

**Data Science Intern at Data Glazier**

**Week 8: Report on Group Project**

**Topic: Bank Marketing (Campaign)**

**Group Name: Campaign Catalysts**

**Specialization:** Data Science

**Batch Code:** LISUM19

**Date:** 26th April 2023

**Submitted to:** Data Glacier

**Understanding the data:**

As part of the CRISP-DM methodology data understanding involves:

* collecting the data
* describing data
* exploring data
* verifying data quality

**Collecting the data:**

This process involves how the data is acquired and accessed showing resources and information about the data collected. This also means you need to have a checklist of the dataset you have acquired, the location, the methods to acquire the datasets, and record any problems encountered and any solutions to the problems for the other users or project members to be aware of.

There are generally four data sets found for the sake of this project.

The first two of these data sets are:

* bank-full.csv with all examples, ordered by date (from May 2008 to November 2010). The bank-full data set contains 45211 X 17 observations.
* bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv. The bank data set contains 4521 X 17 observations and is 10% of the examples (4521), randomly selected from bank-full.csv to test more computationally demanding machine learning algorithms (e.g. SVM).

**Data set location:**

This dataset is public available for research. The details are described in [Moro et al., 2011]. Where it used for Data Mining for Bank Direct Marketing: relating to a direct marketing campaigns of a Portuguese banking institution. [1]

The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed an Application of the CRISP-DM Methodology.

The data is available and can be located at:

* [pdf] http://hdl.handle.net/1822/14838
* [bib] <http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt>

**Data Attributes/Variables/Features:**

The main variables /attributes of the data are:

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 other last two datasets are:

* bank-additional-full with all examples, ordered by date (from May 2008 to November 2010). The bank-additional-full data set contains 41188 X 21 observations.
* bank-additional.csv with 10% of the examples (4119), randomly selected from bank-additional-full. The bank-additional data set contains 4119 X 21 observations and is 10% of the examples (4119), randomly selected from bank-full.csv to test more computationally demanding machine learning algorithms (e.g. SVM).

**Data set location:**

This dataset is publicly available for research. The details are described in [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. [2]

The data is available and can located at:

* [pdf] http://dx.doi.org/10.1016/j.dss.2014.03.001
* [bib] <http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt>

This dataset is based on "Bank Marketing" UCI, and is enriched by the addition of five new social and economic features/attributes collected from a national wide indicators from a 10M population country and published by the Banco de Portugal and publicly available at: <https://www.bportugal.pt/estatisticasweb> and was found to lead to a successful substantial improvement in the prediction process.

This dataset is almost identical to the one used in [Moro et al., 2014] (it does not include all attributes due to privacy concerns).

**The five additional social and economic context attributes / features added:**

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)

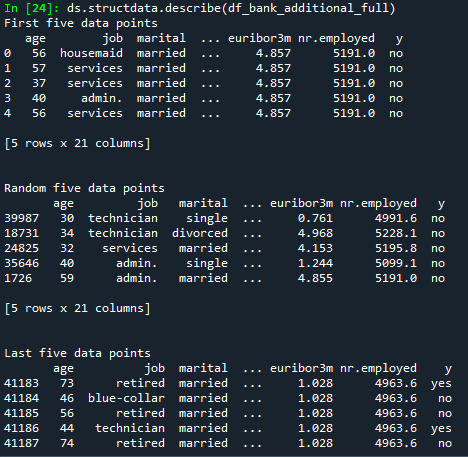
**Choice of data set to be used for the machine learning algorithm creation:**

The bank and the bank-additional datasets ware randomly selected from their corresponding datasets and were meant to be used solely for testing machine learning algorithms.

The bank-additional-full data set seems to be ideal for the purpose of this project in that it is more recent and has more variables which has been known to accurately lead to a successful prediction outcome for the project. During the algorithm creation, this data set chosen could be split into training and testing data ready for the machine learning algorithm or model creation, hence this seems to be the right data set to be used.

**Describing the data:**

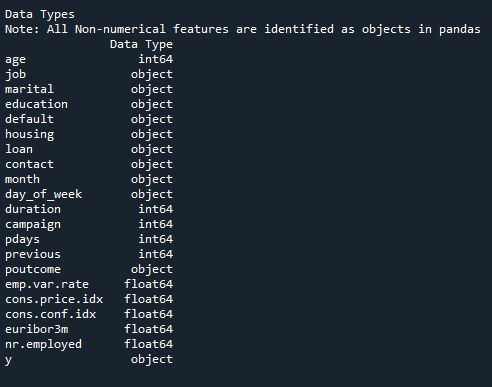
This process involves examining the properties of the data acquired, a description report regarding the format of the data, quantity of data and even the records and fields in each table or datasets.



