Bank Marketing (Campaign)

By Priyadarshani Kamble

LISUM11: 30

Problem Statement:

1 Context

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).

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

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

2 Criteria for success

Understanding the Data: Perform data cleaning/wrangling if needed. e.g., duplicate, and missing data handling. Observe relationships between features by performing exploratory analysis also share other insights of the data.
Modelling approach: This will be supervised machine learning problem which will involve using classifications models to predict to predict if the client will subscribe (yes/no) a term deposit (variable y)

3 Scope of solution space

Predict the **subscribe** (yes/no) attribute which is indicator to show if the client will subscribe a term deposit.

4 Constraints within solution space

- The dataset contains missing values for several columns.
- The data volume for each Genetic Disorder class is not similar.
- The dataset may not have all the parameters that can help predict the genetic disorder accurately.

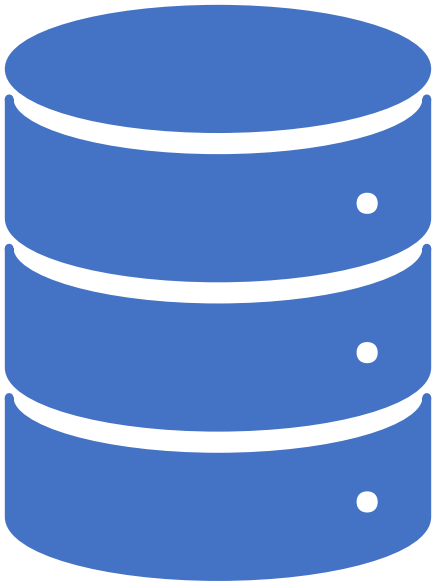
5 Stakeholders to provide key insight

Data science team

6 Key data sources

The data is available in csv file. The Key attributes include age, job, marital, education, default, balance, housing loan, contact, day, month, duration, campaign, pdays, previous, poutcome, y

Dataset Information



Data storage location:

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

Csv file : bank-additional-full.csv

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	5699kb

- There are 10 numeric columns and 11 Categorical Columns
- No missing data found.
- 12 duplicate rows found.

Attribute Information: 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 - balance : average yearly balance, in euros (numeric)

7 - housing: has housing loan? (categorical: 'no','yes','unknown')

8 - loan: has personal loan? (categorical: 'no','yes','unknown')related with the last contact of the current campaign:

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

10 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

11 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

12 - 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:

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; 999 means client was not previously contacted)

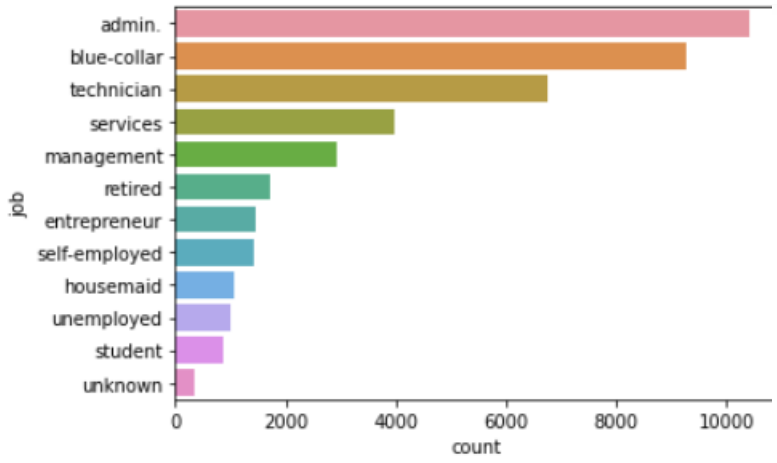
15 - previous: number of contacts performed before this campaign and for this client (numeric)

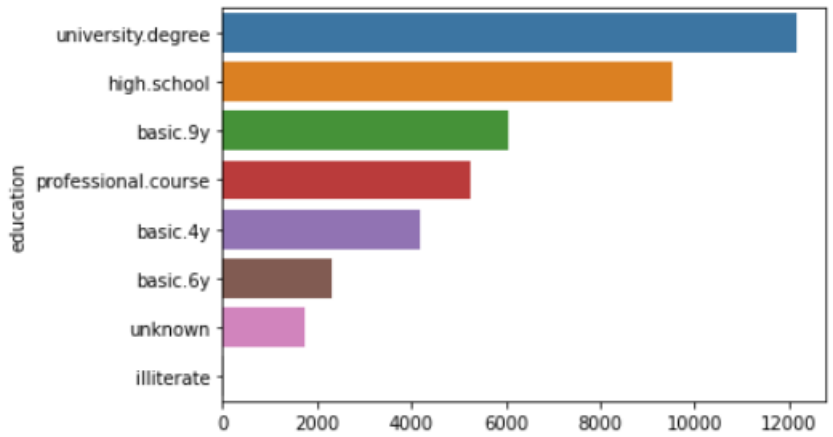
16 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

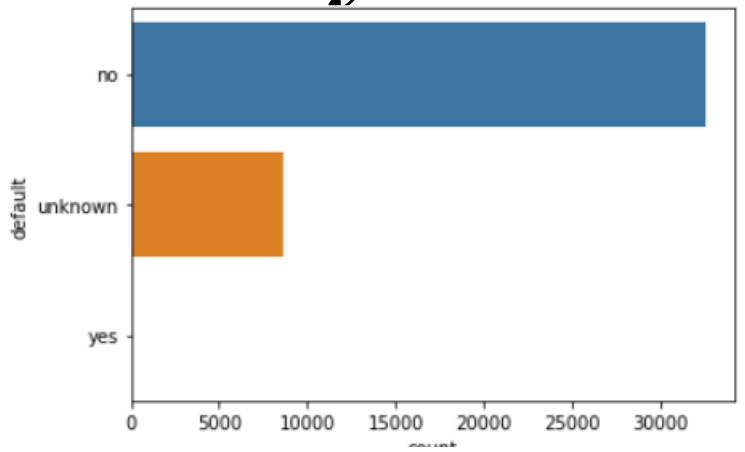
Output variable (desired target):

