**Machine Learning – Classification: Project Report**

**Title:** Classification problem to predict whether the client will subscribe to a term deposit.

**Group Number:** 02

**Participants List:** 1) Dharmil Sangani

2) Anoop Bose

**Problem Statement:** The dataset was sourced from the University of California, Irvine’s repository of machine learning datasets. The data is open source in nature under creative commons license.

The data is related with direct marketing campaigns of a Portuguese banking institution. 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.

The Dataset includes:

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.

**Overall Summary of your Solution:** The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y). With ‘y’ being discrete in nature, we have taken various Classification techniques to predict the dependent variable. The techniques used are ‘Logistic Regression’, ‘Support Vector Machine’ and ‘Random Forest Classifier’.

The project is started with loading the dataset and doing Exploratory Data Analysis so as to clean the data in order to get a good model. We then split the dataset into training and testing in the ratio 70:30. After doing this, the training data is (28330, 14) and testing data is (12357, 14).

We then use the classification techniques. For each of the three techniques, we have calculated ‘Accuracy’, ‘Precision’, ‘Recall’ and ‘F1-Score’. We have compared the three techniques using the ‘F1-Score’ and came to a conclusion that ‘Random Forest Classifier’ is the best amongst the three. The detailed analysis is explained in the points after this.

**Detailed description and Analysis:**

The data is related with direct marketing campaigns of a Portuguese banking institution. 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.

The dataset includes:   
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.

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

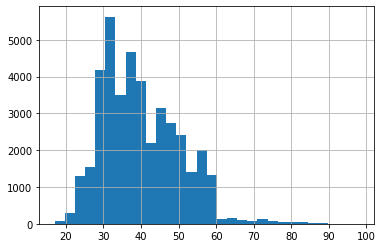
**Attribute Information:**

**Input variables:**

Bank client data:  
1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','bluecollar','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.  
Other attributes:  
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')

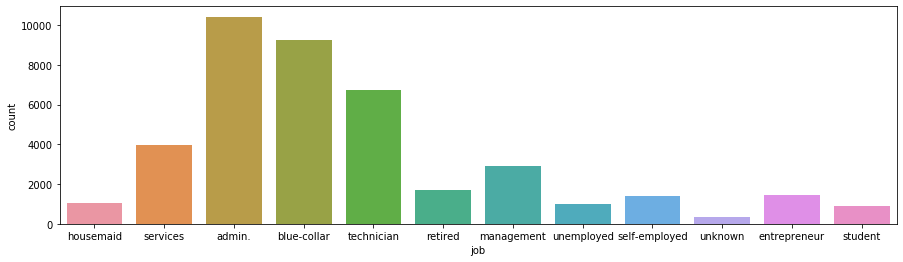
These are the variables/features. We start with the Exploratory Data Analysis. We first check for the null values. There are no null values in particular. However there are many ‘unknown’ values in many columns which I have treated. Following are the analysis of the specific variables;

**AGE:**

****

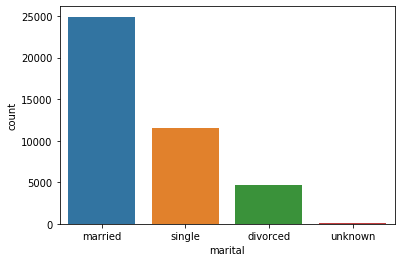
We can conclude that most of the people are in age group 25-50 years.

**JOB:**

****

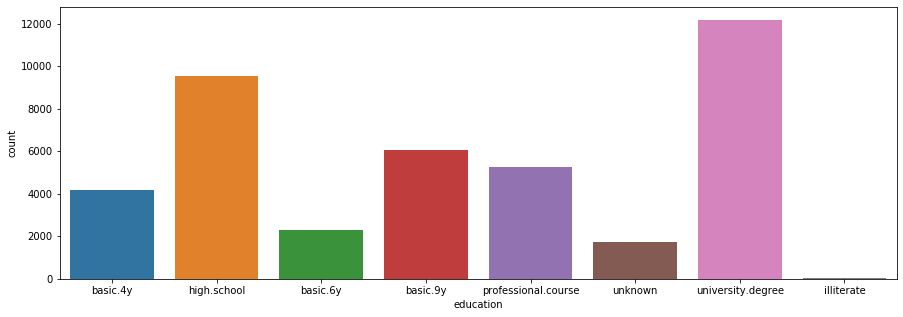
Most of the people in our dataset are working in ‘Services’, ‘Admin’, ‘Blue-collar’, ‘Technician’.

