Exploratory Data Analysis

BANK MARKETING ANALYSIS USING MACHINE LEARNING

14/08/2021

Agenda

Executive Summary

Problem Statement

Approach

EDA

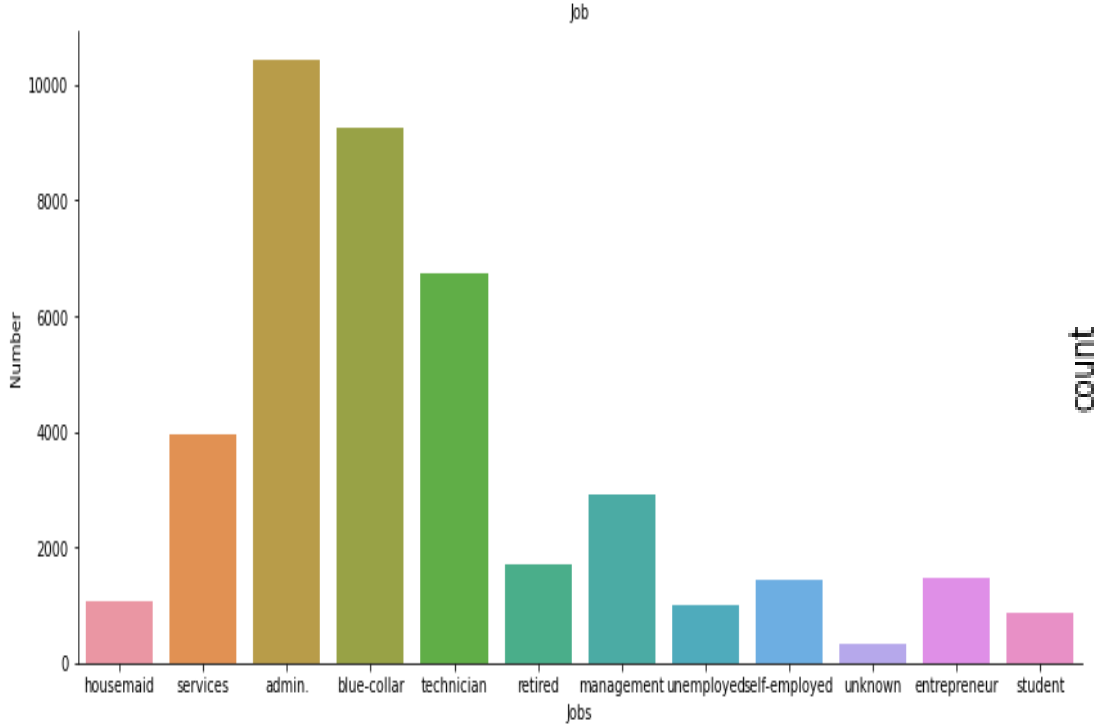
EDA Summary

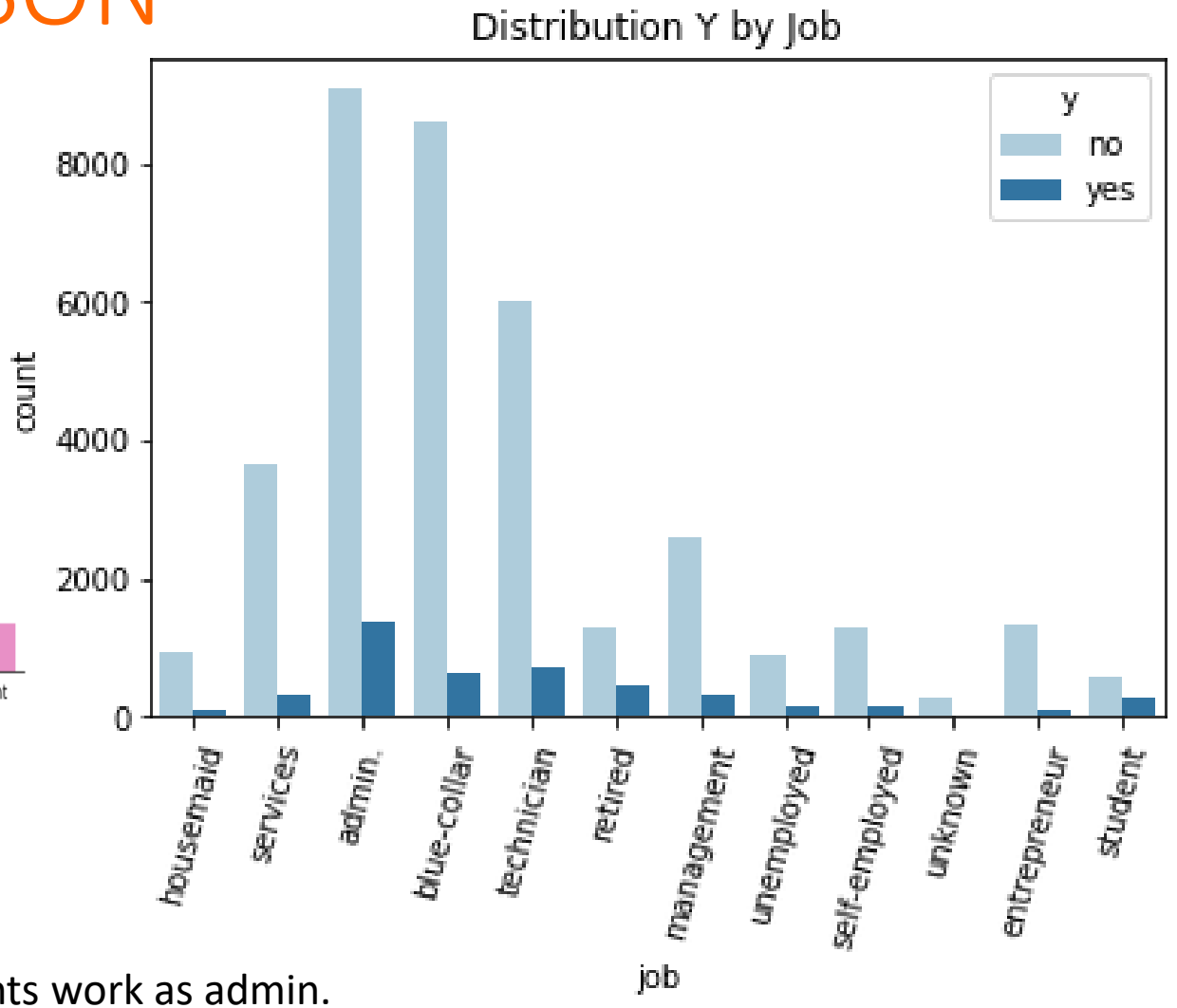
Recommendations

PROBLEM DESCRIPTION

- We are given data related to direct marketing campaigns i.e. phone calls of a bank in Portugal.
- The classification goal is to predict whether a client will subscribe or not(yes/no) to a term deposit (variable y).