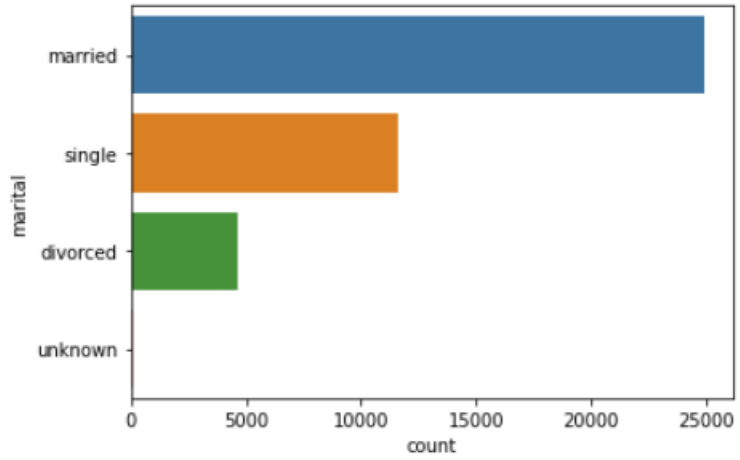
17 - y - has the client subscribed a term deposit? (binary: 'yes','no')

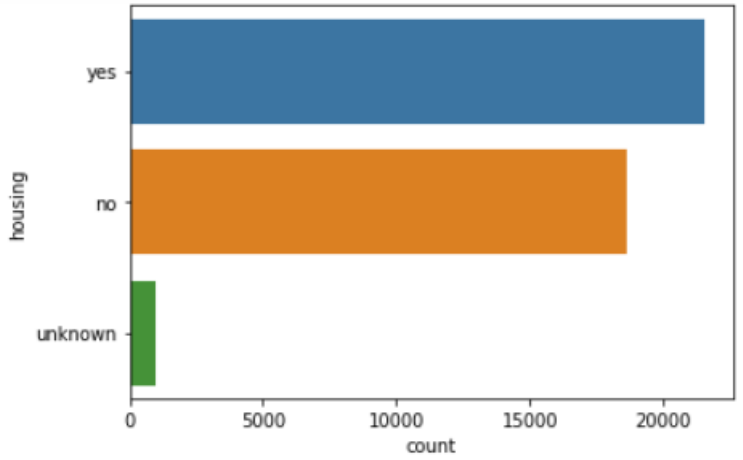
Exploratory Data Analysis – Univariate Analysis

