# **About the Problem:**

We are given the data of direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (target variable y).

# **Type of Machine Learning problem:**

This is a binary classification problem. Our two classes are “yes” denoting that the customer subscribed to a term deposit, and “no” denoting that the customer did not subscribe.

# **Exploratory Data Analysis:**

When working on a new dataset in order to take intelligent action, you need to understand your data. Exploratory data analysis (EDA) allows us to develop the gist of what our data may look like and what kinds of questions can be answered by them. EDA is important because it allows the explorer to make critical decisions about what is interesting to pursue and what probably isn’t worth following up on and thus building a hypothesis using the relationships between variables. EDA helps us to gain insights and help us understand the correlation between the independent variables and target variables.

**Data Set**-

Here I will be using the data set of Bank Marketing Campaign to Predict if the client will subscribe to direct marketing campaign.

As mentioned above, the dataset consists of direct marketing campaigns data of a banking institution. The dataset was picked from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) which is an amazing source for publicly available datasets. There were four variants of the datasets out of which we chose “ bank-additional-full.csv” which consists of 41188 data points with 20 independent variables out of which 10 are numeric features and 10 are categorical features. The list of features available to us are given below:

Problem-

The data is related to direct marketing campaigns of a Portuguese banking institution. Predict if the client will subscribe to a term deposit based on a marketing campaign

The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. There are four datasets:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]

2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.

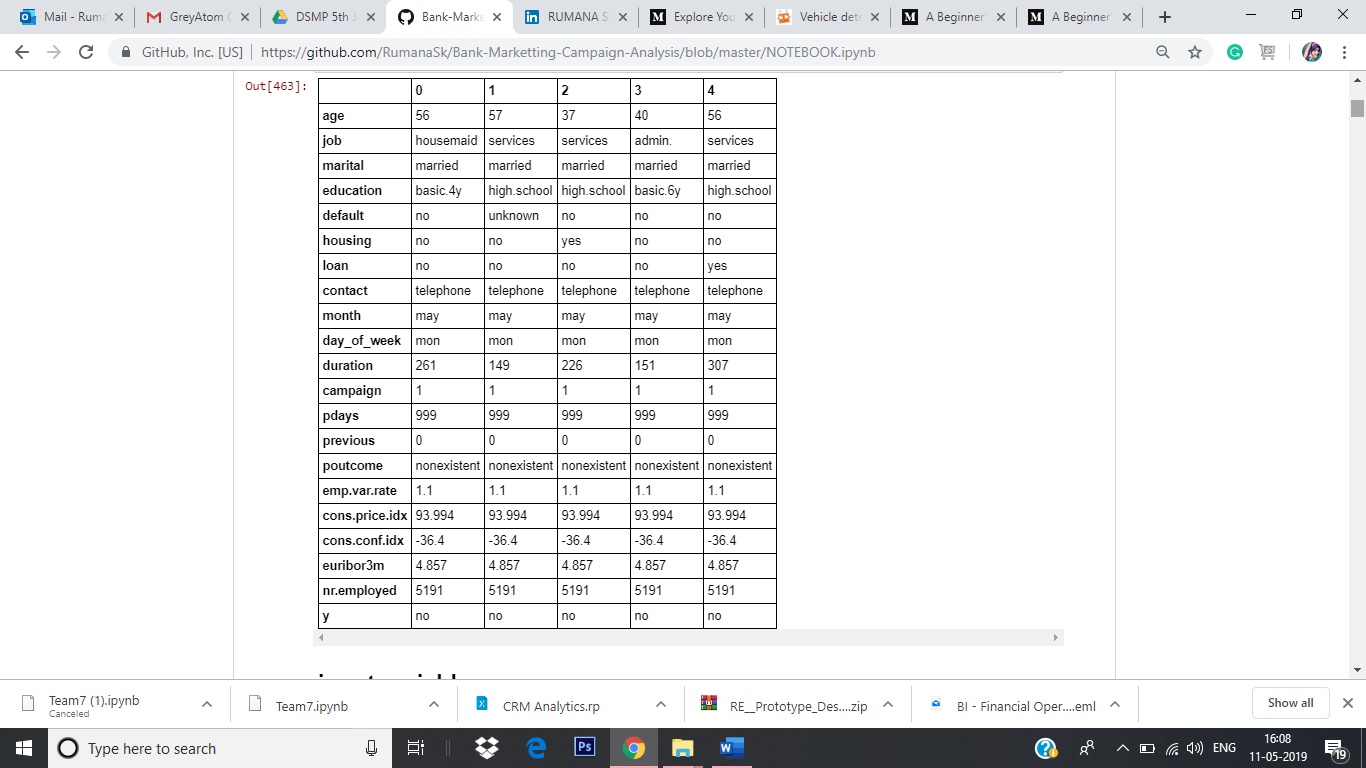
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with fewer inputs).

