Deliverable – Week 8

Group Name: Tony's group

Team:

Name: Tony Nguyen

Email: tonynguyen9707@gmail.com

Country: United States

College: University of Florida

Specialization: Data Analyst

Problem Description

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).

Data Understanding

The data comes from direct marking campaigns from a Portuguese banking institution. The classification goal of this data is to predict if a client will subscribe to a term deposit.

The data consists of 45211 rows and 17 columns/variables (including predictor variable)

Input variables:

bank client data:

1 - age (numeric)

2 - job : type of job (categorical: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "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)

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)

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")

The data can be found here: https://archive.ics.uci.edu/dataset/222/bank+marketing

Note: The exact dataset that will be used is labeled “bank-full.csv”

Citation & Alternative Link:

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.

In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

Available at: [pdf] http://hdl.handle.net/1822/14838

[bib] http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt

Data Type

We have clean data with an assortment of numeric, categorical, and binary data. It is worth noting that the output variable, or predictor variable, “y” is binary.

Data Issues

We see unknown values.

We see outliers and skewness in balance, duration, campaign, and previous. We also see a potential error in pdays.

It is worth noting that despite being outliers, these values are crucial as the data is sensitive which means the outliers could provide useful information.

Data Solution/Approaches

We have three issues we need to solve for our data: unknown values, outliers, and skewness.

Unknown values: Will be deleted as they are missing values.

Outliers: As stated above, outliers in our case were deemed noteworthy. It is best not to remove them.

Skewness: If we calculate the natural logarithmic value of the data, we can normalize the distributions.

GitHub Link

https://github.com/ttn20/DataGlacierInternship/tree/Bank-Marketing-Final-Project