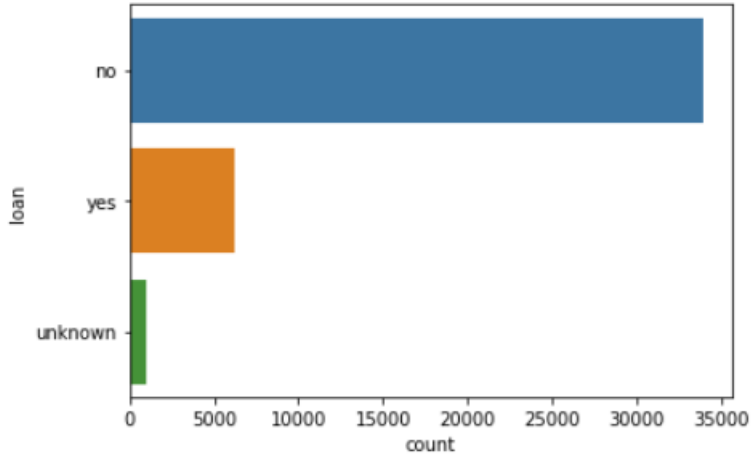






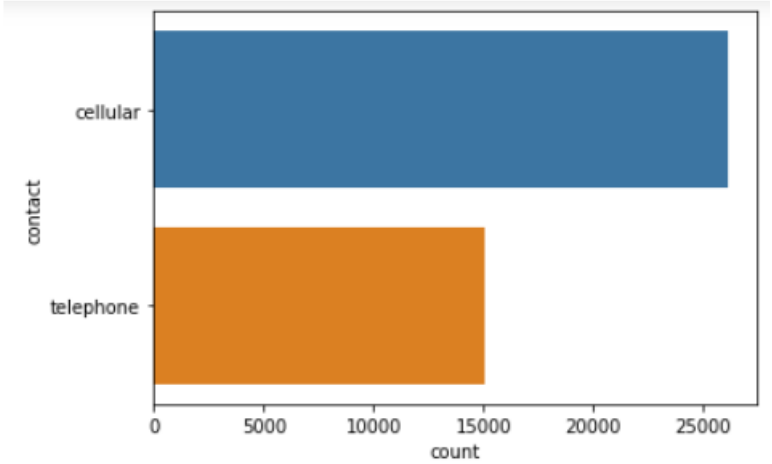


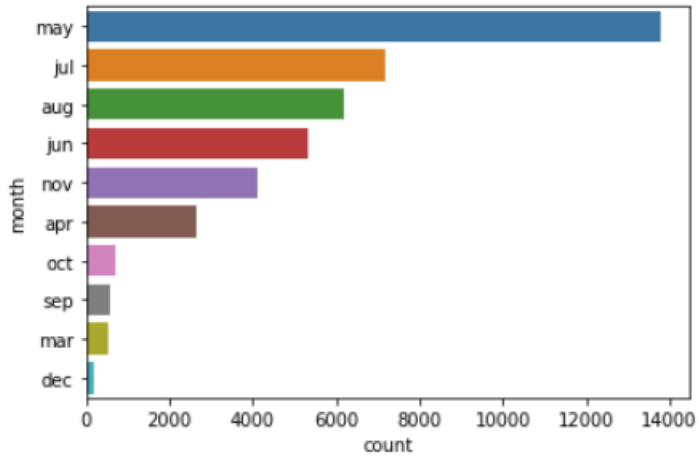


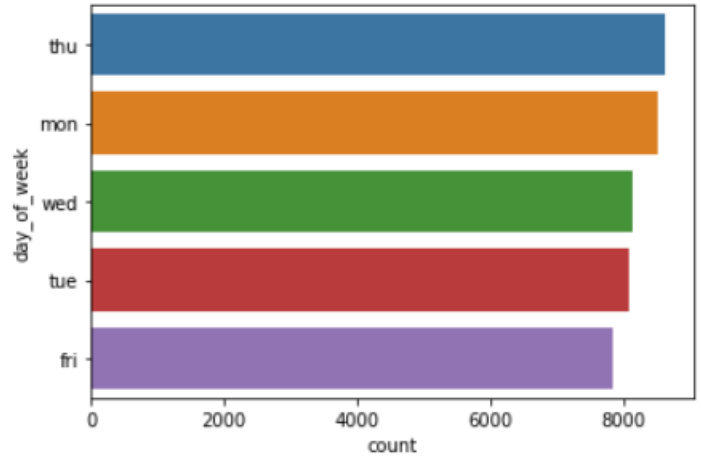


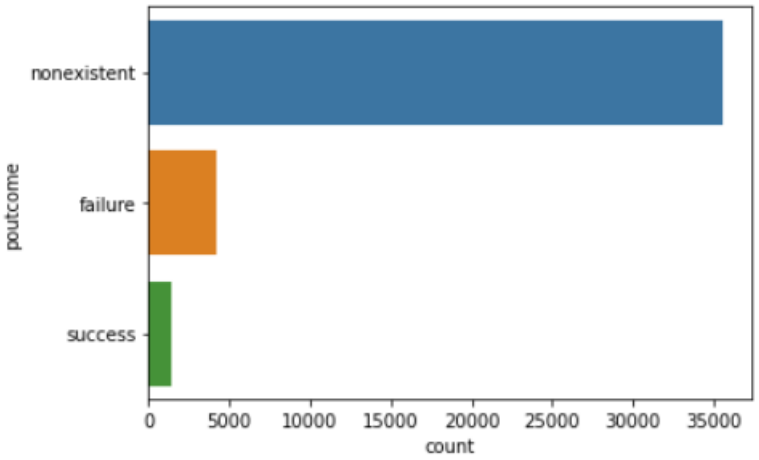
- Most of the clients are working as admin 25% and blue collar 22% job category.
- 60% of the clients are married.
- Most of the clients 29% hold University degree.
- The no of clients who defaulted on a credit, are very less. It shows 80% data for 'No'. % of yes is almost 0. so this feature doesn't seem to be very for prediction purposes and can be dropped from the dataset.
- Housing Shows almost equal % of yes and no.
- Most of the clients do not have personal loan.

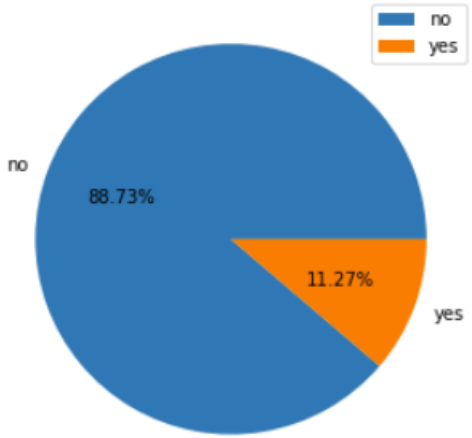
Exploratory Data Analysis – Categorical Features