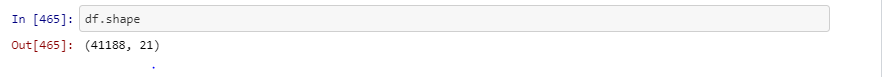
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with fewer inputs). The smallest datasets are provided to test more computationally demanding machine learning algorithms

**Goal:** The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

**Getting to Know about Data**

At the very first of Exploratory Data Analysis, we want to start understanding the data quickly, so here we use df.head(),df.describe() and df.shape().



In this data set, we have 21 columns and almost 41,188 rows.

# **Input variables:**

1. ***age*:** (numeric)  
2. ***job*** **:** type of job (categorical: 'admin.','blue collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3***. marital* :** marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4. ***education*** (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5. ***default:*** has credit in default? (categorical: 'no','yes','unknown')  
6. ***housing:*** has housing loan? (categorical: 'no','yes','unknown')  
7. ***loan:*** has personal loan? (categorical: 'no','yes','unknown')

## related with the last contact of the current campaign:

8. ***contact:*** contact communication type (categorical: 'cellular','telephone')  
9. ***month:*** last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10. ***day\_of\_week:*** last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11. ***duration:*** last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

12. ***campaign:*** number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13. ***pdays:*** number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14. ***previous:*** number of contacts performed before this campaign and for this client (numeric)  
15. ***poutcome:*** outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

## social and economic context attributes

16. ***emp.var.rate:*** employment variation rate - quarterly indicator (numeric)  
17. ***cons.price.idx:*** consumer price index - monthly indicator (numeric)   
18. ***cons.conf.idx:*** consumer confidence index - monthly indicator (numeric)  
19. ***euribor3m:*** euribor 3 month rate - daily indicator (numeric)  
20. ***nr.employed:*** number of employees - quarterly indicator (numeric)

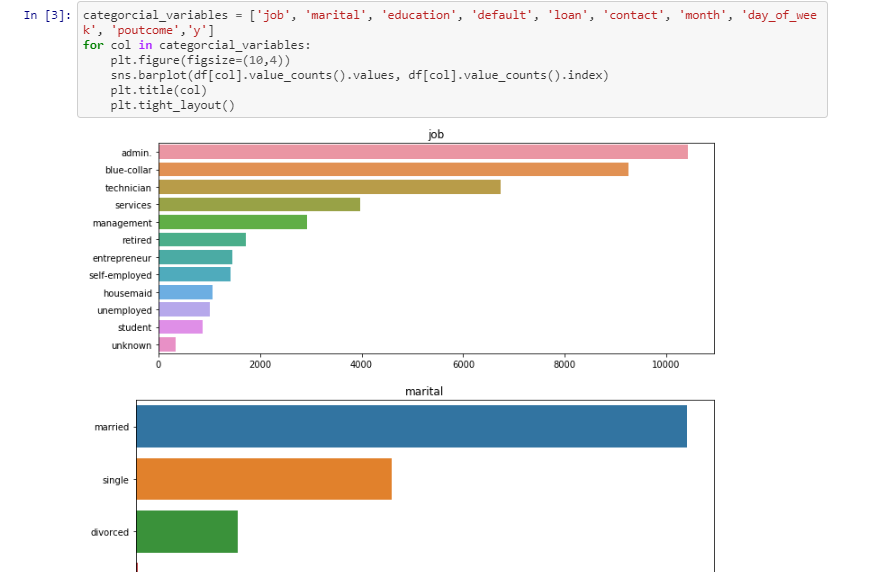
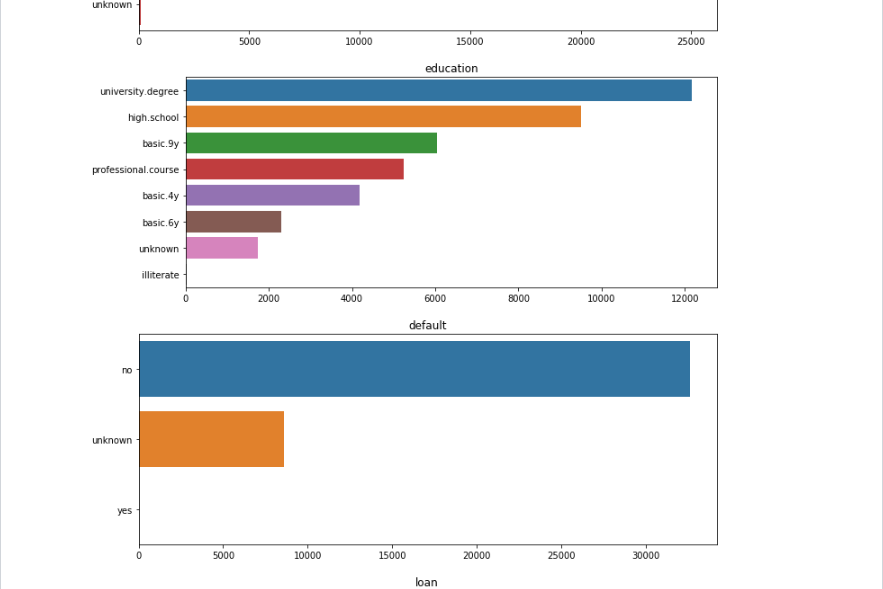
## Output variable (desired target):

