INTERIM PROJECT REPORT

11/25/2021

Jai Bhala
Yash Dodia
Sunny Grover
Deeba Haider
Karan Mehta
Anshuman Pradhan

We will be using the data set **bank-full.csv** from UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Source:
[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

The data is related with direct marketing campaigns (phone marketing) of a Portuguese banking institution. The data provides insights to client attributes, contact attributes and campaign attributes and if the bank term deposit product was subscribed or not. Dataset contains 45211 rows and 17 columns. Variables details are as below:

Seq	Variable Name	Description	Possible values
1	age	age (numeric)	
2	job	type of job	'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'
3	marital	marital status	'divorced', 'married', 'single', 'unknown'
4	education	education level	'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'
5	default	has credit in default?	'no', 'yes', 'unknown'
6	balance	balance	
7	housing	has housing loan?	'no', 'yes', 'unknown'
8	loan	has personal loan?	'no', 'yes', 'unknown'
9	contact	contact communication type	'cellular','telephone'
10	day	last contact day of the week	
11	month	last contact month of year	
12	duration	last contact duration, in seconds	
13	campaign	number of contacts performed during this campaign and for this client	
14	pdays	number of days that passed by after the client was last contacted from a previous campaign	
15	previous	number of contacts performed before this campaign and for this client	
16	poutcome	outcome of the previous marketing campaign	'failure', 'nonexistent', 'success'
17	y	has the client subscribed a term deposit?	'yes' ,'no'

The motivation for choosing this dataset is to understand the effectiveness of marketing campaigns and plan future campaigns accordingly for better reach and outcome.

The main objective is to predict if a particular campaign call will result into bank product (term deposit) being subscribed. We will be using various visualization techniques like Pivot table, Bar chart and Heatmap to explore the data and clean and pre-process data, if required.
The variable of interest is column 'Y'. We will be trying out different models using kNN, Classification Tree, and Logistic Regression and comparing these models based on Accuracy, Sensitivity and Specificity to find the best fit model.
Even in kNN, Classification Tree and Logit Regression, we will be using various models amongst those three. For example, we will be predicting the kNN with different k values and choosing the one with the highest accuracy, specificity, and lowest sensitivity. In Classification Trees, we will be predicting by cp value. After that, we will be using library **randomforest** and predict through it.

However, we will be choosing the best model amongst all 3 and then compare it with each other to select the best model to predict the effectiveness of campaign and understand the important variables to improve future campaigns.