The above shows the first 5, random 5, and the last 5 of data points of the entire data set.



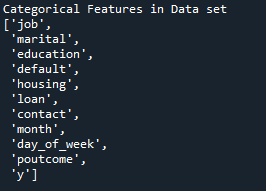
Shows the shape and the size of the data.



The above shows the different data types in the entire data set.

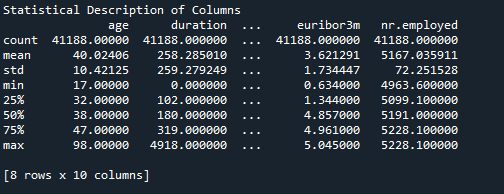


The above shows the numerical features in the data set.



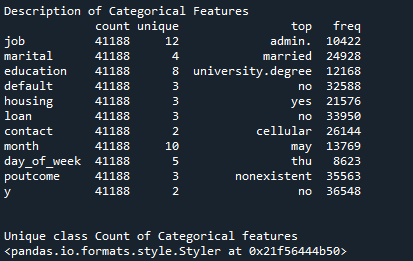
The above shows the categorical features in the dataset.

Below shows a brief statistical description of some of the numerical features.



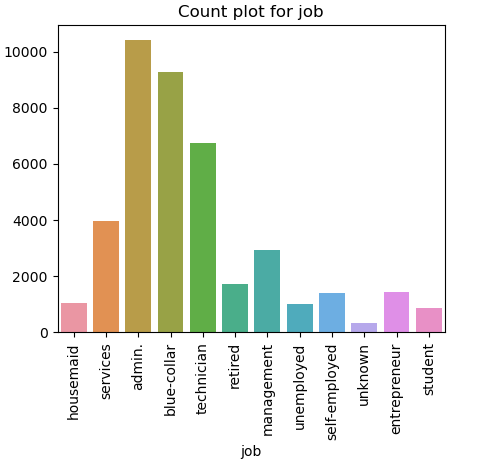
The research was carried out on people with the age range from 17 years to 98 years with the mean age of 40 years.

Below shows the description of the categorical features with their corresponding unique counts

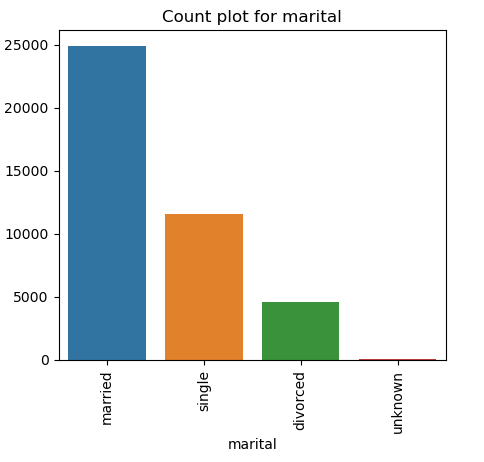


**Exploring the Data:**

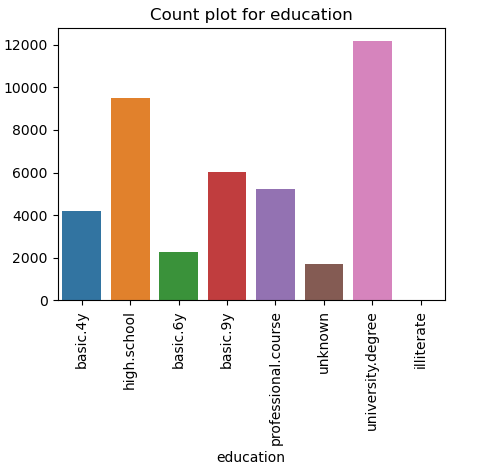
This process involves querying the data using visualization, providing a summary report. This stage, also leads to finding initial hypothesis and their impact on the project.



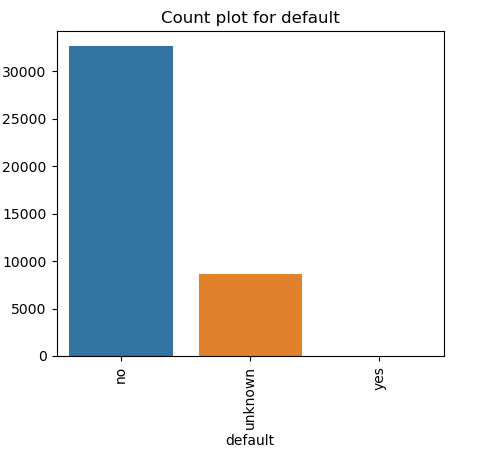
People with administrative job type were more involved in this project, followed by people with blue-collar jobs and technicians. The lease number of people involved in this survey for the project were students, housemaid and the unemployed.



