Apache Spark—Real Time Project—Marketing Analysis

Analyze marketing data for call campaign by bank

Try yourself!

Your client—a Portuguese banking institution—ran a marketing campaign to convince potential customers to invest in bank term deposit.

Information related to direct marketing campaigns of the bank are as follows.

The marketing campaigns were based on phone calls. Often, the same customer was contacted more than once through phone, in order to assess if they would want to subscribe to the bank term deposit or not. The data fields are:

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 to the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular', 'telephone')

9 - month: Month of last contact (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 (example, 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 times a customer was contacted during the campaign (numeric, includes last contact)

13 - pdays: number of days passed after the customer was last contacted from a previous campaign (numeric; 999 means customer was not previously contacted)

14 - previous: number of times the customer was contacted prior to (or before) this campaign (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 customer subscribed a term deposit? (binary: 'yes', 'no')

The data size is huge and the Marketing team has asked you to use Spark to help them get answers for the following questions:

1. Load data and create Spark data frame

2. Give marketing success rate. (No. of people subscribed / total no. of entries)

2a Give marketing failure rate

3. Maximum, Mean, and Minimum age of average targeted customer

4. Check quality of customers by checking average balance, median balance of customers

5. Check if age matters in marketing subscription for deposit

6. Check if marital status mattered for subscription to deposit.

7. Check if age and marital status together mattered for subscription to deposit scheme

8. Do feature engineering for column—age and find right age effect on campaign

The total time required to complete this task is 8 hours.

Note: The dataset required for this project can be accessed either from main folder or downloaded from the “Download Center”.