 by the authors of the technique Freund and Schapire. More recently it may be referred to as discrete Adaboost because it is used for classification rather than regression.

cross\_val\_score

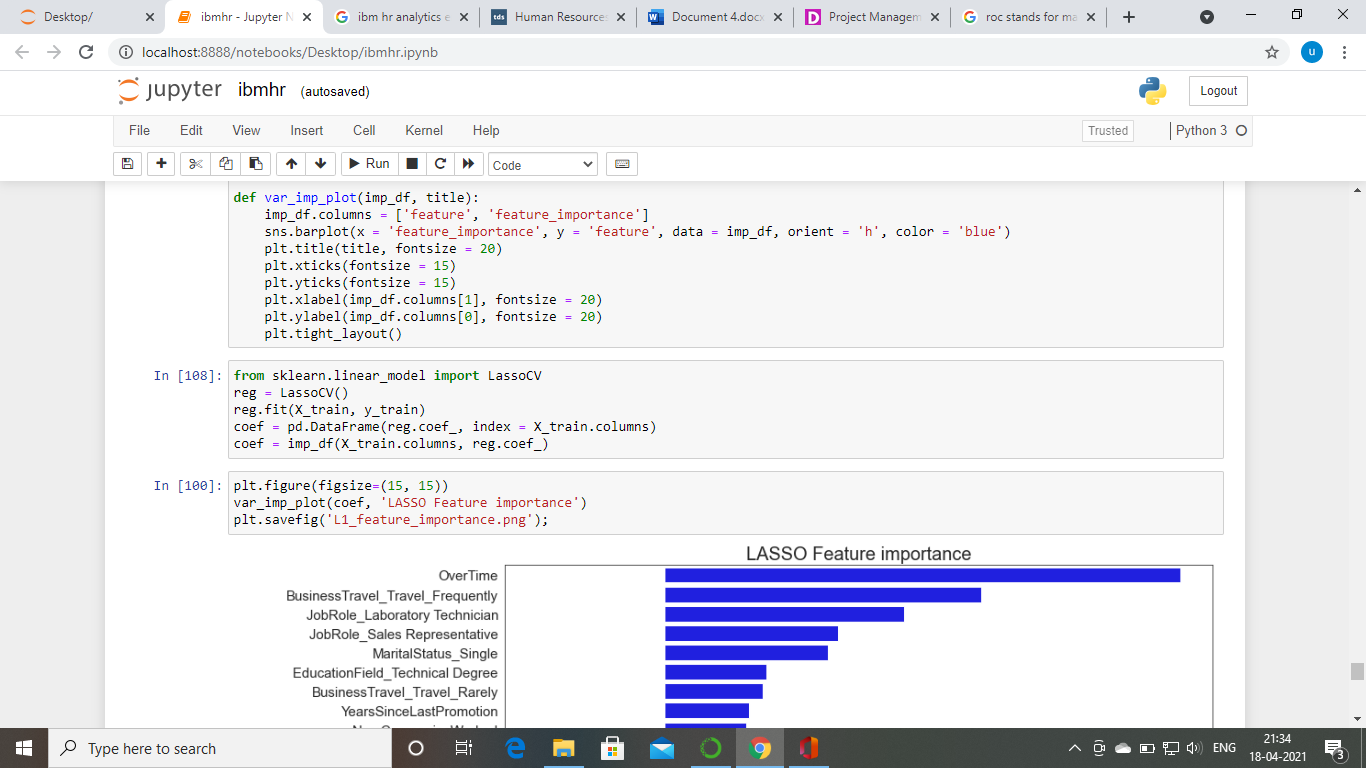
cross\_val\_score returns score of test fold where cross\_val\_predict returns predicted y values for the test fold. For the cross\_val\_score() , you are using the average of the output, which will be affected by the number of folds because then it may have some folds which may have high error (not fit correctly).