**MARITAL:**

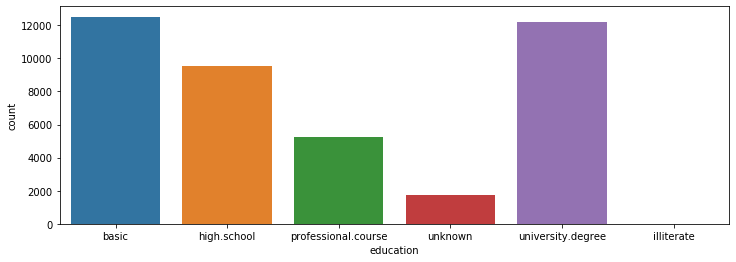
****

The married people in our datasets are approx. 25000.

**EDUCATION:**

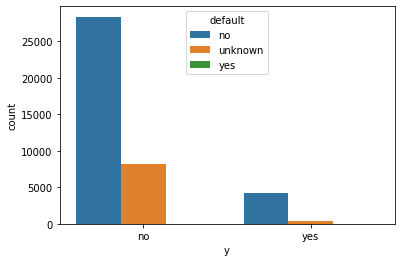
****

We have converted ‘basic.4y’, ‘basic.6y’, ‘basic.9y’ to ‘Basic’. The modified chart is shown below:



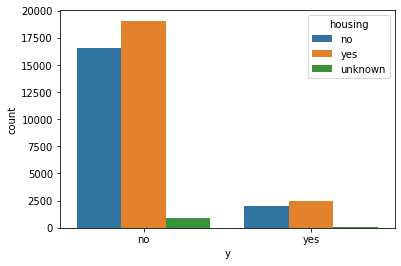
More than 12000 of the people have completed their basic education. We have 12000 people who University Degree and approx. 10000 who have completed High School.

**DEFAULT:** Has credit in default? (categorical: 'no','yes','unknown')



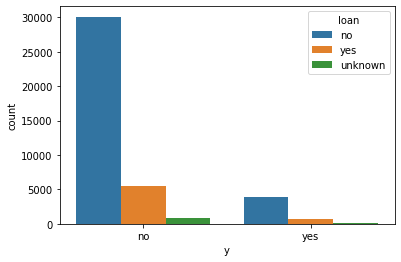
Most of the people have not taken credit. There are only 3 people who have taken credit. We impute the unknown as ‘NO’. Thus, the default column is less relevant as only 3 people have taken credit. So, we drop the ‘Default’ column.

**HOUSING:** has housing loan? (categorical: 'no','yes','unknown')



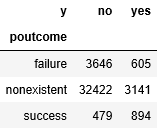
There are very less unknown columns. Also most of them have not taken the Housing loan. Therefore, we impute the ‘unknown’ with ‘NO’

**LOAN:** has personal loan? (categorical: 'no','yes','unknown')



Most of the people have not taken personal loan. Therefore, we impute the ‘unknown’ with ‘NO’.

**Poutcome:**



Since most of the values are nonexistent, we drop this column.

**Also, we drop the columns ‘Contact’, ‘month’, ‘day\_of\_week’, ‘pdays’ as they are of less importance.**

**Duration:** The data present in 'Duration' is in seconds. However we cannot take the column as it is as that will give more weightage to more number of seconds. Thus we use scaling. We do this by importing StandardScalar from sklearn.preprocessing.

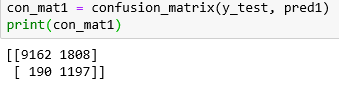
**Converting Categorical to Numerical:** We convert the categorical columns to numericals. The columns include ‘job’, ‘marital’, ‘education’, ‘housing’, ‘loan’ and target variable (‘y’). To do this, we import LabelEncoder from sklearn.preprocessing. We don’t use OneHotEncoder as it will lead to many features for each value of a particular column which may affect out model.