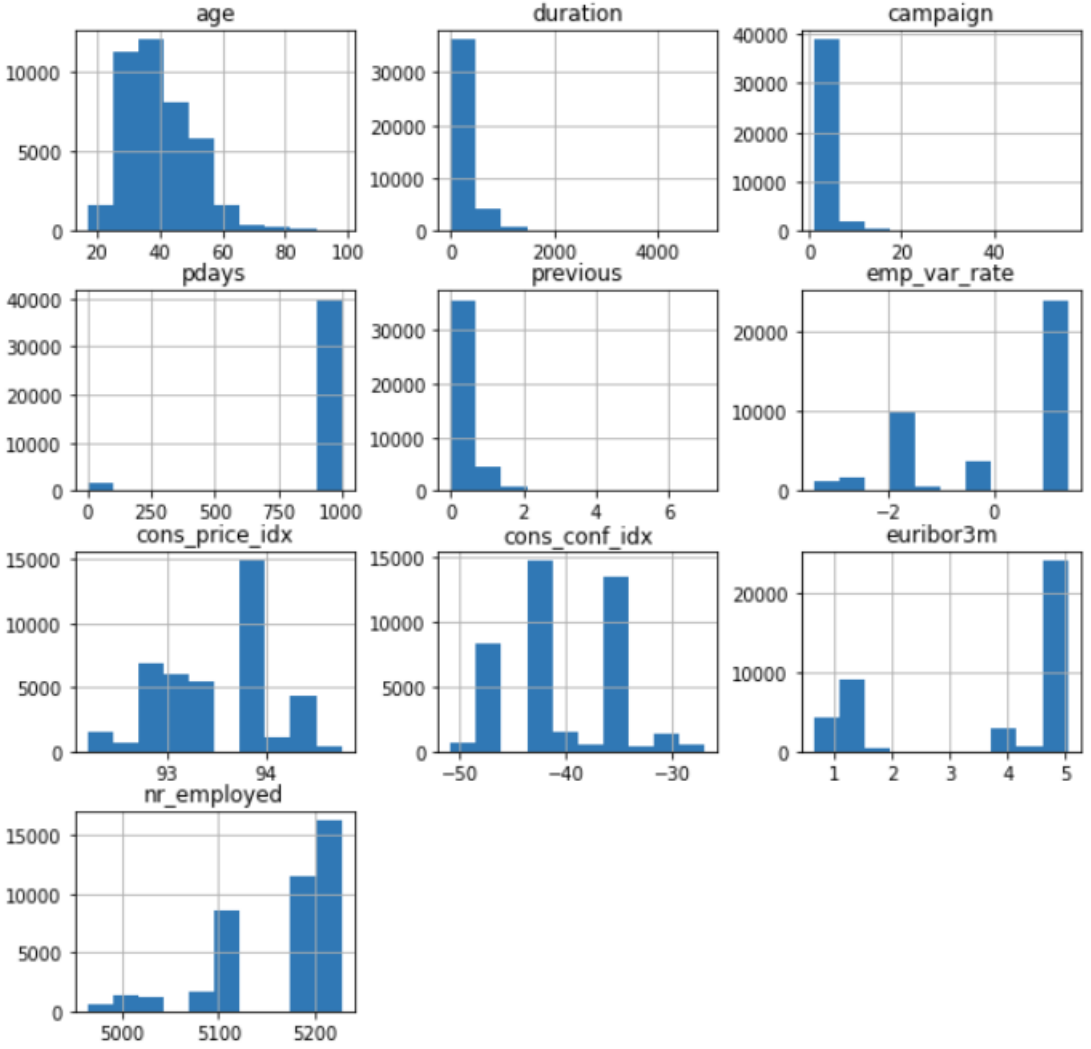






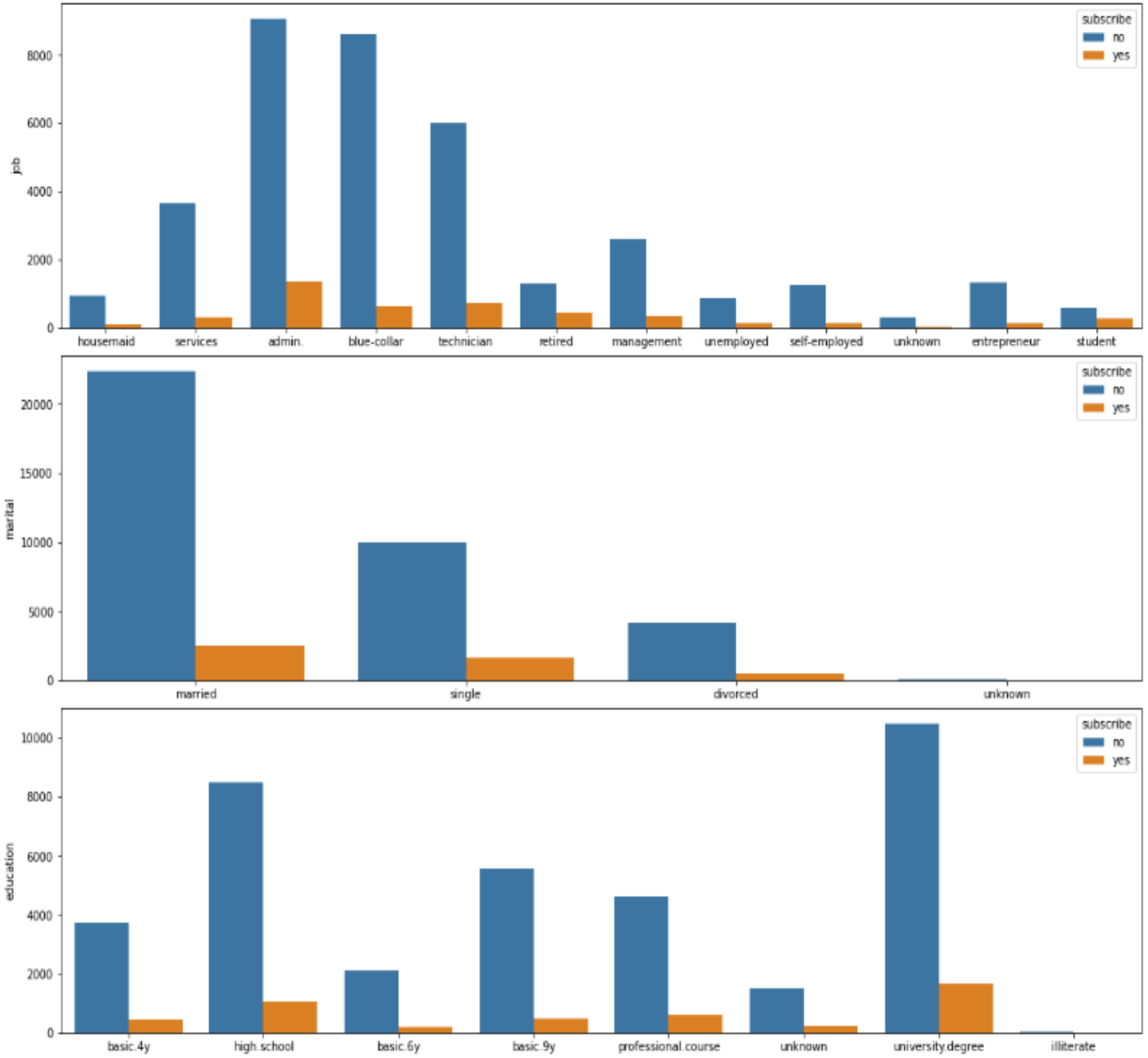
-More than 63% of all clients were contacted through cellular phone.
-Most of the clients were contacted in the month of May 33%.
-Here we see equal distribution of the data in the graph and the % amongst the days. So, there is no significant day which shows more activity than others.
-More than 86% of clients were never covered by previous marketing campaigns.
-We see there is imbalance in data. only 11.70% clients have subscribed to a term deposit.