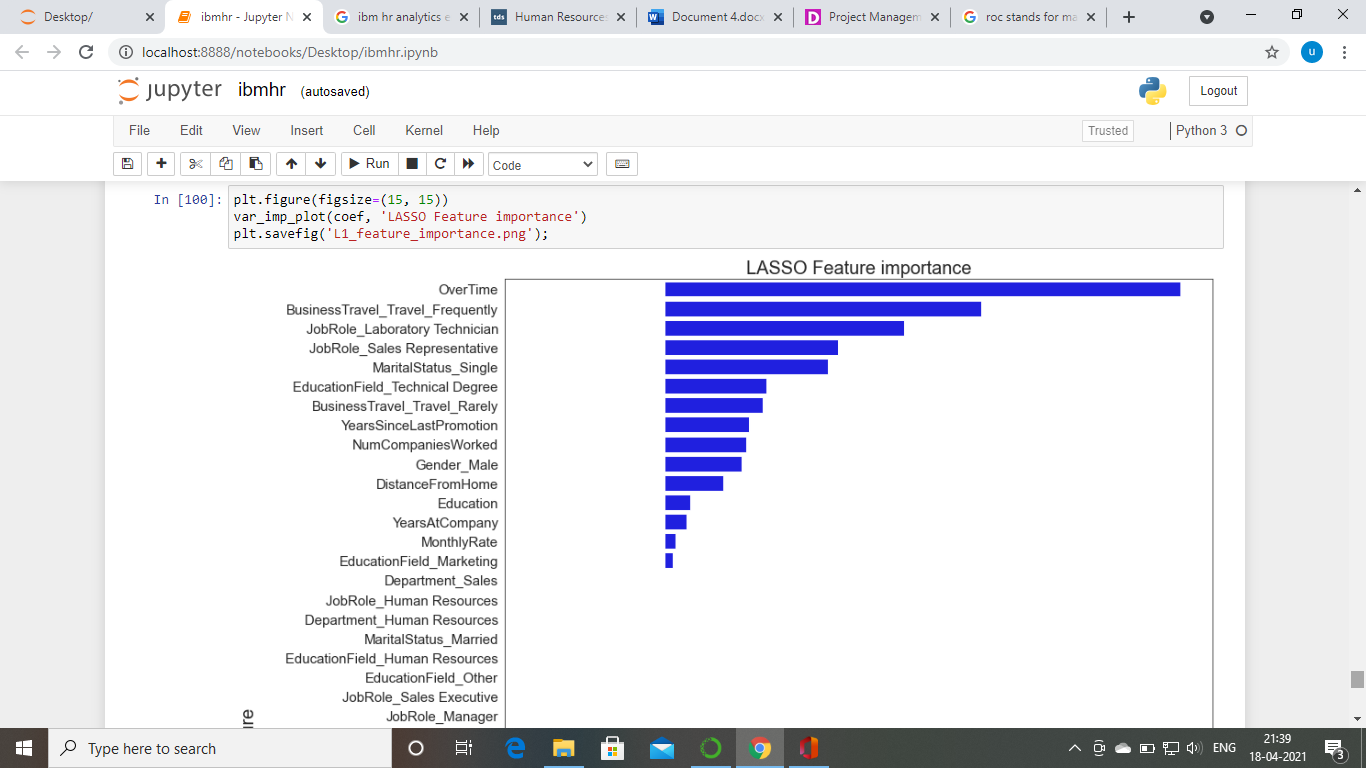


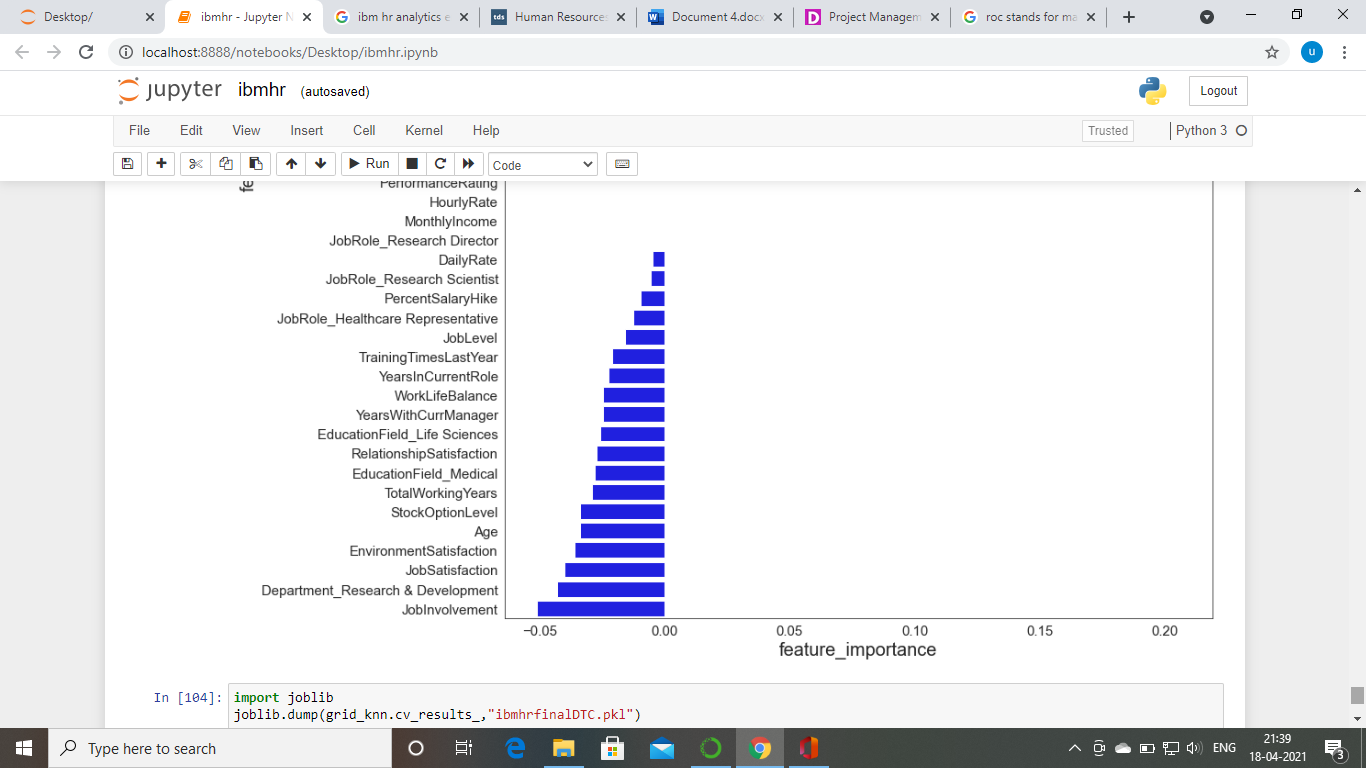
A receiver operating characteristic curve, or **ROC** curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. ... **ROC** analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

The **ROC curve** shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give **curves** closer to the top-left corner indicate a better performance. ... The closer the **curve** comes to the 45-degree diagonal of the **ROC** space, the less accurate the test.

**AUC** - **ROC** curve is a performance measurement for the classification problems at various threshold settings. **ROC** is a probability curve and **AUC** represents the degree or measure of separability. ... By analogy, the Higher the **AUC**, the better the model is at distinguishing **between** patients with the disease and no disease.







# Conclusion

HR Analytics is gaining traction in organisations that embrace digital transformation. The scope has expanded from analytics of employee work performance to providing insights so that decisive improvements can be made to organisational processes. While some level of attrition is inevitable, it should be kept at the minimal possible level