Portugal bank data analysis

Shashikesh Swarnkar

**Organization and Data Overview**

A bank in Portugal was rolling out a new term deposits product for its existing customers. They wanted to understand which of their existing customers to call and target for this new product so that ROI would be high. In the past they had connected to their customer base through phone calls for various other products. Results for these previous campaigns had been made available for over 50,000 customers. The record is provided in the form of dataset which contains annual track record of the customers characteristic, campaign characteristic, previous campaign information. The dataset of this campaign contains outcome of previous marketing campaign against the month they were contacted.

**Data Source**

The data set is collected from UCI (University of California Irvine) machine learning repository, and it is a hypothetical data for training purpose. The reason of choosing this data is that as being a hypothetical dataset, it has no error, no missing values. Thus, it will allow in performing all of testing, analysis it will help me in applying the statistics and regression analysis, which will be helpful in gaining more knowledge in data analytics field. This project will help in understanding the application of statistical models, regression analysis and other sort of analysis.

**Challenges**

Customers have recently started to complain that bank’s marketing staff bothers them with irrelevant product calls, and this should immediately stop.

There is no prior framework for manager to decide and choose which customer to call and which one to leave alone.

**Data variables**

* 1 -Age (numeric)
* 2 - Job: type of job (categorical: “admin.”, “unknown”, “unemployed”, “management”, “housemaid”, “entrepreneur”, “student”, “blue-collar”, “self-employed”, “retired”, “technician”, “services”)
* 3 - Marital: marital status (categorical: “married”, “divorced”, “single”; note: “divorced” means divorced or widowed)
* 4 - Education (categorical: “unknown”, “secondary”, “primary”, “tertiary”)
* 5 - Default: has credit in default? (binary: “yes”, “no”)
* 6 - Balance: average yearly balance, in euros (numeric)
* 7 - Housing: has housing loan? (binary: “yes”, “no”)
* 8 - Loan: has personal loan? (binary: “yes”, “no”)

Related with the last contact of the current campaign:

* 9 - Contact: contact communication type (categorical: “unknown”, “telephone”, “cellular”)
* 10 - Day: last contact day of the month (numeric)

Direct Marketing Campaign: Details and Phase I Tasks

* 11 - Month: last contact month of year (categorical: “Jan”, “Feb”, “Mar”, . . ., “Nov”, “Dec”)
* 12 - Duration: last contact duration, in seconds (numeric)

other attributes:

* 13 - Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
* 14 - Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
* 15 - Previous: number of contacts performed before this campaign and for this client (numeric)
* 16 - Poutcome: outcome of the previous marketing campaign (categorical: “unknown”, “other”, “failure”, “success”)

Output variable (desired target):

* 17 - y - has the client subscribed a term deposit? (binary: “yes”, “no”)

**Implementation**

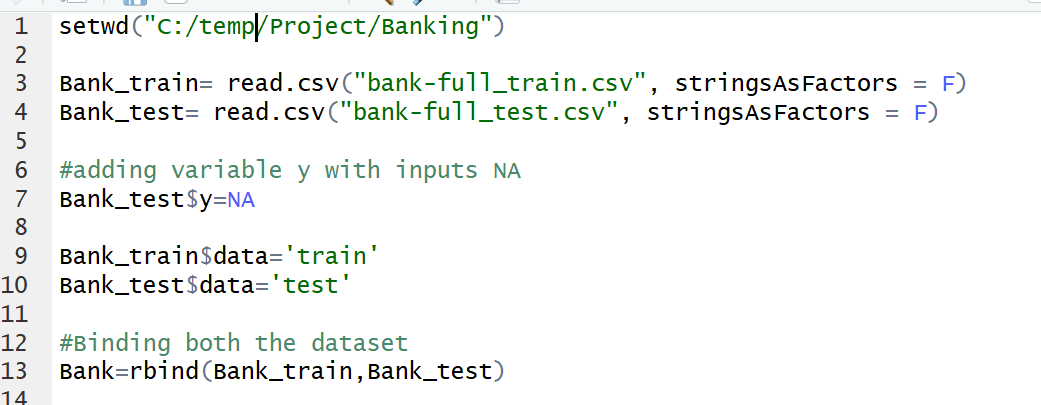
The dataset is cleaned and modelled in R studio. The data consists of 2 .csv files: Bank-full\_train.csv and Bank-full\_test.csv.