Exploratory Data Analysis – Numerical Features

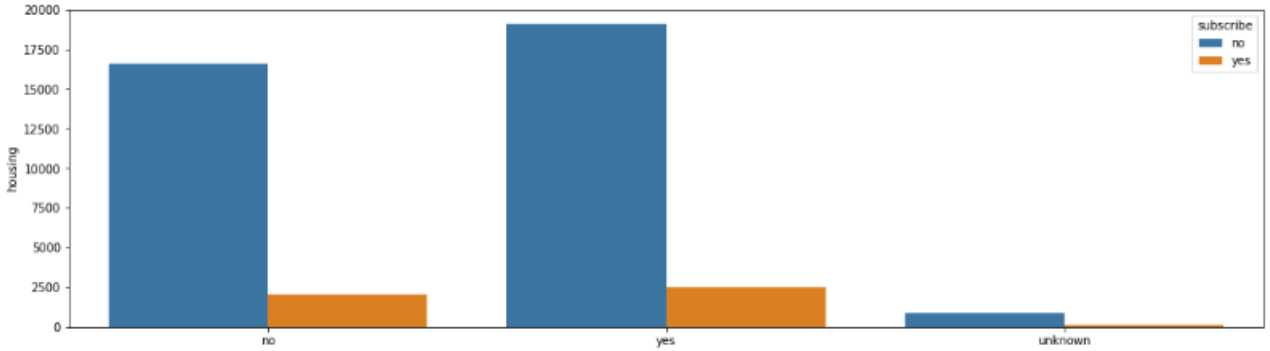


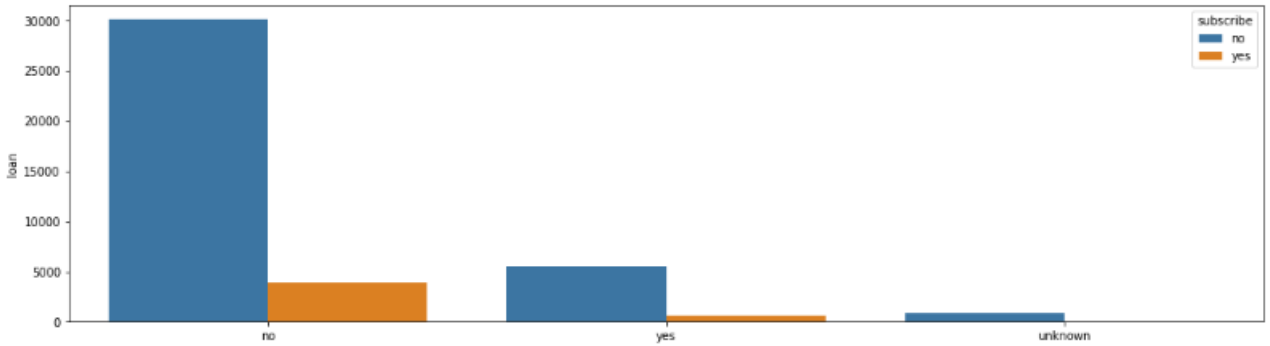
- The graph of age shows normal distribution.
- The graph of pdays, previous, duration, campaign shows skewness.
pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted).
- pdays columns shows above 96% values are having 999 value. So, this column should be removed.
- The variable "duration" will need to be dropped before we start building a predictive model because it highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed.
- Other graphs shows sevral spikes in the data.

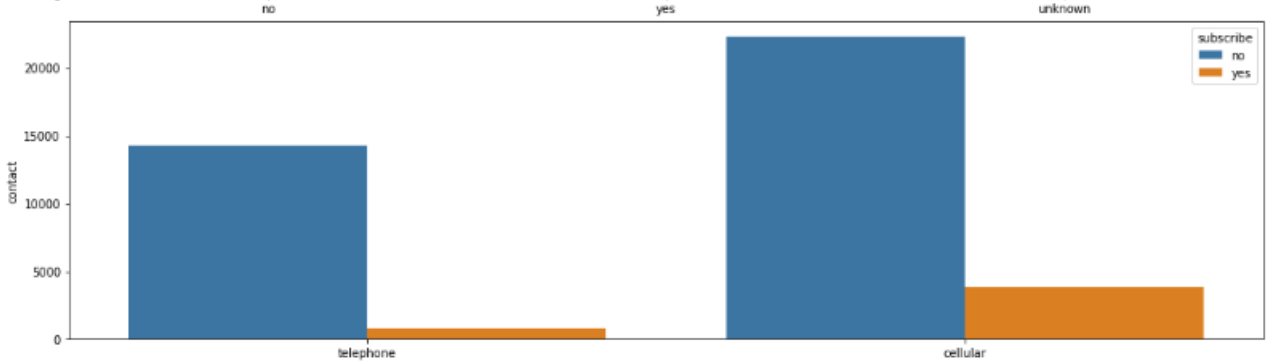
Exploratory Data Analysis – Bivariate Analysis

