**Direct Marketing Campaign: Details and Part I Tasks**

A Portugese bank is rolling out term deposit for its customers. They have in the past connected to their

customer base through phone calls. Results for these previous campaigns were recorded and have been

provided to the current campaign manager to use the same in making this campaign more effective.

**Challenges that the manager faces are following:**

• 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 her decide and choose which customer to call and which one to leave

alone

She has decided to use past data to automate this decision, instead of manually choosing through each

and every customer. Previous campaign data which has been made available to her; contains customer

characteristics , campaign characteristics, previous campaign information as well as whether customer ended

up subscribing to the product as a result of that campaign or not.

Using this she plans to develop a statistical model which given this information predicts whether customer in

question will subscribe to the product or not. A successful model which is able to do this, will make her

campaign efficiently targeted and less bothering to uninterested customers.

Help her build this model . Data details are given as below

1 - age (numeric)

2 - job : type of job (categorical: “admin.”,“unknown”,“unemployed”,“management”,“housemaid”,“entrepre

neur”,“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)

1

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”)

**Part 1 Tasks**

1. Import File “bank-full.csv” , ensure that categorical variables are imported as characters , not factors.

Take a quick look at type of variables in the data.[20]

2. Find out mean ,std deviations, q1 , q3 and IQR for the variables age and balance. Check if they are

following normal distribution. If they are following normal distribution then calculate outlier limits as :

mean ± 3\*sd, otherwise calculate outlier limits as [q1 - 1.5 \* IQR, q3 + 1.5 \* IQR]. Find out which

observations exceed these limits, remove them from the data. [20]

3. Prepare a percentage cross table for job and y & month and y [separately]. Round of percentages to

two digist. End Results will looke like this. Notice that the percentages for each job category add up to

one. This shows relative frequency of your resposne across job categories. You need to get similar table

for months as well. [20]

## y

## job no yes

## admin. 0.88 0.12

## blue-collar 0.93 0.07

## entrepreneur 0.92 0.08

## housemaid 0.91 0.09

## management 0.86 0.14

## retired 0.77 0.23

## self-employed 0.88 0.12

## services 0.91 0.09

## student 0.71 0.29

## technician 0.89 0.11

## unemployed 0.84 0.16

## unknown 0.88 0.12

Job categories which have realtive similar percentages can be merged together to reduce number levels. Take

a call on which job categories to be merged together. Create a new categorical variable from job with fewer

levels with that consideration. Do same for variable month.

4. Bonus Question [ it is not mandatory to attempt this, if you do , you get bonus marks ]: Write a

function which takes a dataset and categorical variable names as input and returns dataset with dummy

variables for that categorical variable. [20]

5. Create pi-chart for the variable education with response y as fill. [20]

6. Using function above [written in Q4] or otherwise create dummy variables for the categorical variables.

**Part 2 Tasks**

1. Break your data in to random parts train and test, train should contina 70% of the observation. Make

sure this random sampling is reproducible. [10]

2. Remove predictor variables with VIF>5 from the train data. [10]

3. Build logistic regression model for the response y with remaining variables. Use function step to drop

insignificant variables from your model. Check if p-values in resulting model for all the variables are

less than 0.05. If they are not , get rid of such variables sequentially. For thus obtained final model, get

scores and save them in the train data.[30]

4. Get cutoff of the score using KS method. Check performance of the model thus obtained on the test

data.[20]

5. Build a random Forest model on the same train data and report if performance of this model is better

than logistic regression model on the test data. Get a variable importance plot. [30]

6. Bonus Question: [ it is not mandatory to attempt this, if you do , you get bonus marks ] Pick top

6 variables from the variable importance plot and build a logistic regression model with just those

variables [ get cutoff using KS method]. See how it performs on the test data. [20]