21. **y** - has the client subscribed a term deposit? (binary: 'yes','no')

## 

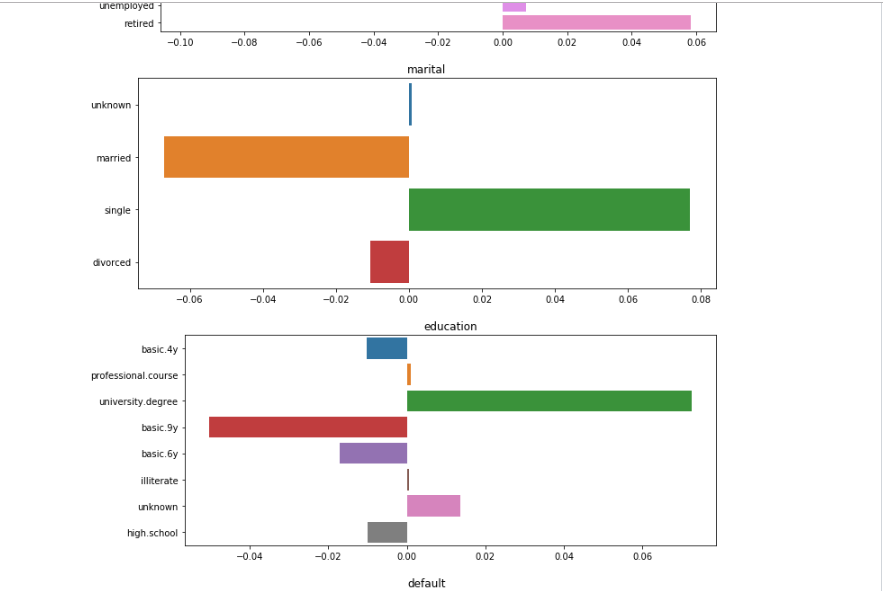
## Categorical Variables:

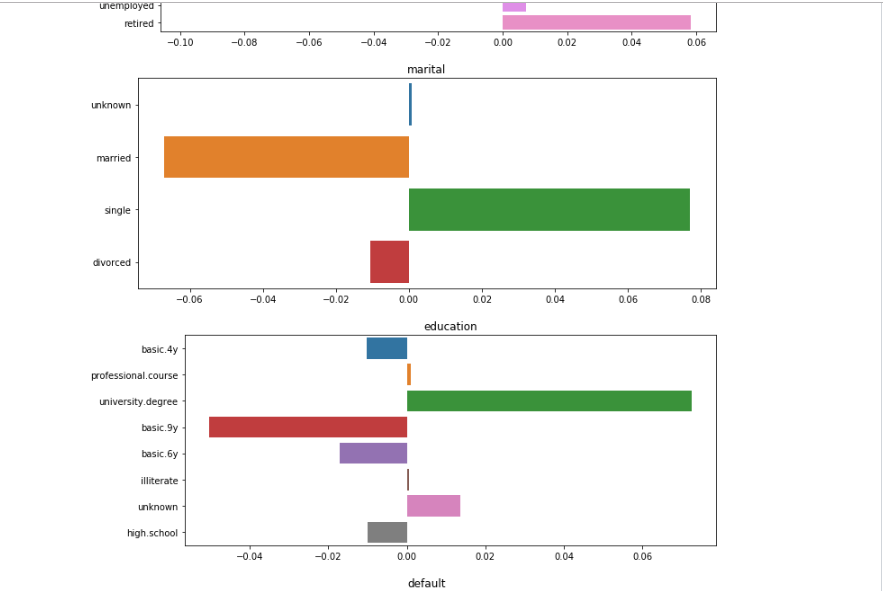
We first start the exploratory analysis of the categorical variables and see what are the categories and are there any missing values for these categories. Here, we used the seaborn package to create the bar graphs.

## List of normalized relative frequency of the target class per category.

Normalized distribution of each class per feature and plotted difference between positive and negative frequencies. Positive values imply this category favours clients that will subscribe and negative values categories that favour not buying the product.



**Handling Missing Values**

There are unknown values for many variables in the Data set. One of the ways is to discard the row but that would lead to reduction of data set and hence would not serve our purpose of building an accurate and realistic prediction model.

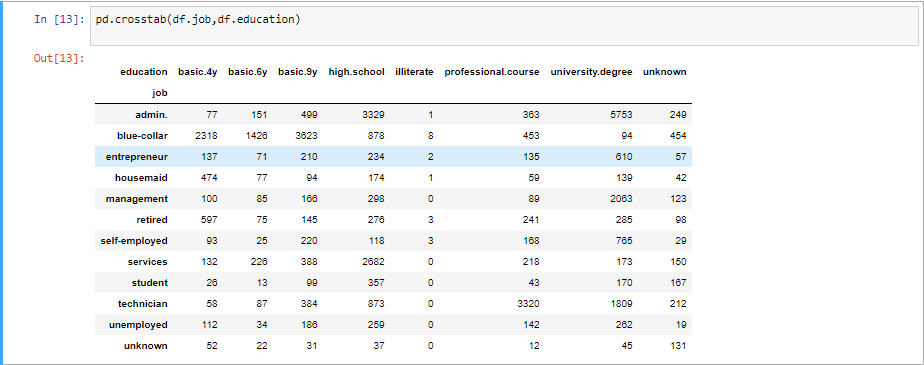
Other method is to infer the value of the unknown variable from the other variables. This a way of doing an imputation where we use other independent variables to infer the value of the missing variable. This doesn't guarantee that all missing values will be addressed but majority of them will have a reasonable which can be useful in the prediction.

Variables with unknown/missing values are: 'education', 'job', 'housing', 'loan', 'deafult', and 'marital'. But the significant ones are 'education', 'job', 'housing', and 'loan'. The number of unknowns for 'marital' is very low. The unknown for 'default' variable are considered to be recorded as unknown. It may be possible that customer is not willing to disclose this information to the banking representative. Hence the unknown value in 'default' is actually a separate value.

Therefore, we start with creating new variables for the unknown values in 'education', 'job', 'housing' and 'loan'. We do this to see if the values are missing at random or is there a pattern in the missing values