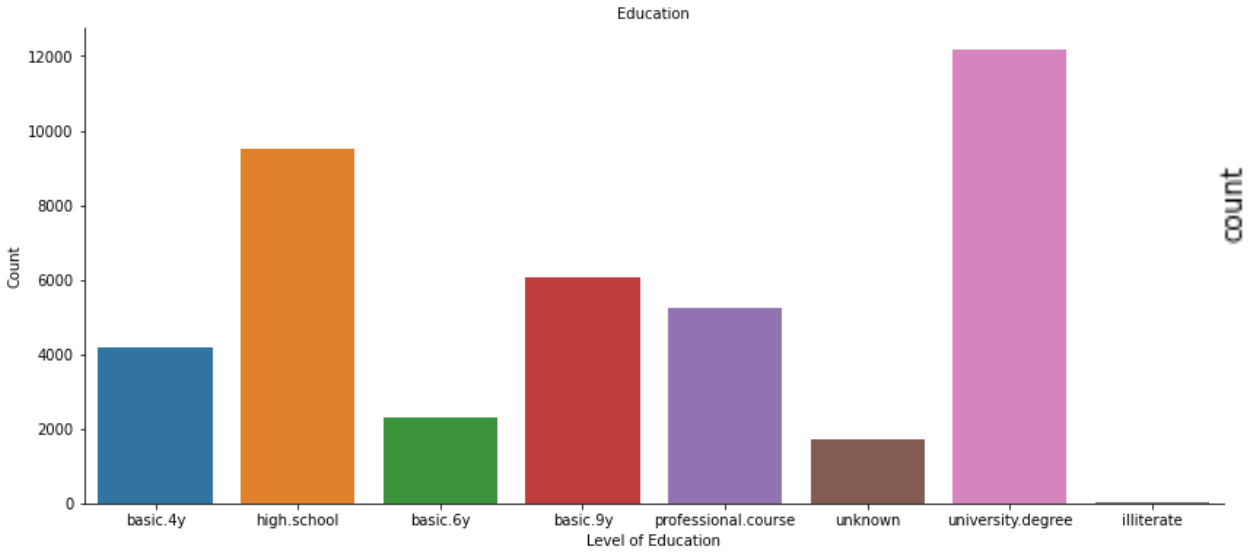
FEATURE COMPARISON

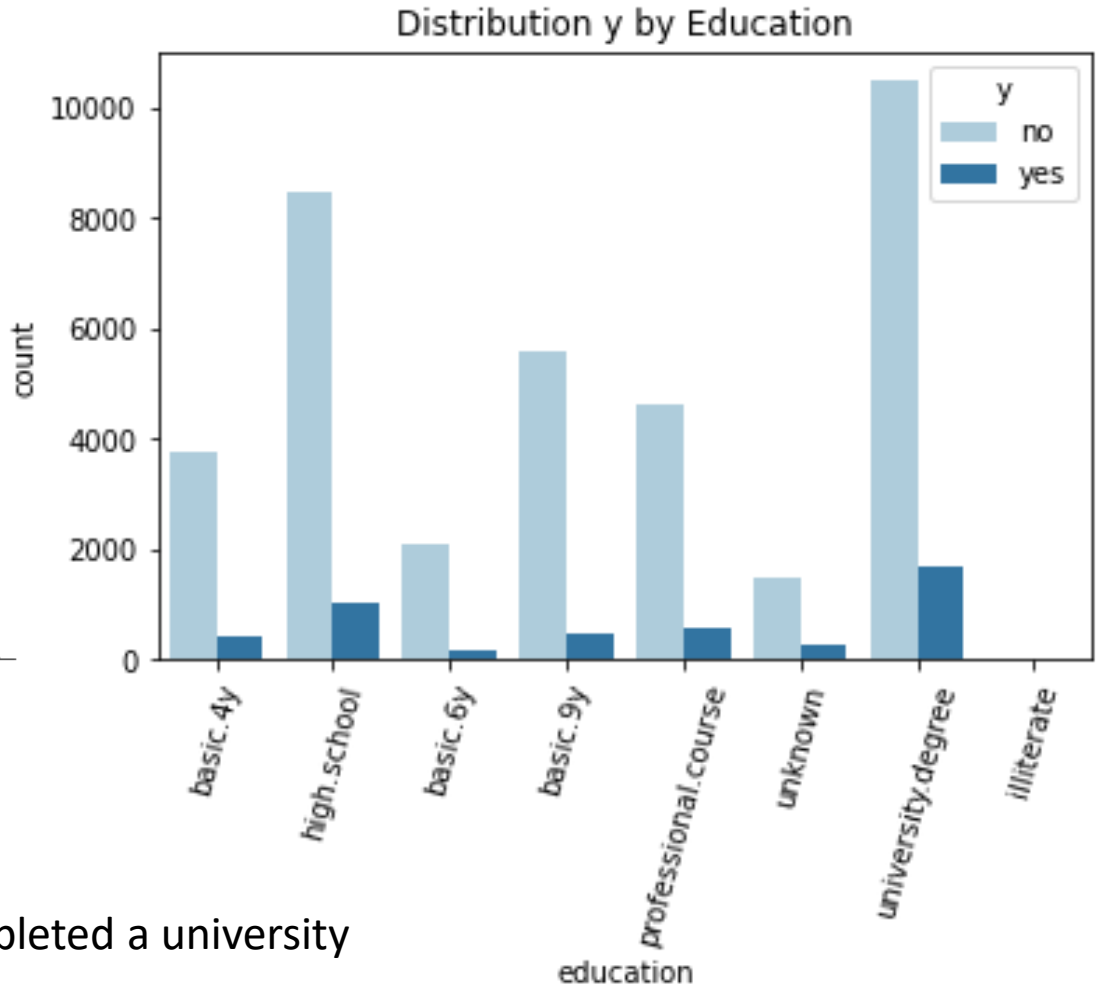




From the above we can see that majority of the clients work as admin.
Equally those who work in admin subscribed more to the term deposit.