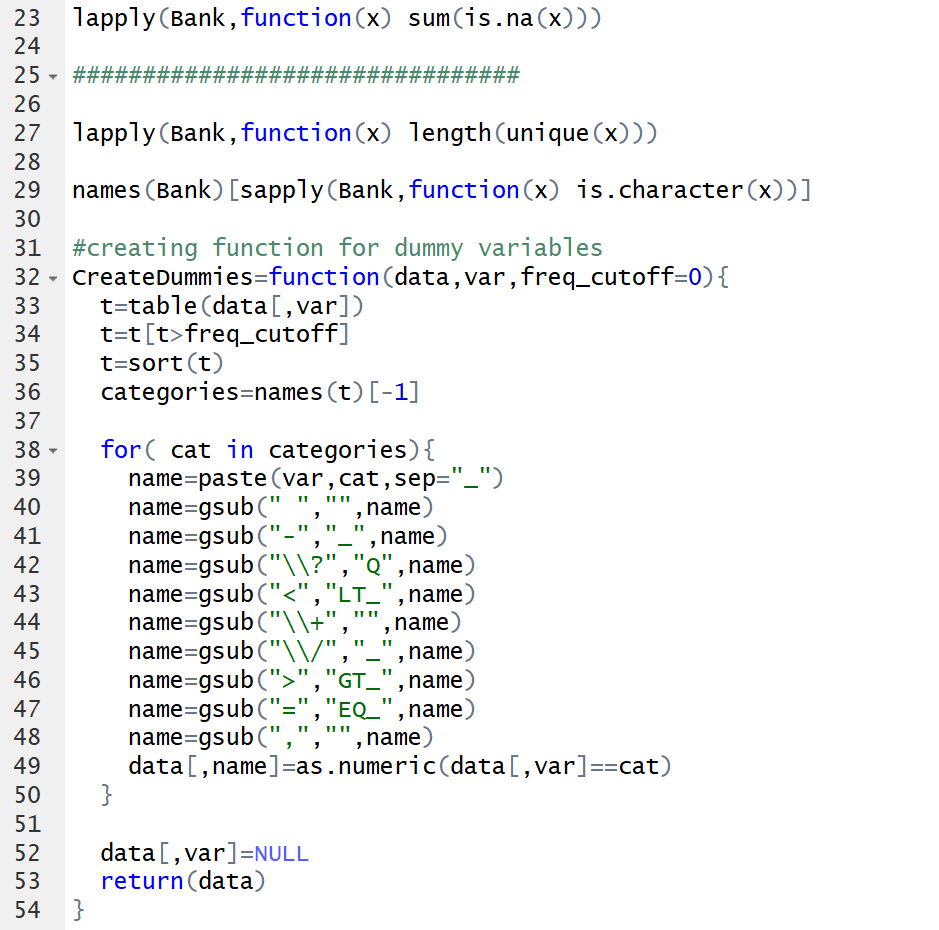
The train data has the output variable ‘y’ whereas the test data does not have output variable ‘y’. The implementation plan includes creating model and training it on train data, then deploying the model on test data to generate the output variable. As the output variable is binary, thus making Logistic Regression best fit for this dataset.