## Imputation:

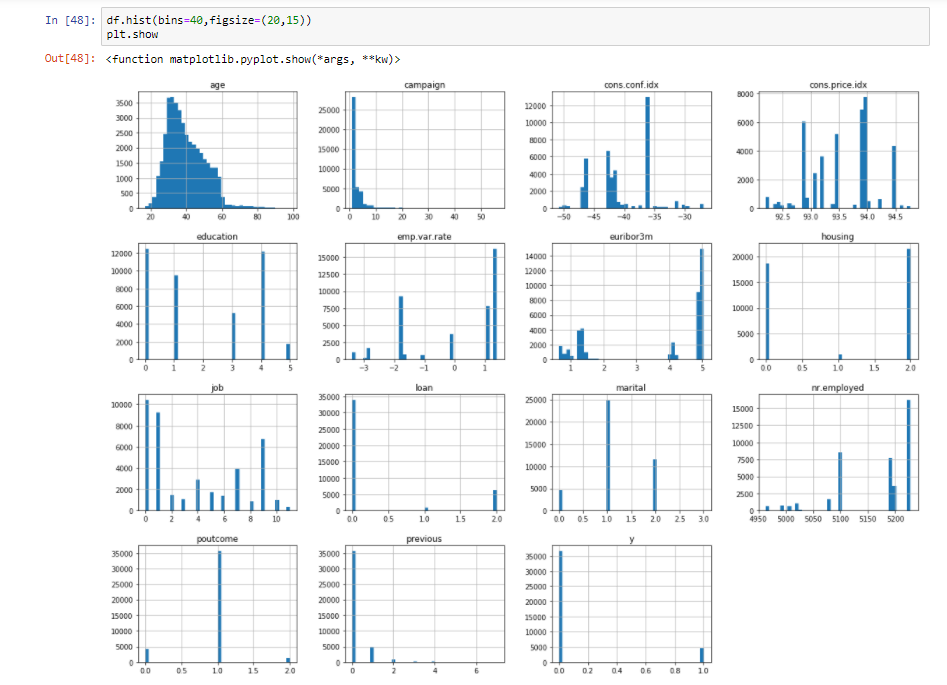
Now, to infer the missing values in 'job' and 'education', we make use of the cross-tabulation between 'job' and 'education'. Our hypothesis here is that 'job' is influenced by the 'education' of a person. Hence, we can infer 'job' based on the education of the person.



* Most of the blue-collar job are having the education basic.4y, basic.6y, basic.9y
* admin jobs have the highest university degree that is 5753

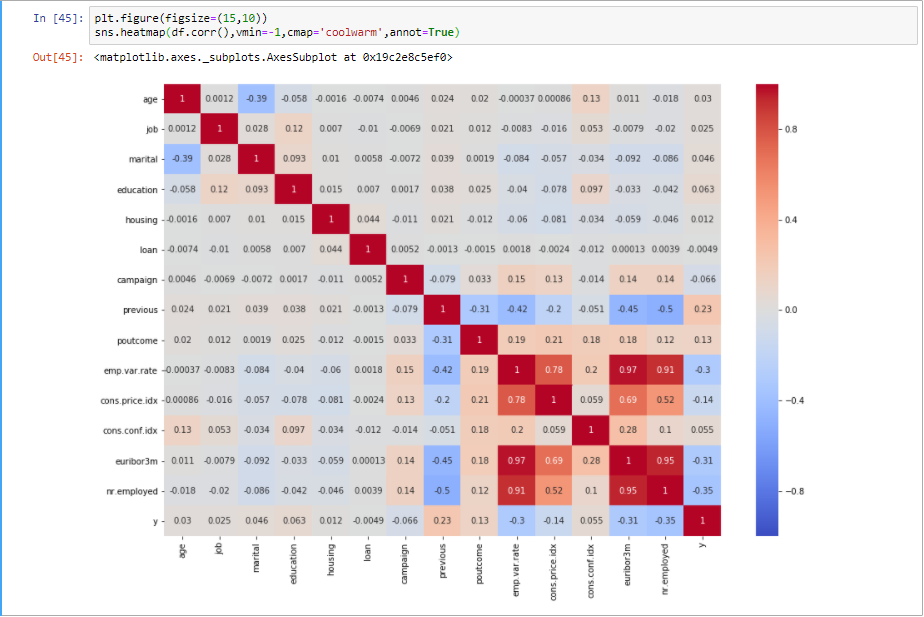
### Data Visualization

Visualising the data is an important step of the data analysis. With a graphical visualisation of the data we have a better understanding of the various features values distribution:



The correlation matrix is an important tool to understand the correlation between the different characteristics. The values range from -1 to 1 and the closer a value is to 1 the bettere correlation there is between two characteristics.

