Group Name: Bank Marketing Campaign

Members:

Sebastian Bucheli, smbucheli@gmail.com, Ecuador, Fundación Ecociencia, Data Science.

Nolan Pilozo-Hibbit, npilozo@hotmail.com, Canada, University of Waterloo, Data Science

Problem description

There is information about a marketing campaign of a Portuguese bank and they have data of phone calls. The main objective is to predict -with the collected data- if a person will subscribe to the product.

Data understanding

The data is related to 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 assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The data being used has been compiled into one larger data set. The four datasets are:

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 in [Moro et al., 2014]

2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.

3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).

4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

What type of data you have got for analysis

Attribute Information:

Input variables:

bank client data:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue-collar','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')

Attribute type breakdown:

Attribute set	Numeric	Categorical	Binary
Bank client data	2	10	0
Other attributes	3	1	0
Social and economic context attributes	5	0	0
Output	0	0	1
Total	10	11	1

What are the problems in the data (number of NA values, outliers , skewed etc)

There are a lot of categorical data, making it difficult to find correlation between the various variables. Within these categorical variables exists another issue of a lot of these variables containing the value ‘unknown’ to fill in areas where the answer to an

attribute question is unknown (i.e. do they have a housing loan, what is their highest level of education, etc…).

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

This is a very clean dataset provided by UCI Machine Learning and the problems are not in NA values.

As was described above the dataset has a lot of categorical variables and outliers are not a big problem to deal with.

The main overcome problems are to find patterns among the huge number of categorical features and dealing with imbalanced data.

The approach we are applying by first is to encode the categorical data into numeric arrays because we can't perform a model with raw categorical features, and then obtain the ratio of imbalanced data for applying the correct parameters in the models. We are going to try an ensemble method for the model and also a clustering model because these methods work well with classification problems.