EDUCATION

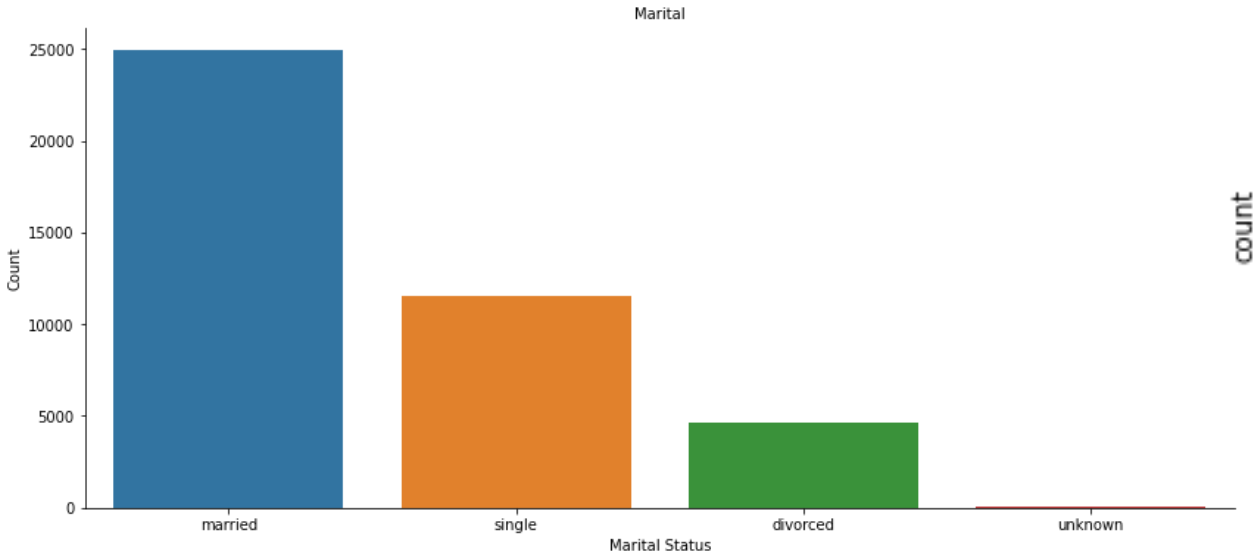


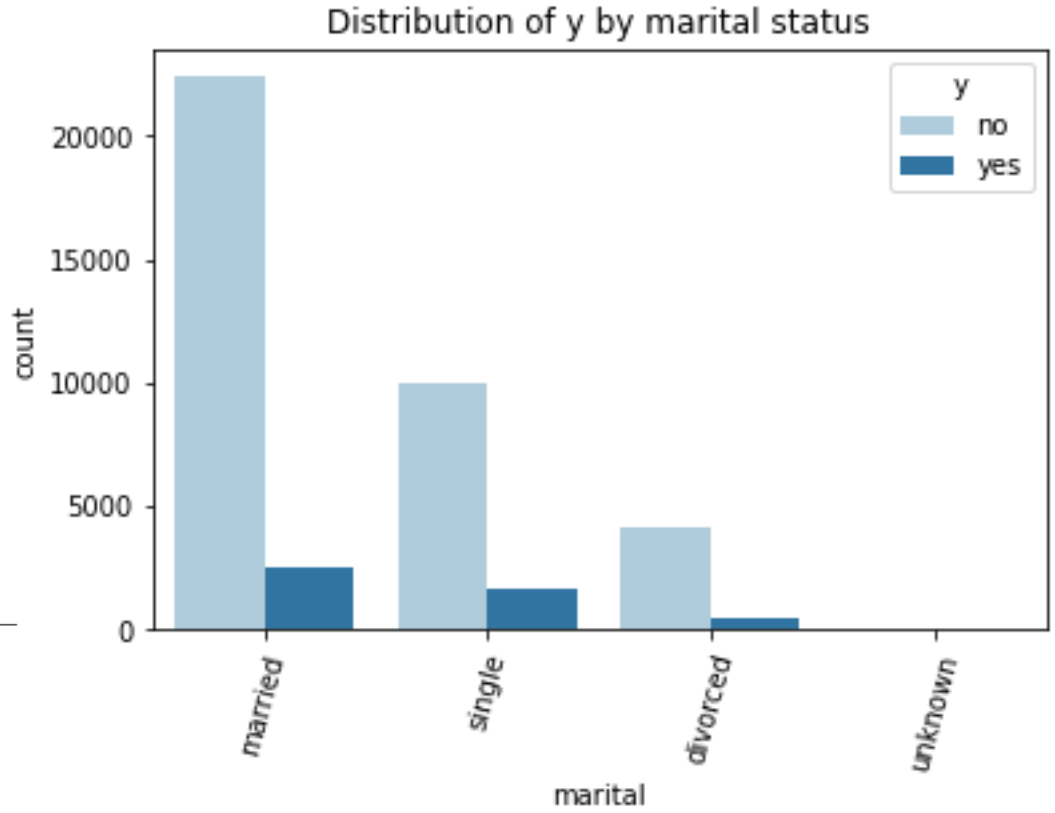


From the plot most of clients who were contacted completed a university degree.

Most of those who subscribed completed a university degree .

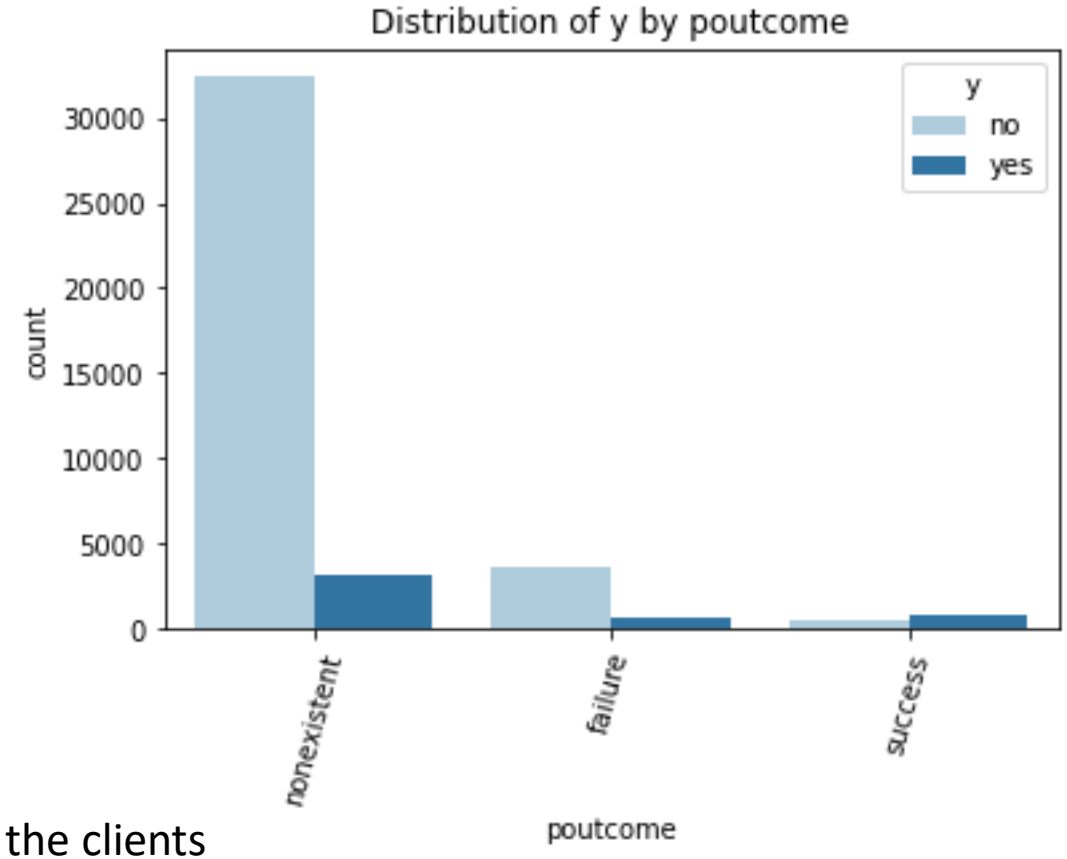
MARITAL





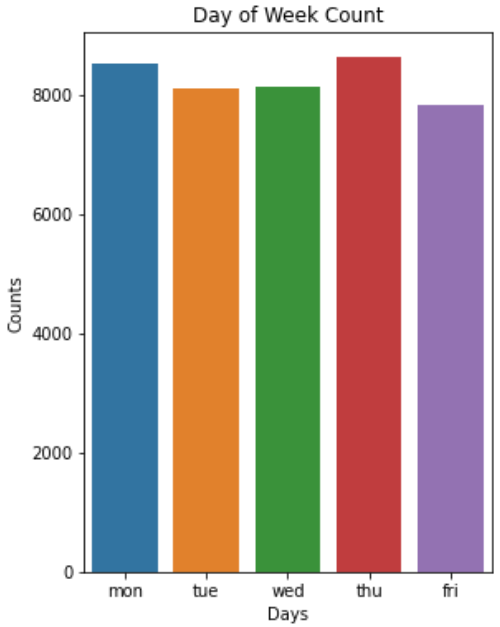
Majority of the clients were married and they subscribed slightly more than those who were single.

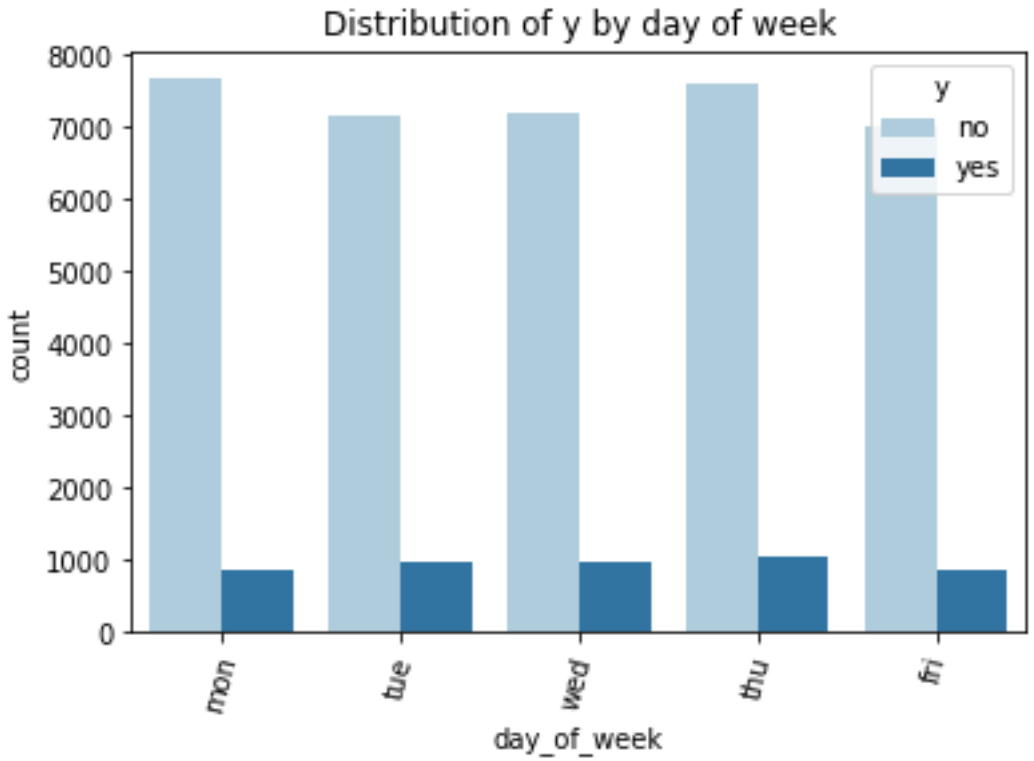
POUTCOME



Considering the outcome of the previous campaign, majority of the clients contacted were new clients and they subscribed more compared to those who had been contacted before.