**TRAIN-TEST SPLIT:** We split the data in 70:30 ratio.

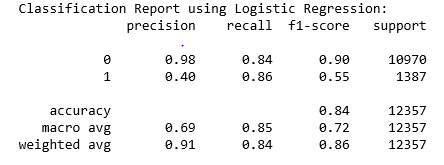
**MODEL BUILDING:**

1. **Logistic Regression.**

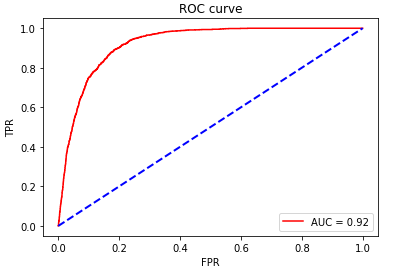
Confusion Matrix:



Classification Report:

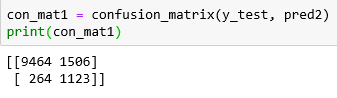


AUC-Curve:

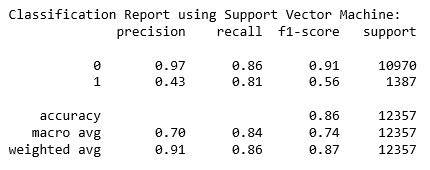


1. **Support Vector Machines:**

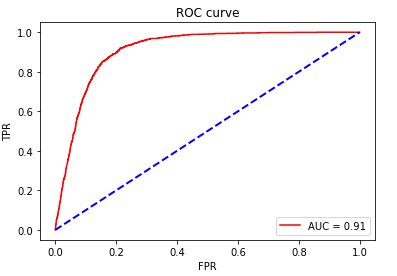
Confusion Matrix:



Classification Report:

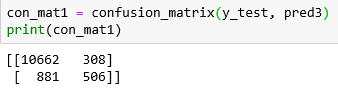


AUC-Curve:

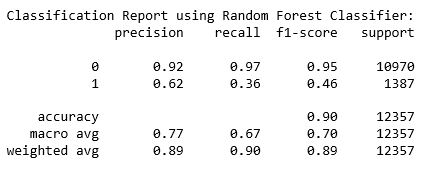


1. **Random Forest Classifier:**

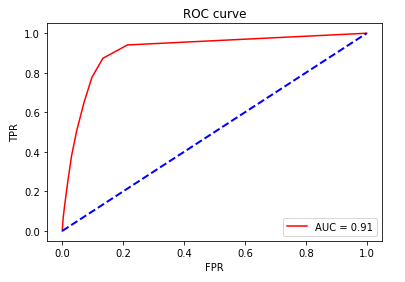
Confusion Matrix:



Classification Report:

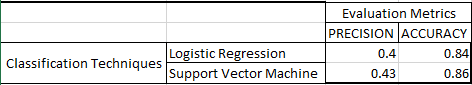


AUC-Curve:



**Identification of the best model backed by the logic for selecting it:**

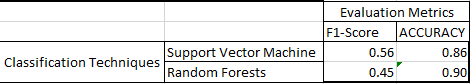
1. First, we compare Logistic Regression and Support Vector Machine. Since both of the models have more number of False Positive, we use ‘Precision’ to evaluate and compare. Our interest lies in class ‘1’. And the precision in SVM for Class1 is more than that of Logistic Regression. We have given the class weight as balanced. So we can compare the accuracy. SVM gives us a higher accuracy than Logistic Regression.



Note: The Precision value is as per our study of interest (Class 1)

1. We, now compare Support Vector Machines and Random Forests. For SVM, the number of False Negatives are more whereas for Random Forests, the number of False Positives are more. An appropriate evaluation metric will be the F1-score. Our interest lies in class ‘1’. The F1-score of Support Vector Machine is higher than Random Forest.

Here also we have given class\_weight as ‘balanced’. So we can compare the accuracy and Random Forest has higher accuracy.In our dataset, our interest lies in ‘True Positives’ and ‘True Negatives’. So we use the evaluation metric as ‘Accuracy’.



1. The AUC Curve for all the three techniques are almost same.
2. From the above three points, we can conclude that ‘Random Forest Classifier’ will be an appropriate Machine Learning Technique to solve the problem whether client will subscribe to the term deposit.

**Lessons learnt from the project:**

1. There are many techniques we can run for a classification problem. Each and every algorithm has its own pros and cons.
2. The algorithm we select depends on our dataset and the ratio of our classes of the target variables.
3. Exploratory Data Analysis has a big impact on the accuracy of the model. No ML technique can be used without doing EDA first.
4. Accuracy will be a good evaluation metric if the classes are balanced and we are interested in ‘True Positives’ and ‘True Negatives’.
5. Precision is useful when our False Positives are high.
6. Recall is useful when our False Negatives are high.
7. F1-score can also be used to evaluate as it takes the harmonic mean between precision and recall.