Married people were more involved in this project follwed by single and the divorced which recorded the least number of peole for this bank marketing campaign survey.



More people with a university degree took part in this bank marketing survey, followed by high school students. illiterate people recorded the least number considered in this bank marketing campaign survey.



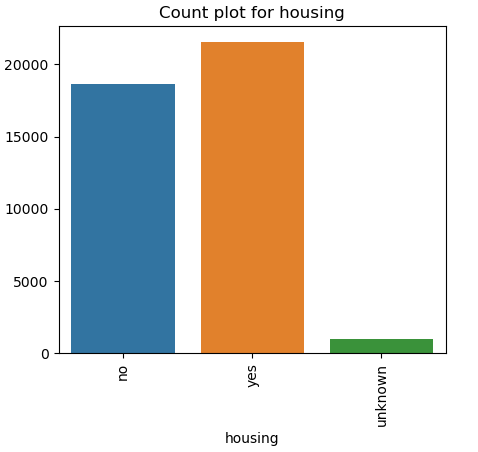
The number of people that defaulted on credit were much more than the people that did not creating a class imbalance which typically is always a fundamental issue when it comes to classification for machine learning process.

Class Imbalance Data:

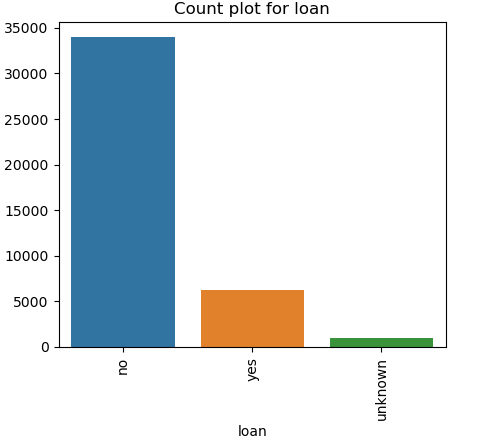
The below shows the Class distribution and could be clearly seen as a skewed distribution

which has become one of the data mining challenges when it comes to a pure binary

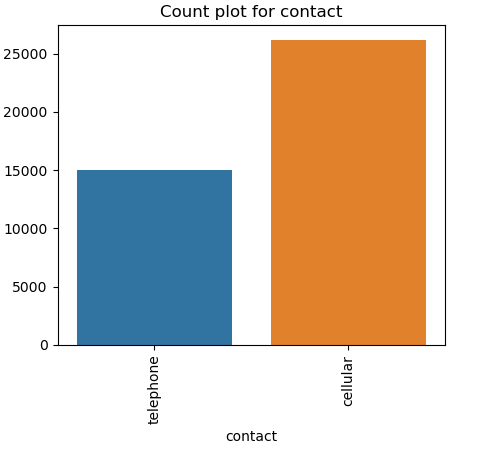
classification scenario



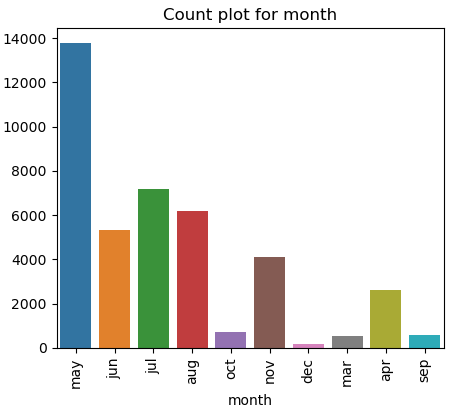
There were slightly more people with housing loan than people without.



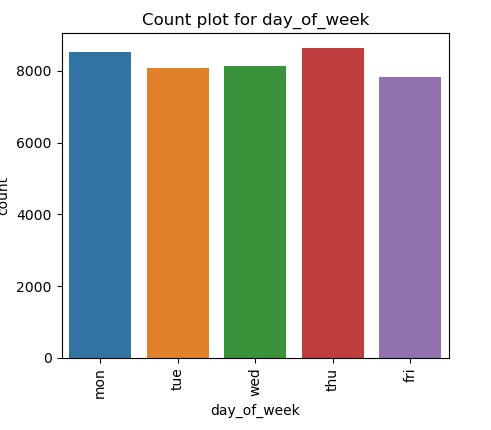
The number of people who had no loan were about ten times the people who had a loan, meaning more people without a loan took part in this survey than people with a loan.