In [17]



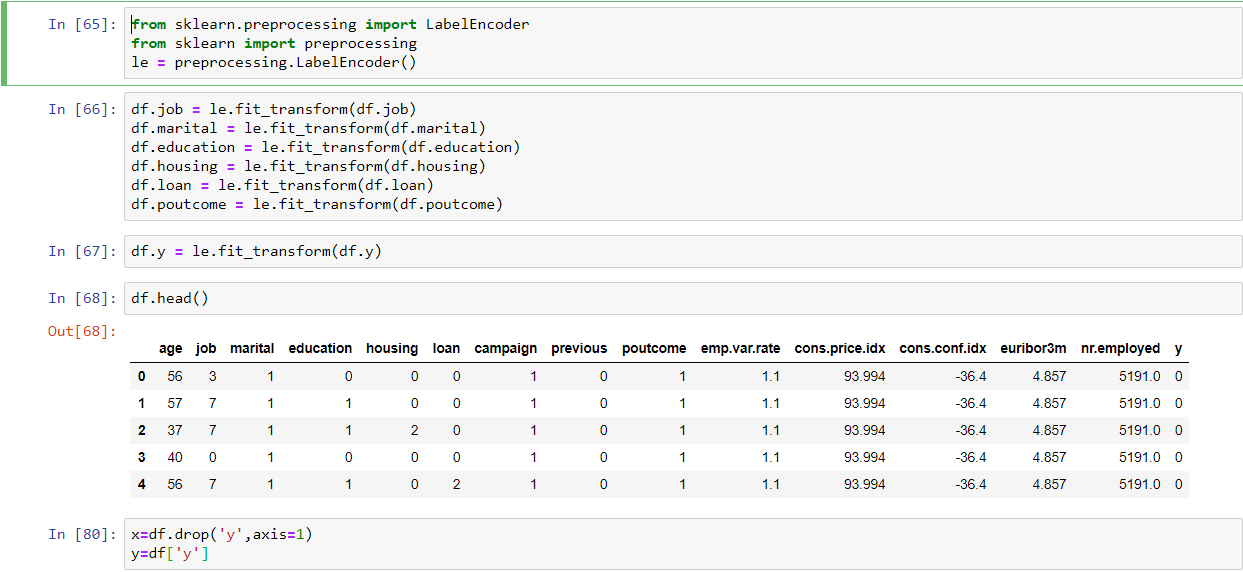
The emp.var.rate, cons.price.idx, euribor3m and nr.employed features have very high correlation. With euribor3m and nr.employed having the highest correlation of 0.95!

Now with that basic EDA done, let’s move forward to the next step which is Data Preprocessing.

### Pre Processing

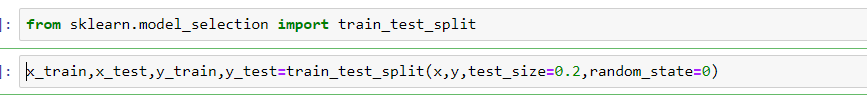
For encoding the levels of a categorical features into numeric values. LabelEncoder encode labels with value between 0 and n\_classes-1.

Converting Categorical variable into numeric using Label Encoder



### Train and Test split

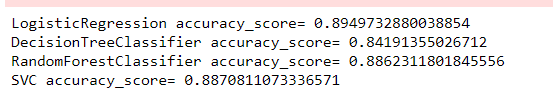
Now that the data has been prepared, we are ready to train our model and make predictions. Let’s first split the data into the training and testing sets:



Now it is very important to split your dataset into train, test and CV datasets,  Otherwise we will have a data leakage problem.

### Building Predictive Model

We will try four classification algorithms, i.e. Logistic Regression, Support Vector Machine, Decision Trees, and Random Forest, and then compute their accuracy score choose the classifier with the highest accuracy:



## Conclusion

We finally find accuracy score of 89.49% using Logistic Regression algorithm

# Results and Recommendation

* In "Duration" variable people are saying more yes. People having longer conversation with the bank showing the higher interset. The bank should foucs on the clients who have significant call duration.
* Agent should also target clients of job category "housemaid","services","technician"as these people are more against taking risk.
  + According to the plot for both logistic regression and random forest, we can tell that the most influential
  + variables are duration, nr.employed, euribor3m, and emp.var.rate.