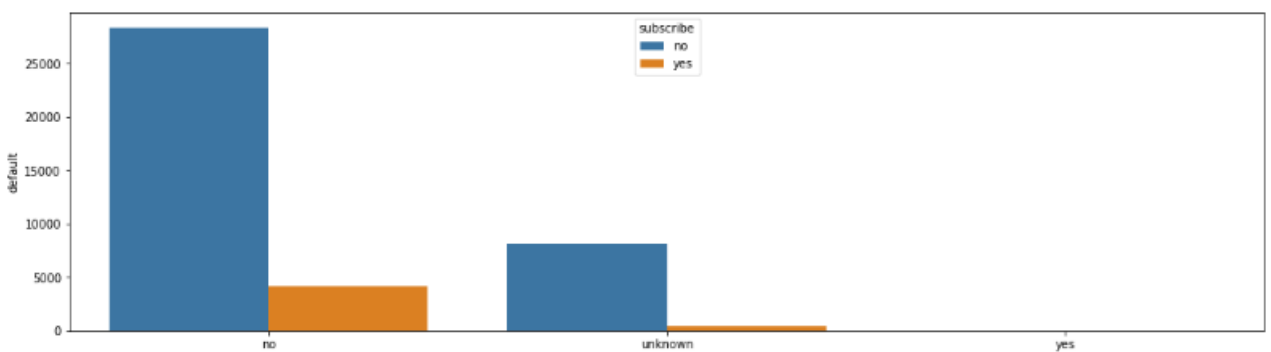


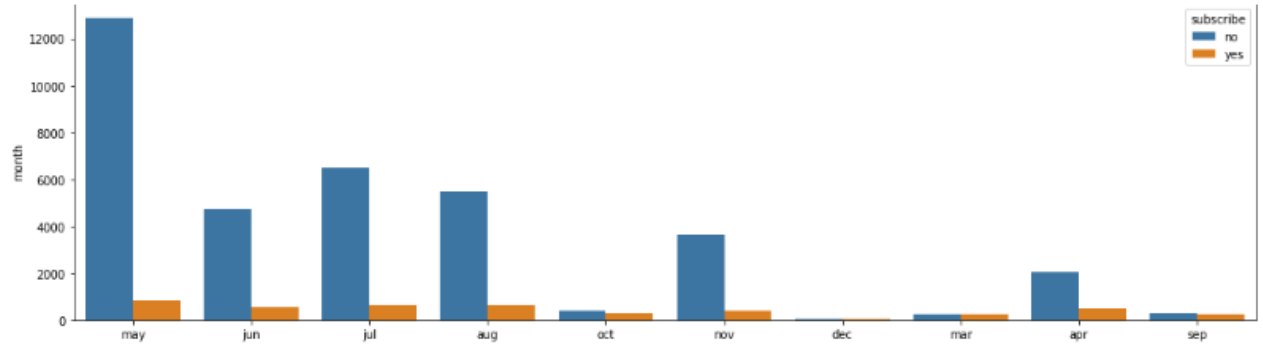
- The % of clients showing interest in 'term deposit' is more when they are married, holds university degree and no defaulted to credit.
- The retired people seem to have higher % of 'Yes' for term deposit than other job category clients.
- The clients having housing, loan have higher % or saying no to term deposit.
- The months may, jun, juy, aug shows more clients responding to term deposit
- If poutcome = success then the % of 'yes' to term deposit is high
- % of interest in 'Deposit' is more when clients are contacted via cellular mode.
- Clients having longer durations shows more interest in term deposit