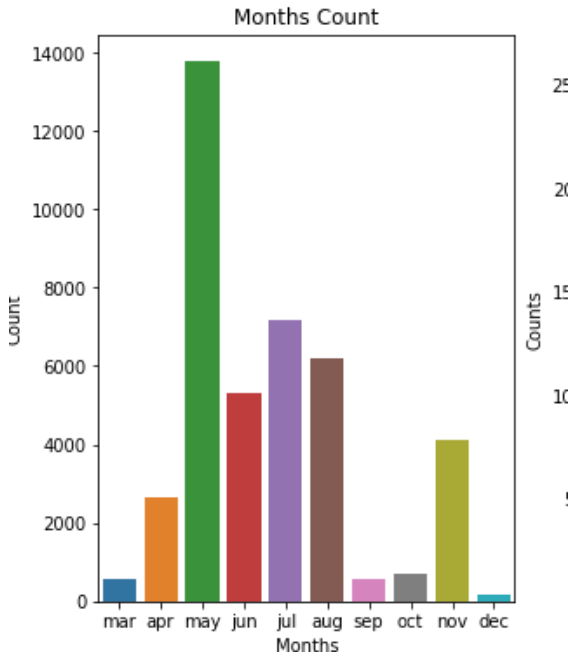
DAY OF WEEK

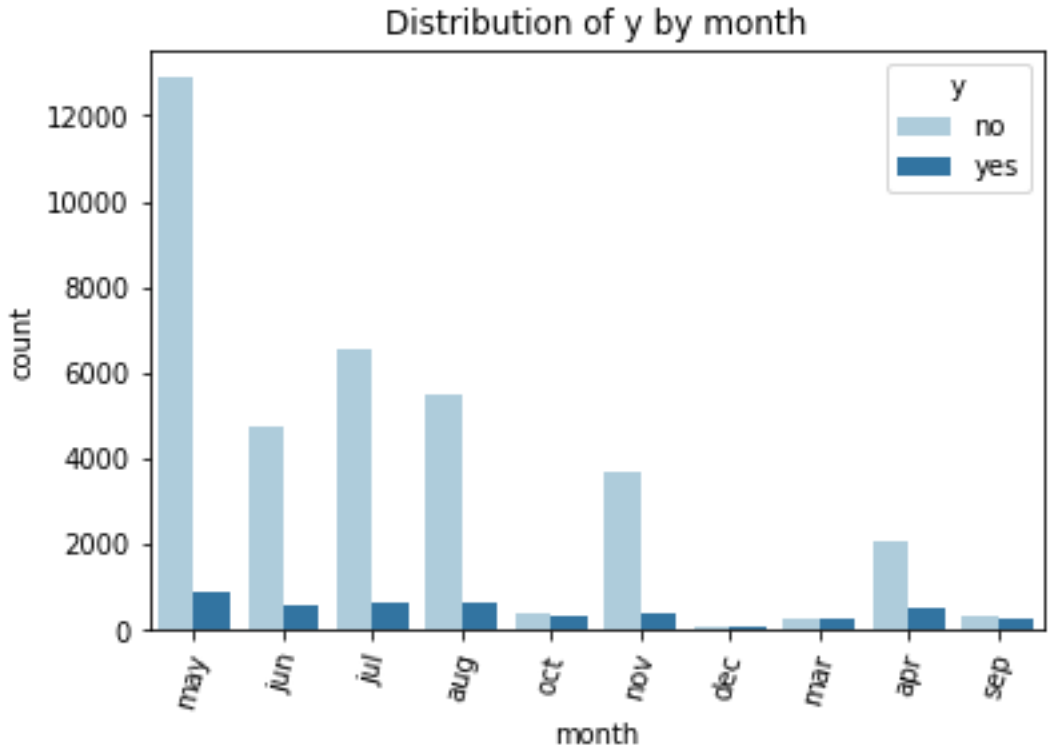




Clients were contacted on Monday, Tuesday, Wednesday, Thursday and Friday with the least number of clients being contacted on Friday.
Almost an equal number of clients contacted subscribed .

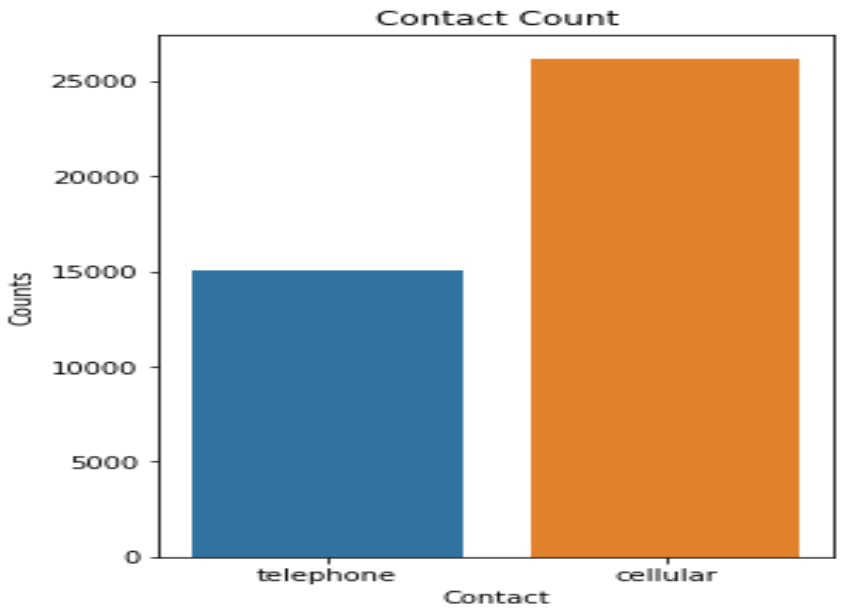
MONTH

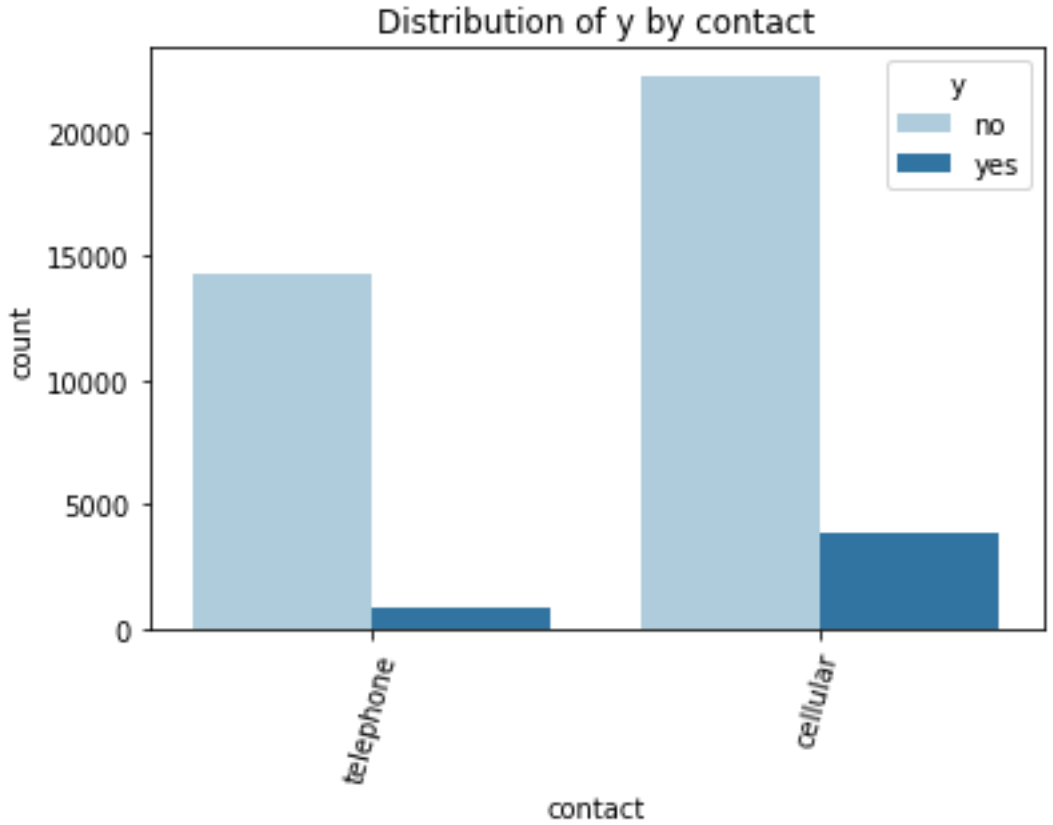




Clients were contacted most in the month of May and more clients equally subscribed in May.