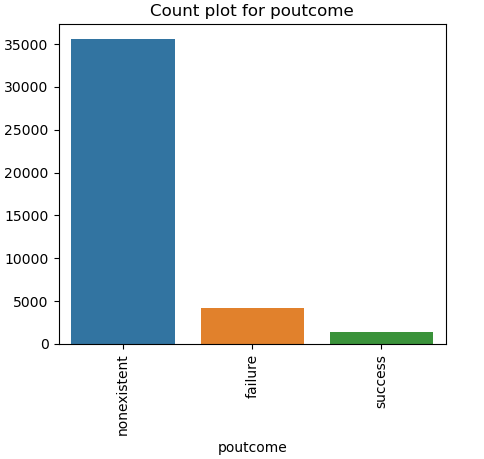
There was more contact made by cellular than telephone where for every three people contacted by telephone five more were contacted by cellular.



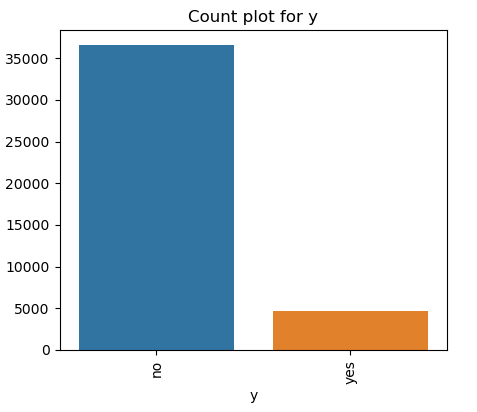
The month of May saw more last contacts been made than any month of the year according to the survey, followed by June and August with December recording the least number of contacts been made.



There were slightly equal and active number of contacts made for the days of the week.



There were more failures than success for the outcome of the previous campaign recorded with a high number of non-existent records with no activity.



The number of clients that subscribed for a term deposit were ten times more than clients that did not subscribe for a term deposit creating yet again an imbalance class problem.

Class imbalance is one of the challenging and interesting problem in machine learning study.

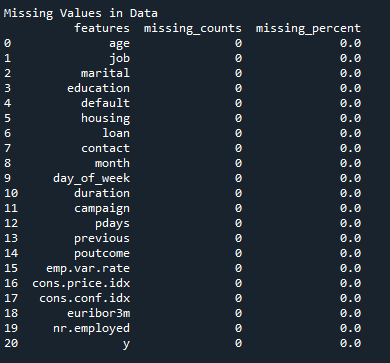
The majority of machine learning algorithms designed previously usually assume that the

training sets are well-balanced, and implicitly assume that all misclassification errors cost

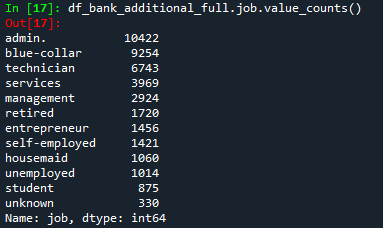
equally. But data in real-world is usually imbalanced. [3]

**Verify Data Quality:**

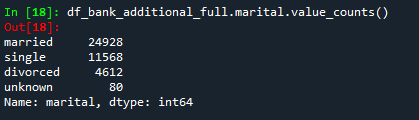
This process examines if the data is complete. If the data has errors or are there missing values and if there is, what is the percentage of the missing values against the overall obtained data.



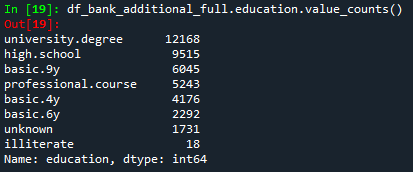
There are practically no missing values present in the data set, however there more unknown records present in the data set and this could be seen in the below for the different variables in the data set.



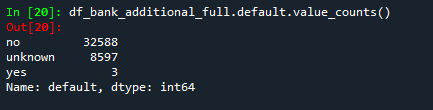
There are 330 unknown records for the job variable.



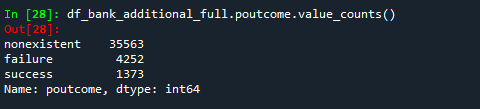
There are 80 unknown records in the marital variable.



There are 1731 unknown records for the education variable.



There are 8597 unknown records for the default variable.



There are 35563 non-existent unknown records for the poutcome variable.



However, there are 12 duplicates of records present in the data set. This will be address and dealt with in the data preparation stage, the next stage of the project process.

These missing values, unknown values, duplicates and any anomalies in the data set will be addressed and dealt with in the next stage of the project process, where the sole attention will be focused on data cleaning and transformation

**References:**

[1]: [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.

[2]: [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

[3]: Data Mining on Imbalanced Data Sets from the 2008 International Conference on

Advanced Computer Theory and Engineering journal