Assignment: BU51039-: Predictive and Prescriptive Analytics

Part B: Predictive Analytics (60 % of total weight)

You have been provided with marketing data from a bank that has been making phone calls to customers asking them to subscribe a product service.

There are two datasets: **train.csv** and **test.csv.** These are respectively the training (70%) and test data (30%). These made available in the folder Assessment B.

Often, more than one contact with the same client was required to assess if the customer would subscribe. The objective of the analyst is to develop a system to predict whether an individual will subscribe to this new line of product. The bank will then use this system to target individuals who are more likely to subscribe.

The data provided contains 17 columns and the data in them can be broken down into 4 categories:

Category	**Columns**	**Further description**
Customer Data	1-8	Personal details on the client contacted
Contact Data	9-12	Details of most recent contact made to this client
Contact Summary Data	13-16	Details of previous contact to the client
Result	17 (**column Q**)	Did the customer Subscribe to the new product

Field #	**Field Name**	**Description**	**Data Type**
1	age	Age	numeric
2	job	type of job	categorical
3	marital	marital status	categorical
4	education	education	categorical
5	default	has credit in default?	categorical
6	Balance	Balance on the current account	numeric
7	housing	has housing loan?	categorical
8	loan	has personal loan?	categorical
9	contact	contact communication type	categorical
10	day-of_the_week	contact day of year	categorical
11	month	contact month of the week	categorical
12	duration	contact duration, in seconds	numeric
13	campaign	number of contacts performed during this campaign and for this client	numeric, includes last contact
14	pdays	number of days since the client was last contacted from a previous campaign	numeric; 999 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
17 **TARGET VARIABLE**	y	has the client subscribed to the new product?	Binary (1 and 0)

Variables 1-17 have been used in the paper when the data has been published~~:~~. The full dataset (bank-additional-full.csv) was described and analyzed in:

- S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.001.

  Available at: [pdf] http://dx.doi.org/10.1016/j.dss.2014.03.001

  [bib] http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt

The target variable is **y** which can take two values **0** (**Not subscribed**) and **1** (**subscribed**). The other variables in the first 16 columns are *potential* features.

Tasks

You are tasked to build up a model for the outcome (y) using the data you have been provided. You would like to come up with a model or some models that does a good job at forecasting. You may also want to consider what the main features (X) that are important in predicting the outcome variable. This is because the marketing team would like to know what are the characteristics of individuals they should be looking for in order to target these individuals in order to maximise the subscriptions of these models.

Your report (no more than 2000 words) may be structured into the following headings with your own subheadings:

Introduction

Methods

Results

Corporate Purpose.

Conclusion

Submission Instructions

The submission for this coursework requires 2 files:

(i) The report in word or pdf format and;

(ii) A software script which contains the code and comments for your project. This will enable replicability of your results when marking. For instance, if you estimate your model using **R**, you will want to submit your R script (or RMD file). The same applies if you use a **python script**, **stata do file**, **Eviews command script**.

You are welcome to use Microsoft Word, latex or RStudio to write your report.

If you write your report in Microsoft word, please follow the following instructions:

- **Font:** Times New Roman
- **Font size:** 12
- **Line and Paragraph spacing:** 2
- Justify text to both the left and right margins
- All pages should be **numbered**
- Any referencing style can be used

Grading

The grade shall be based on the following factors:

1. Your report is well written and understandable (60 %)
2. Script for replicability (10 %)
3. The reliability of your model (30 %) to predicting the outcome variable (y).

Suggestions

You are provided with you some SUGGESTIONS on contents you could write in each section. You are more than welcome not to follow these suggestions and adopt your own.

Introduction:

- What is the purpose of this study? Why is it important for the bank?
- What is the problem at hand?
- What are the algorithms/methods the report will be using?
- How will the forecasts be assessed?

Methods

- Usually, you may want to write this section last (If you have tried various models and want to explain only one (or possibly more methods).
- You don't need to get too technical – as we are simply looking at the **application** of these methods/models.
- You may also want to describe the measures of forecast evaluation you will be using.
- (The field of data science comes up with many different algorithms/models/approaches on a regular basis. You can use a model that has not been discussed in class. But please make sure you reference the paper).

Results

- In this section you interpret the output of your model.
- If you adopt a model where you include the features – you may want to explain why it would seem these features work well (and why others not).
- How does your model(s) perform in terms of forecasting ?
- You may also want to compare the models you have used against other competing models.

Corporate Purpose and limitations of study

- Explain how these numbers can translate into a revenue making operation for the bank.
- Reflect on the suitability of the dataset for the purpose of identifying potential money-making opportunities for the bank. For instance, you can address limitations of the datasets.

Deadline

This task will be made available to you on the 17th March and will have a deadline on the **07h April 11:45**.