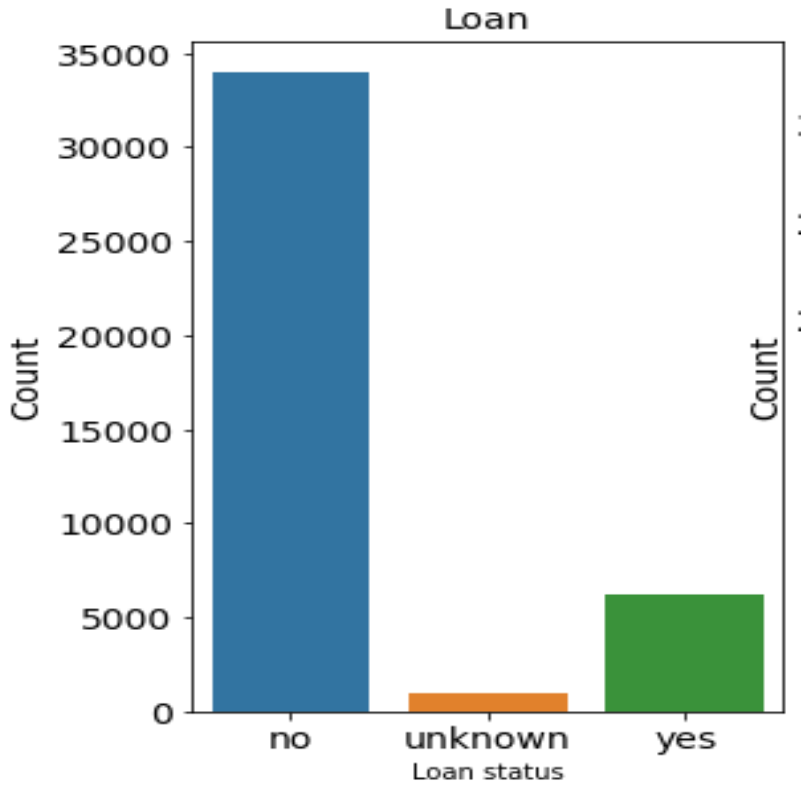
CONTACT

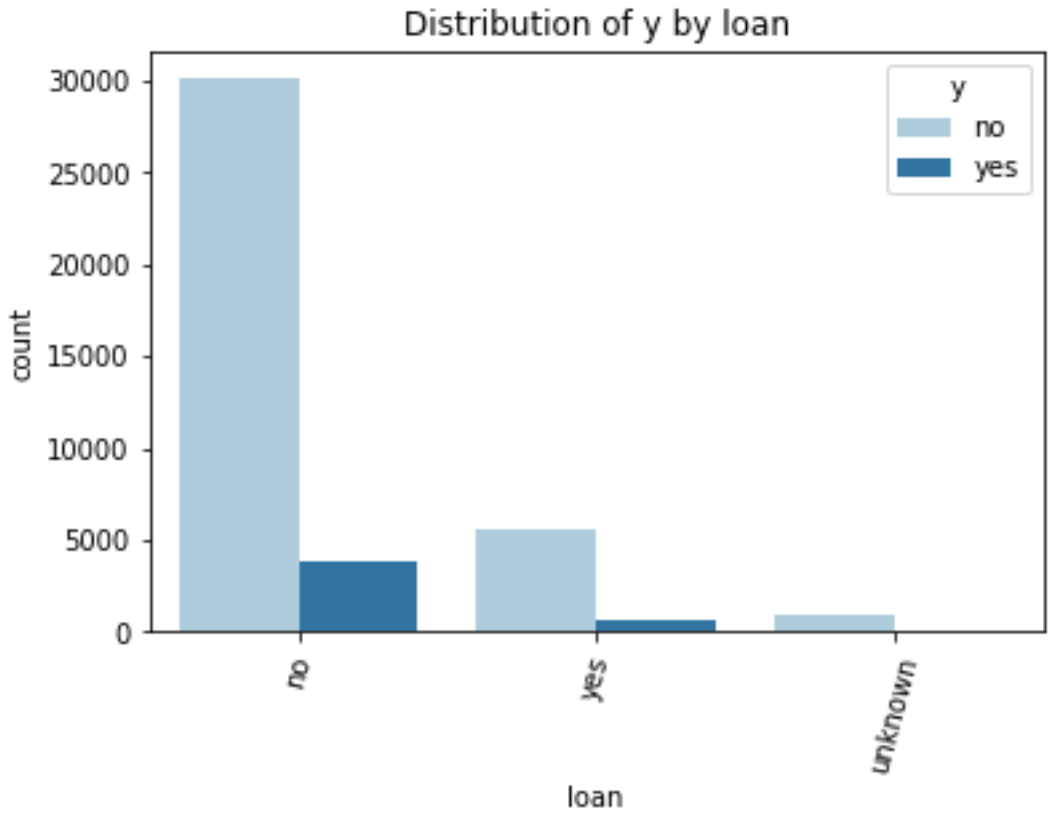




Majority of the clients were contacted by cellular and they subscribed more compared to those who used telephone.