Following are the steps taken to achieve the logistic regression:

* Imported both the .csv files into R studio. After importing the train and test data into R studio, created a new column for test dataset (Bank\_test) as output variable ‘y’ and imputed values as ‘NA’. After adding variable ‘y’ to test data the binding of test and train data would be easy. Now, binding of test and train has been done, which saves time and effort as there is no need to clean both the datasets individually. Before binding the 2 datasets, it is necessary to add some identifier to know which observations are from train data or test data, this is achieved by adding another variable ‘data’ and it has 2 values ‘train’ and ‘test’. The observations from train data has value ‘train’ and observations from test has value ‘test’.

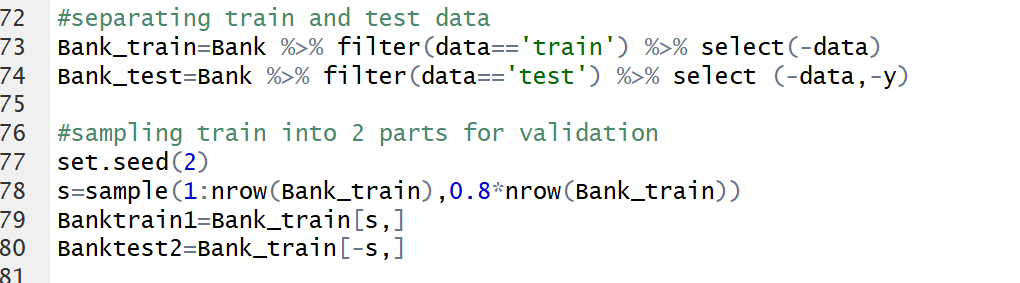


* After binding, numerical data needs to be handled. The numerical variables except ‘pdays’ do not need any changes. The variable ‘pdays’ has value -1, which means that the customer was not contacted previously. Thus, -1 can be converted to 0 which would be easier to handle. Thus, converting the -1 to 0 and keeping the rest of the values as it is.
* Next, we need to handle categorical data. As, regression models need numeric variables. Thus, we need to create dummy numeric variables out of categorical variables present in the data set. The categorical data is viable for any model as considering just numeric values won’t provide the best model as it is only based on some part of data, whereas keeping the categorical values in the data provides better understanding and gives more variable to generate better model. Following is the function to do so.