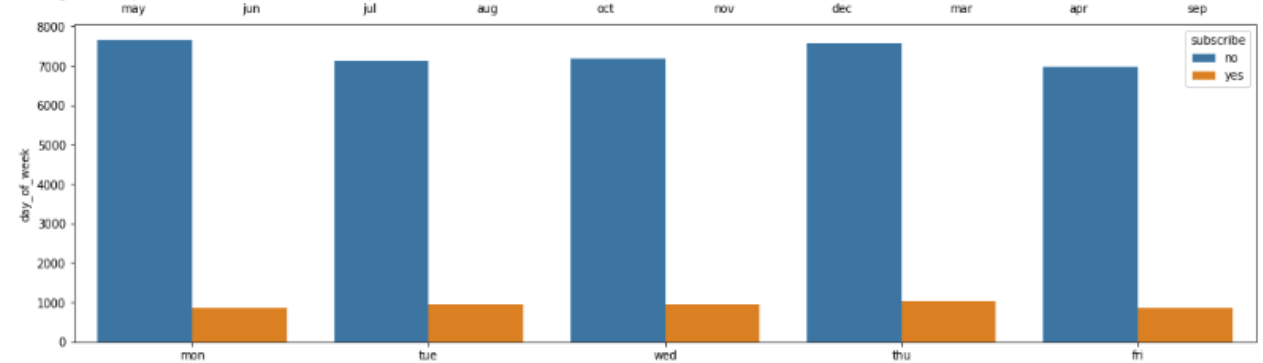

housing
subscribe
no
yes
no
yes
unknown

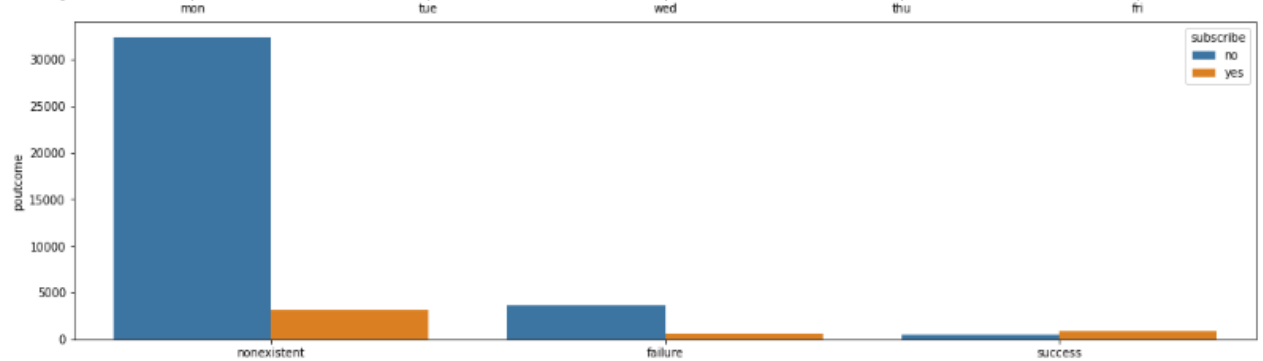

loan
subscribe
no
yes
no
yes
unknown

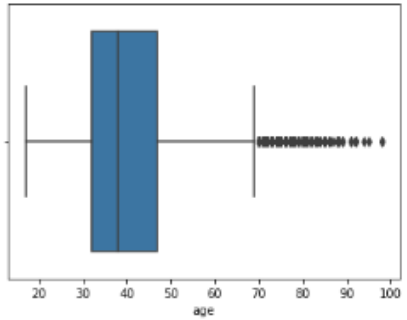

contact
subscribe
no
yes
telephone
cellular

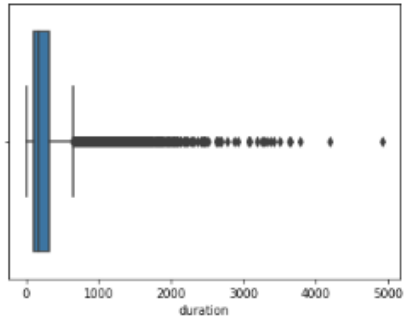

default
subscribe
no
yes
no
unknown
yes

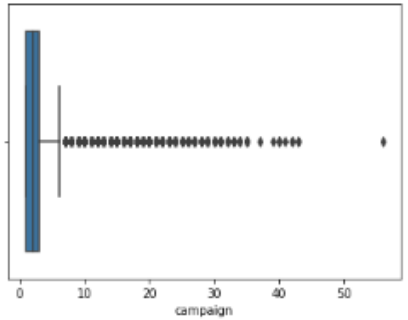

month
subscribe
no
yes
may
jun
jul
aug
oct
nov
dec
mar
apr
sep

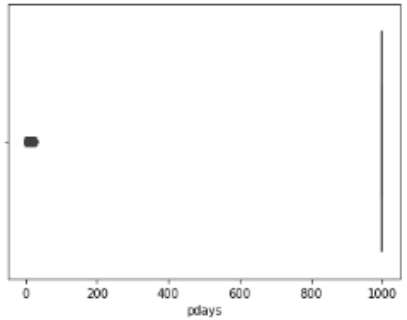

day_of_week
subscribe
no
yes
mon
tue
wed
thu
fri

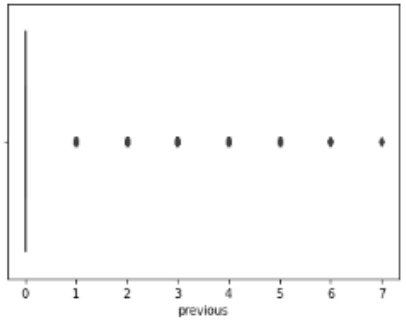

poutcome
subscribe
no
yes
nonexistent
failure
success

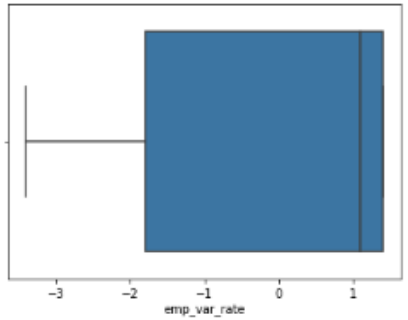


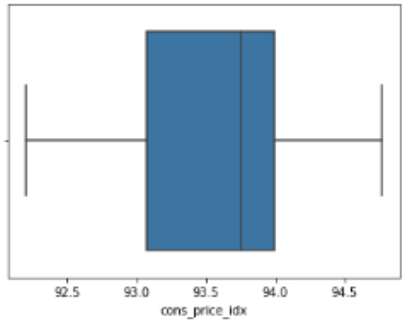


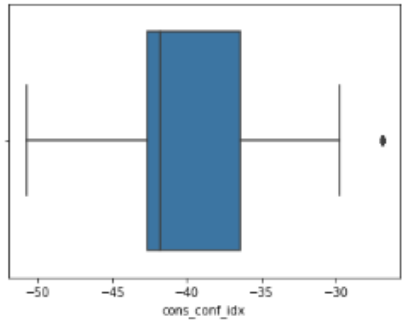


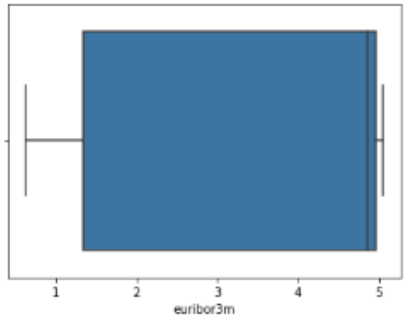


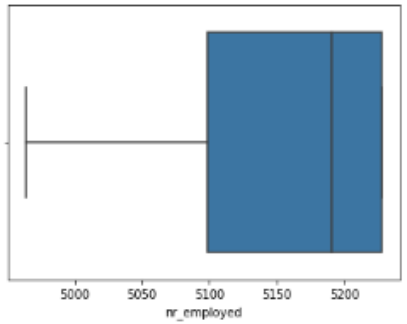