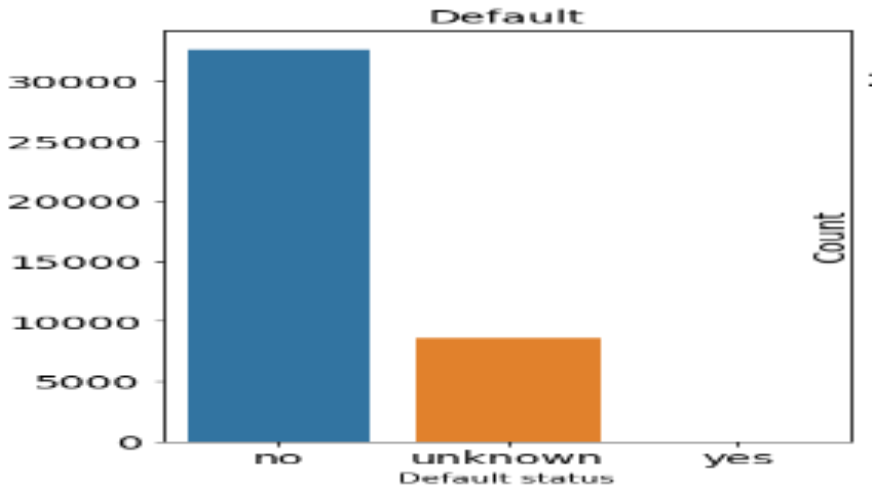
LOAN

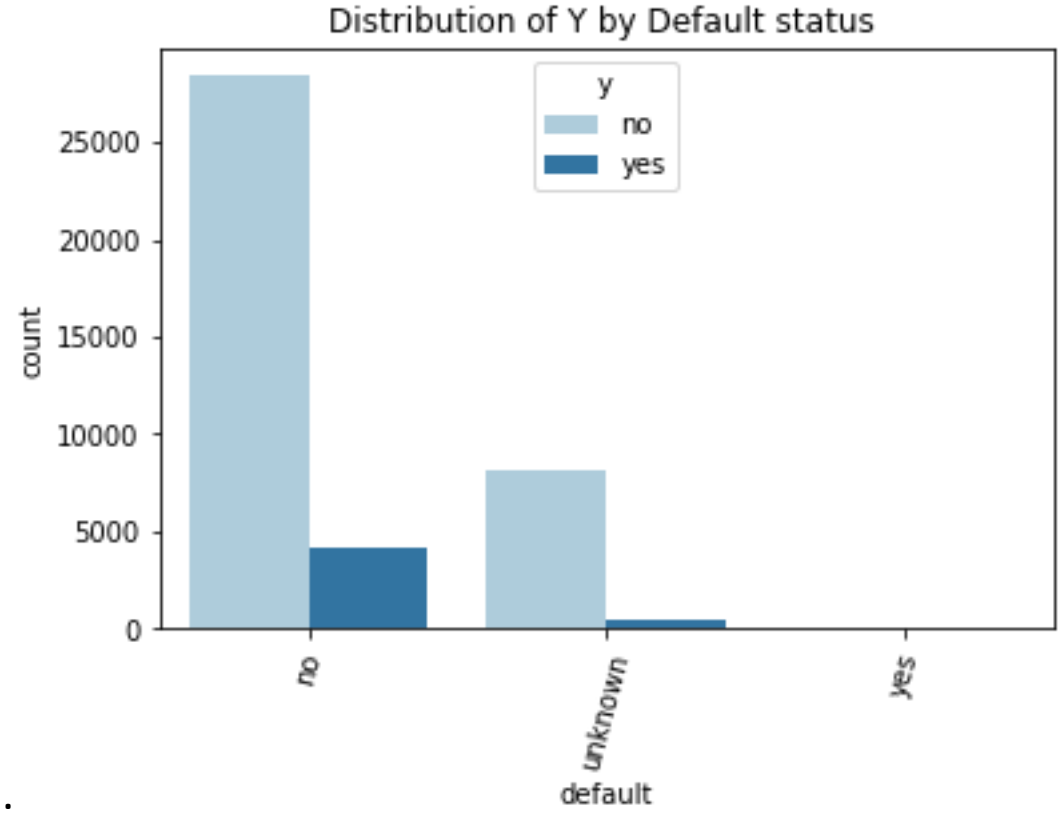




The bank mostly contacted clients who did not have loans.

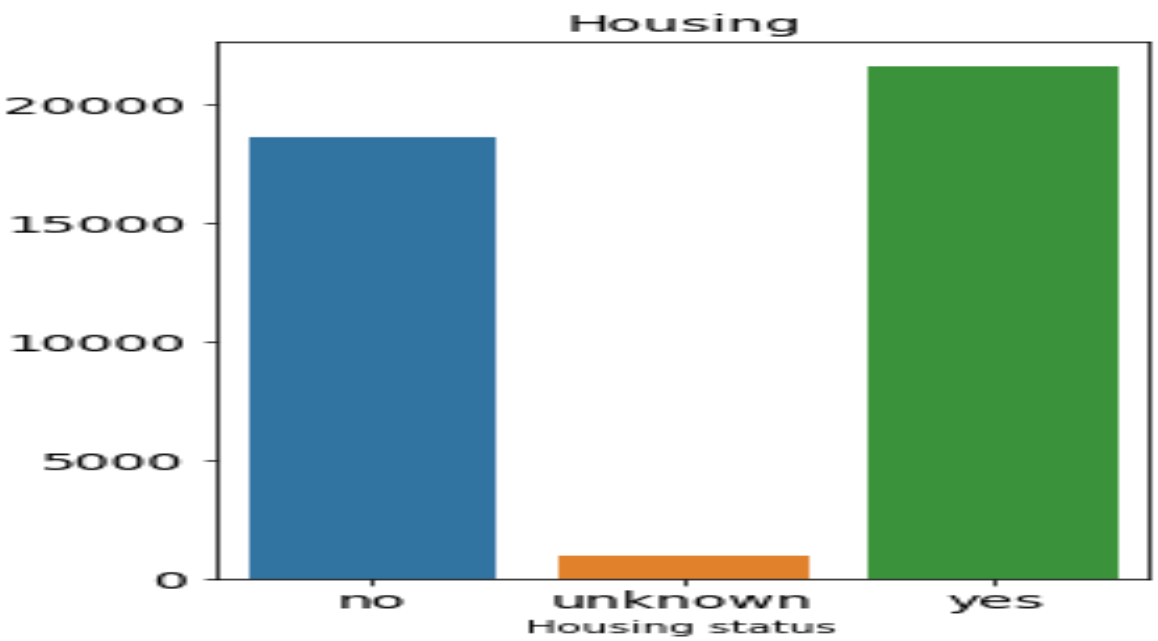
DEFAULT

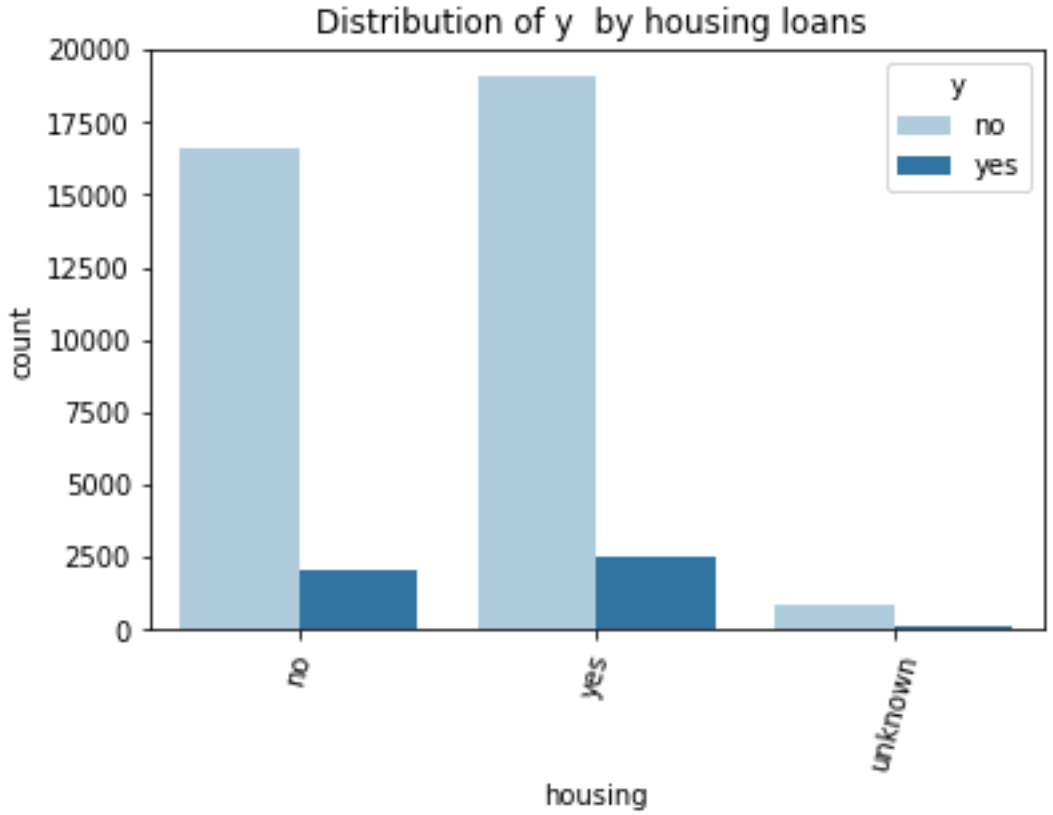




The contacted mostly clients who had not defaulted .
Most of the clients who had not defaulted in their credit and they subscribed more compared to those whose default status was unknown.