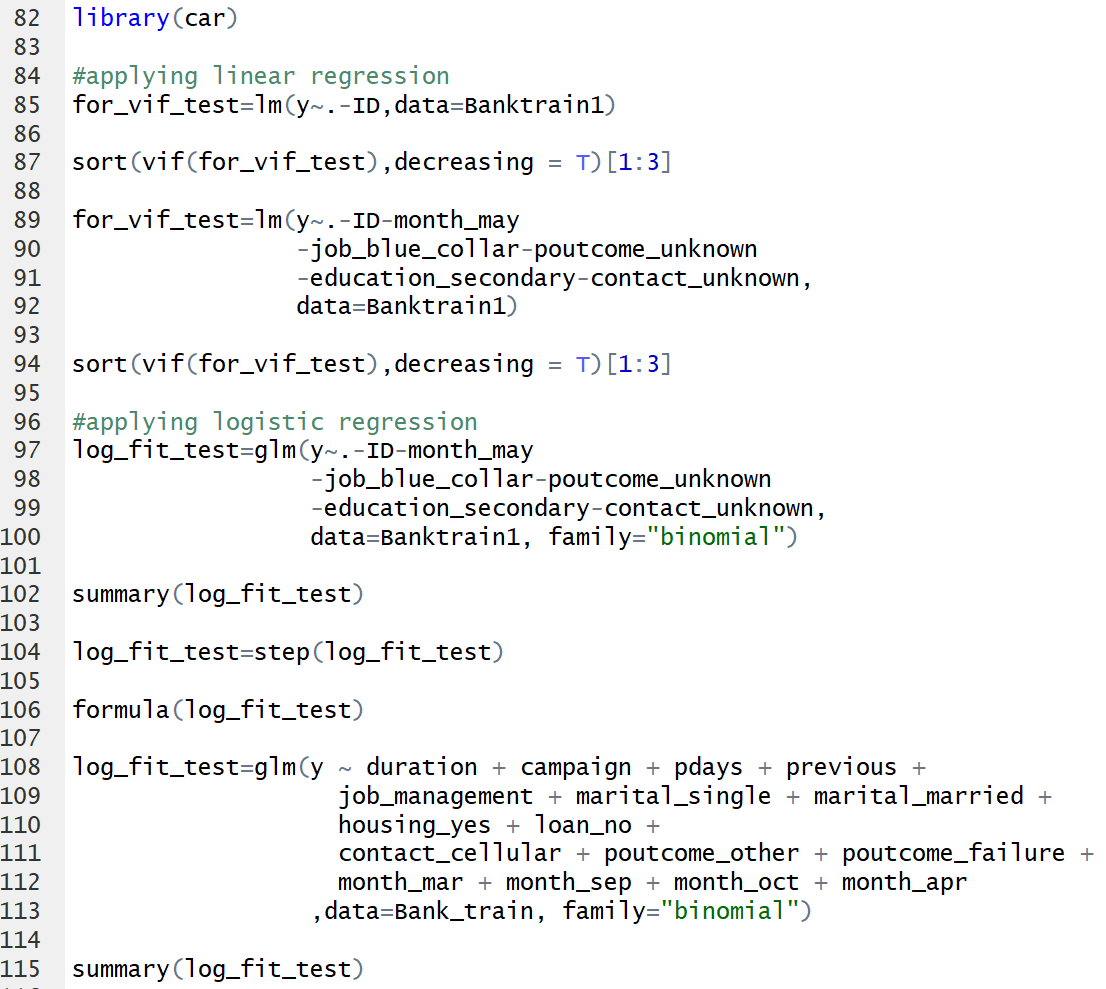




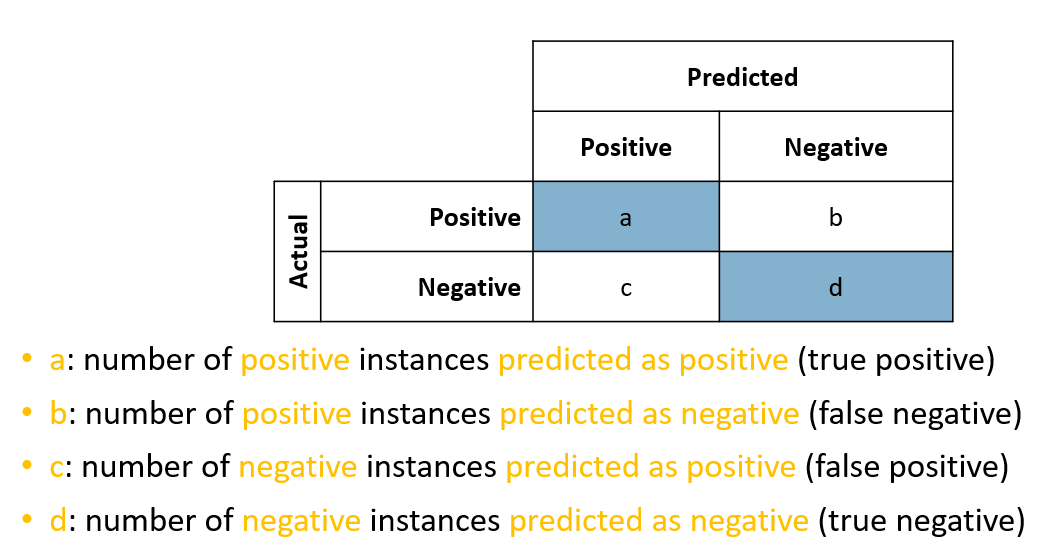
* As the output variable is a character variable having values either ‘yes’ or ‘no’. Thus, converting it to binary variable, 1 represents yes and 0 represents no.
* Now the data has been cleaned, for validation purpose the train datasets (Bank\_train) has been separated into 2 parts: 80% of the train dataset (Banktrain1) is used to build model and remaining to test the model (Banktest2).



