Data Science Internship 2022

Data Science: Bank Marketing (Campaign)

Submitted by: Big Analytics Group

Group Members:

Names	Email	College/Company
Taimoor Razi	taimoor.r10@gmail.com	Middle East Technical University, Turkey.
Ogwu Augustine	ogwuaugust@gmail.com	University of Jos, Nigeria.
Akshar Chaklashiya	chaklashiya.akshar@gmail.com	Lambton College, Toronto.
Sabrawi Hussein	Sabrawi.hussein@gmail.com	Everteam Company.

Submitted to: Data Glacier

Due Date: 19th August 2022

Table of Content

Project Lifecycle

Tasks

- Business Understanding
- Data understanding
- Exploratory data Analysis
- Data Preparation
- Model Selection & Model Building
- Performance reporting
- Deploy the model
- Converting ML metrics into Business metric and explaining result to business
- Presentation for non-technical persons.

Project Deadline

· 30th September 2022

Business Understanding

Problem Statement

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Why ML Model

ABC Bank wants to use ML model to shortlist customers whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only on those customers whose chances of buying the product is more.

This will save resources and their time (which is directly involved in the cost (resource billing).

Data Understanding

Dataset Information

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 classification goal is to predict if the client will subscribe (yes/no) to a term deposit (variable y).

Data Intake Report

Group Name: Big Analytics

Report date: 16-08-2022

Internship Batch: LISUM11: 30

Version: 1.0

Data intake by: Taimoor Razi

Data intake reviewer: NA

Data storage location: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Tabular data details:

Total number of observations	45211
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	4.39 MB

Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	.csv
Size of the data	5.56 MB

Proposed Approach:

· The data is downloaded from the UCI Machine Learning Repository.

· The bank-full dataset has no null or duplicate values. The bank-additional-full has no null values but has 12 duplicates. These 12 duplicates were removed.

· Both the datasets (bank-full and bank-additional-full) are appended together.

· The resulting dataset does not contain any duplicate values. However, null-values are created after combining both the datasets as there are some additional features/columns that are present in the bank-additional-full dataset and not in the bank-full dataset.

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 a housing loan? (categorical: 'no','yes','unknown')

7 - loan: has a 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')

References

Repository Link

[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