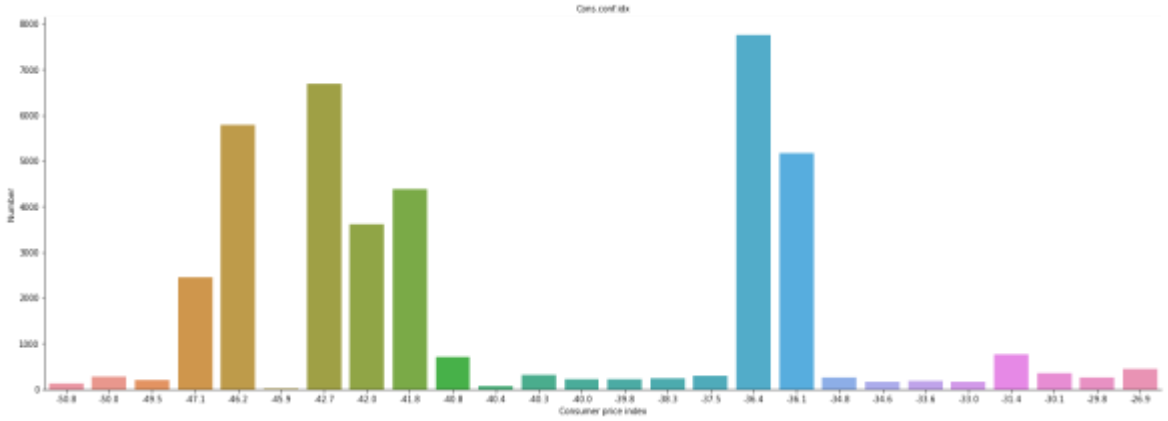
HOUSING

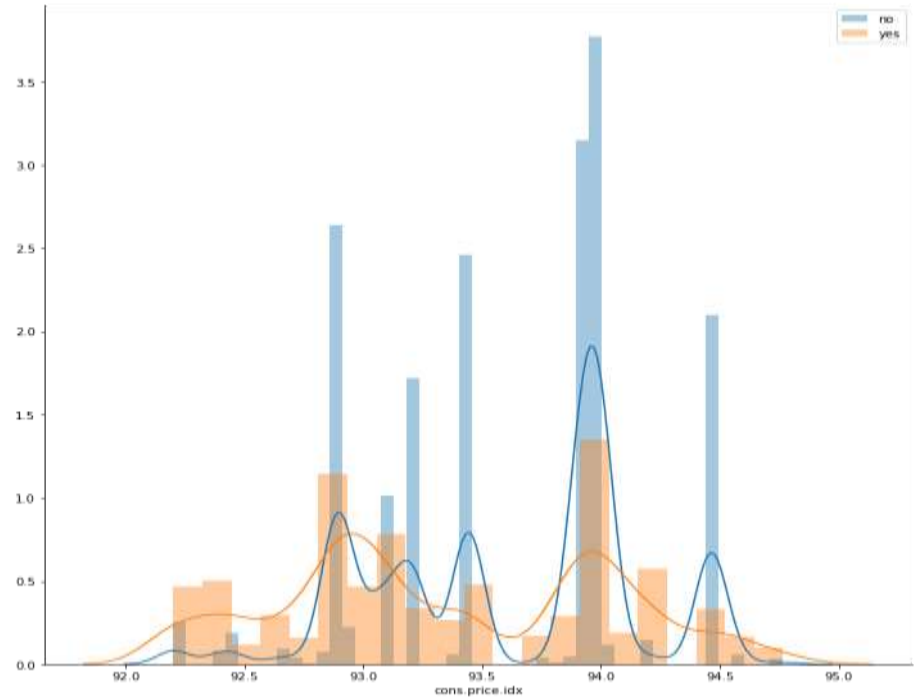




From the plots majority of clients had housing loans .
Those who had loans subscribed slightly more than those who didn't have loans.

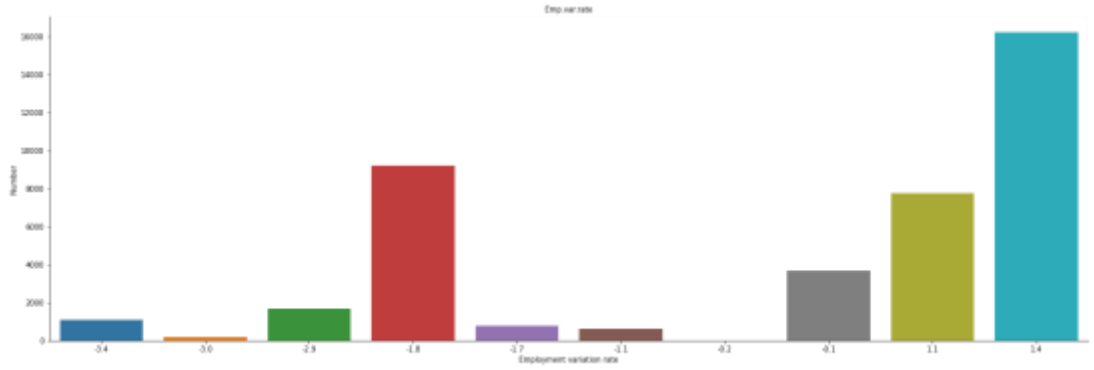
CONS.PRICE.IDX

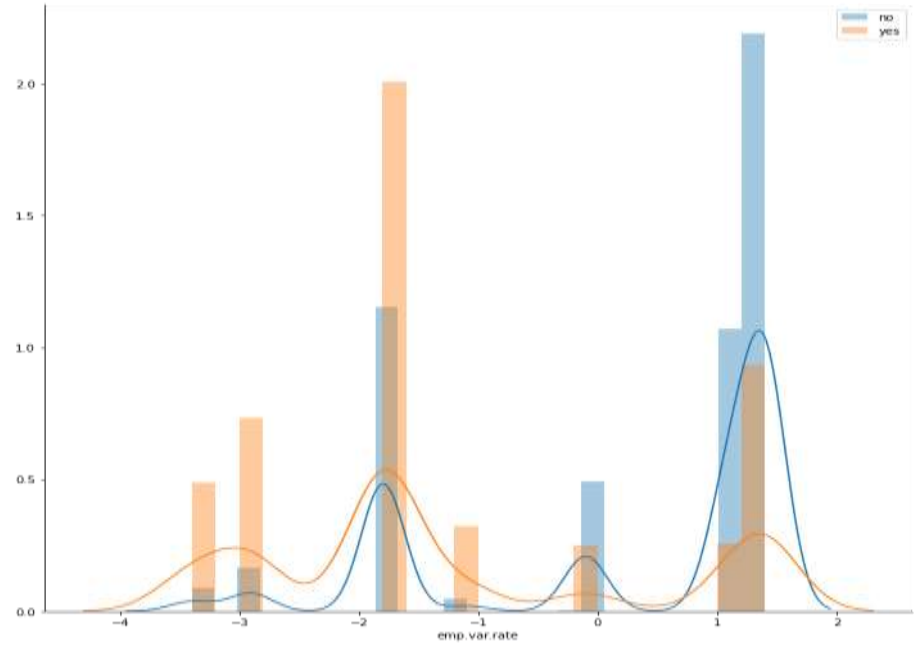




From the plot we can see cons.price.idx would be useful in prediction.