* Collinearity can be defined as two variables being highly linear to one another. This leads to unstable regression estimates and high error. To avoid such situation VIF (Variation inflation factor) is used. As it measures the variance inflation among predictors that exists due to collinearity. If VIF is 1 for a give variable among other predictor variables then there is no collinearity between them, if VIF is about 4 for given variable among other predictor variables then it needs some investigation and if the VIF is 10 then collinearity is very high and needs correction.
* To address collinearity, applied linear regression to see how much each predictor variable is linear to one another and removing the highly linear variables based on AIC (Akaike Information Criterion is used to get approximation of relative quality for a given dataset) and p value.



* After the variables are finalized for the model, logistic regression is applied on 80% of the train data (Banktrain1) and tested on remaining 20% of the data (Banktest2).
* After validating the model, the logistic regression is applied on complete train data (Bank\_train) and predicted ‘y’ variable for hypothetical dataset (Bank\_test). As predicted ‘y’ variable in the form of probability (between 0 and 1). Thus, it is necessary to convert these values into appropriate 0’s and 1’s. This has been done by using confusion matrix.



Following are the terms for Confusion matrix:

Accuracy: Proportion of correct predictions.

Sensitivity (true positive rate TP): Proportion of positive cases correctly classified as positive.

Specificity (true negative rate TN): Proportion of negative cases correctly classified as negative.

False positive rate FP: Proportion of negative cases incorrectly classified as positive.

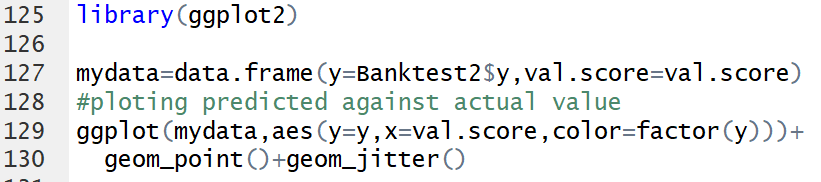
False negative rate FN: Proportion of positive cases incorrectly classified as negative.

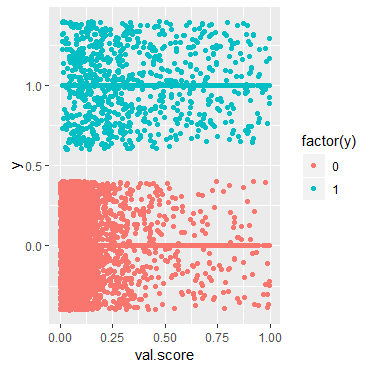
Precision: Proportion of predicted positive cases that were correctly classified.

Recall: Recall is percentage of prediction that was recalled successfully.

**Results**

* Following is the code used to plot point graph 1 of Banktest2 data.