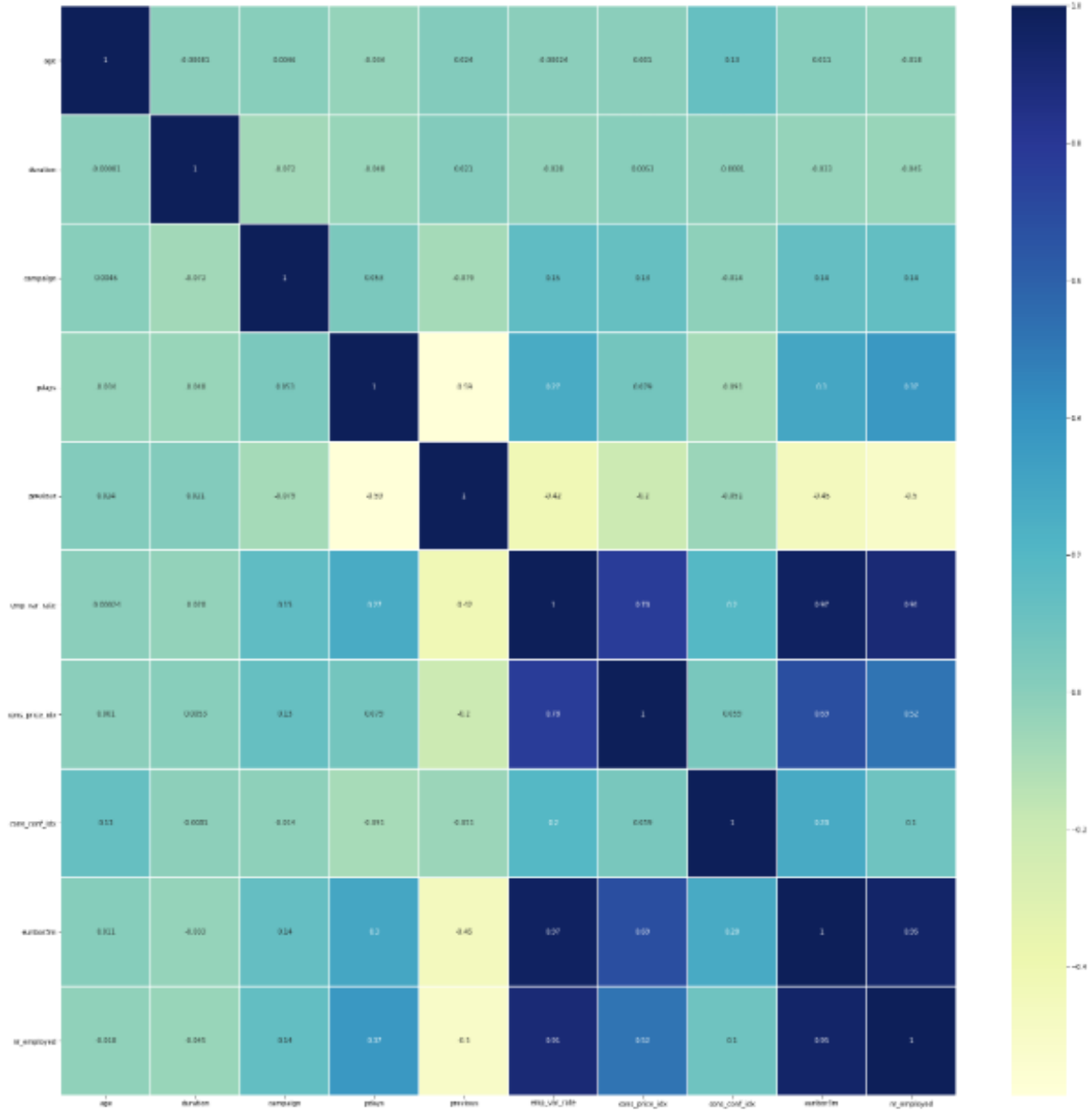






- The features like age, pdays, duration, campaign, duration has some outliers

Exploratory Data Analysis – Multivariate Analysis



As we see in above graph , euribor3m is highly corelated with 'emp_var_rate' and nr_employed. So, I dropped 'emp_var_rate' and nr_employed as they will only add redundancy and overfitting.

Model Selection

→ As this is a classification problem, I think we should use logistic regression, random forest classifier, gradient boosting classifier and XGB classifier. We will see the performance of these models and decide if it needs further tunning. Looking at the target variable ‘Subscribe” data, it is imbalanced data. So, we should surely try some data imbalance techniques like class_weight=balanced, over/under sampling.