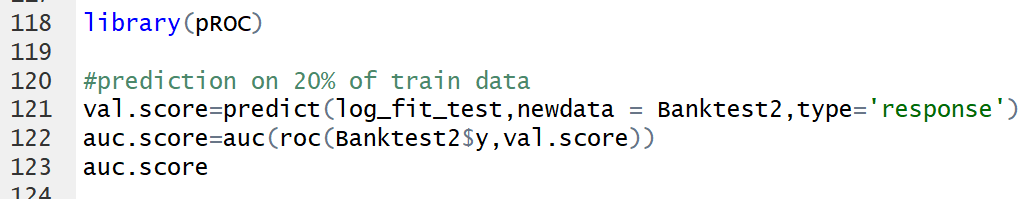




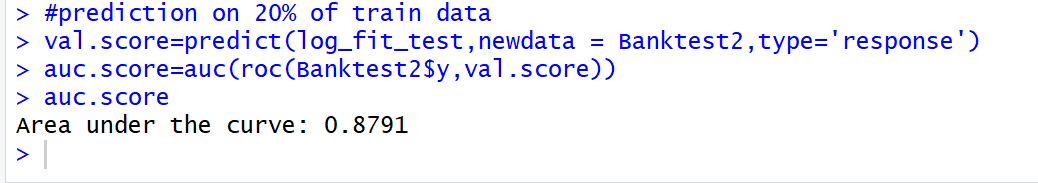
Graph 1. Point graph

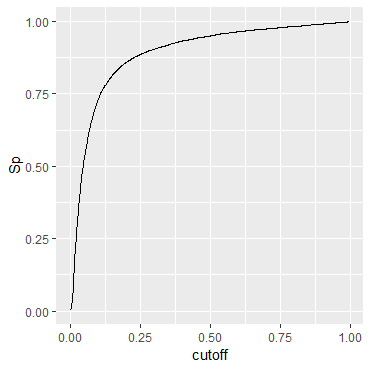
As per the above graph, responses 0 are bunched around high score and responses 1 are bunched around low scores. By observing the above graph values accumulated below 0.5 would be considered 0 and above 0.5 as 1.

* Following is the code for AUC and ROC (Graph 2) for Banktest2.



Provided is the AUC output on the test data (Banktest2) which was used for validation. As AUC is about 88% which states that the model is appropriate and can be used on future datasets.

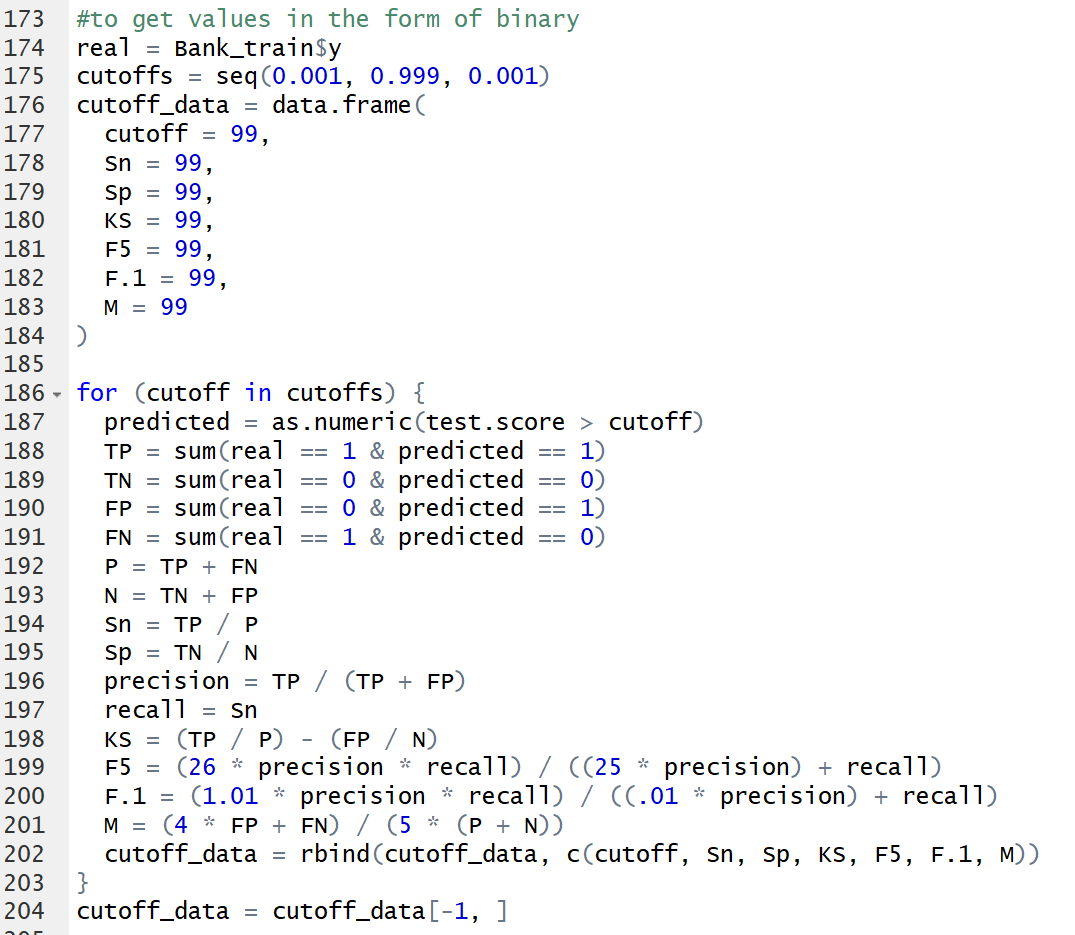


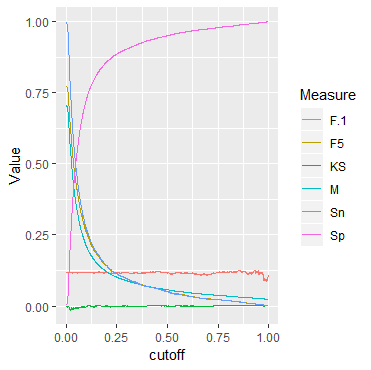


Graph 2. ROC graph

Note: The above results are of Banktest2 (20% of the Bank\_train data) because it has both the ‘y’ variables i.e. predicted and the real.

* Following is the code for cutoff, using the four basic numbers, we have calculated the other performance matrices values (KS, M, F.1, F5) to decide cutoff on Bank\_test.





Graph 3. Performance matrix

**Analysis/ Research Question/ Hypothesis**

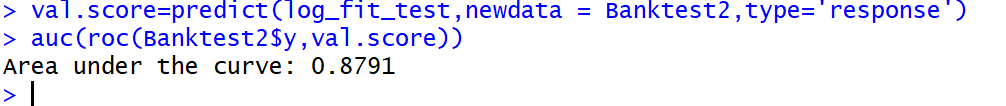
* **Predict ‘y’ i.e. has the client subscribed a term deposit" for the bank for the year 2018-2019.**



Based on the explanation above, please find predicted y values on hypothetical test data (Bank\_test).

* **Identify accuracy and ROI of the response from the dataset.**

AUC (Area Under the Curve) is used to understand the accuracy of the ‘y’ variable based on the modelling that has been created on training data.



AUC is 0.8791 which translates to 87.91% of accuracy of model.

Accuracy is used to understand how well the model performs on test data by comparing the output variable ‘y’ of actual results to output result. In this case model produced about 80% of accurate results which is calculate by predicted ‘y’ that are accurate divided by total number of observations and ROI about 68% (this means about 68% customers may enroll to future campaigns, who previously disagreed) which is beneficial for the bank, as ROI and accuracy of the model is within mentioned range. This proves that the model is efficient to use for future datasets. Thus, it can be used for future campaigns and save advertisement cost.

* **Which customer faced success as the outcome of previous marketing campaign?**

About 1000 customers were happy with previous marketing campaigns.

* **Which campaign has average response to yes?**

Campaign 1 has most numbers of positive responses (59 times).

**Conclusion**

The project helped in getting hands on experience with R studio which would be beneficial in future. As the dataset is a hypothetical data, for future projects I would like to select real time datasets which will be more challenging and provide more opportunities to gain knowledge on how such datasets are handled in real life scenarios. Also, I would like to apply different modelling techniques to cross check the performance of each model to have higher accuracy and efficiency.

Reference

Collinearity Diagnostics, Model Fit & Variable Contribution (n.d.) Retrieved from

<https://cran.r-project.org/web/packages/olsrr/vignettes/regression_diagnostics.html>

Akaike information criterion (n.d.) Retrieved 11th March 2019.

<https://en.wikipedia.org/wiki/Akaike_information_criterion>