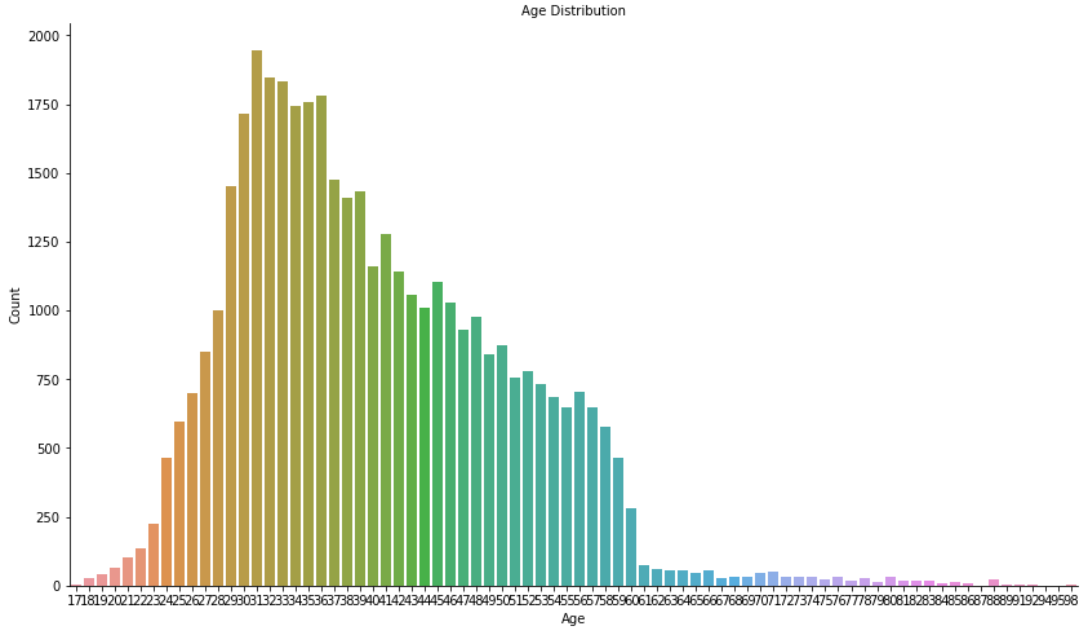
EMP.VAR.RATE

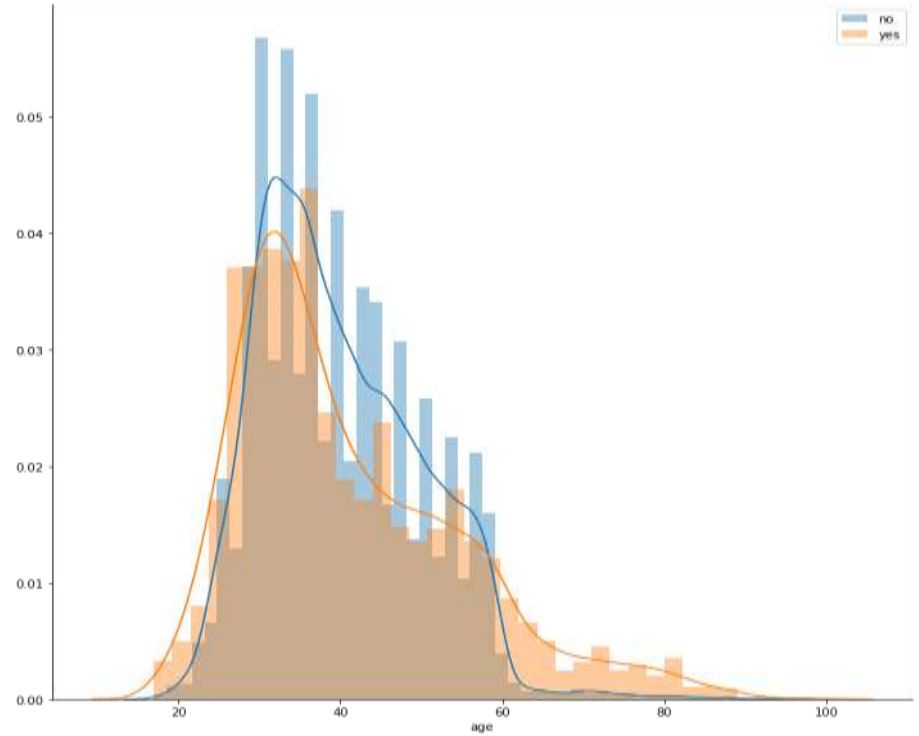




From the plots we see that emp.var.rate would be very useful in prediction.

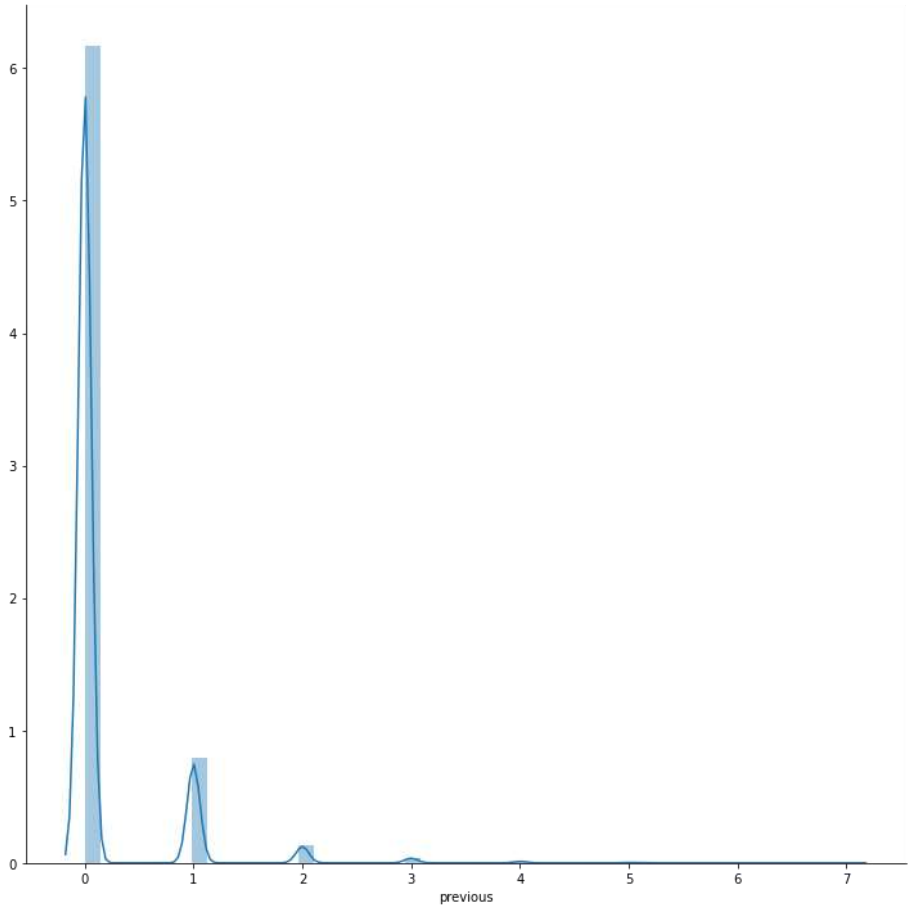
AGE

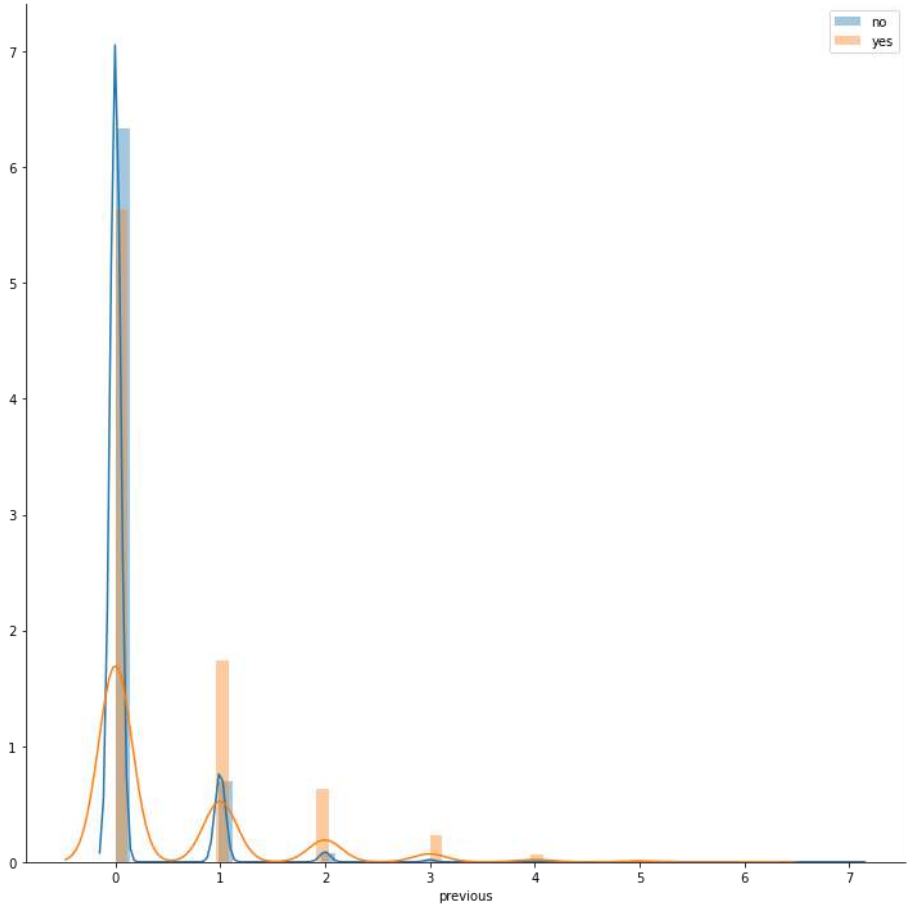




Clients age ranged from 17-98 yrs.

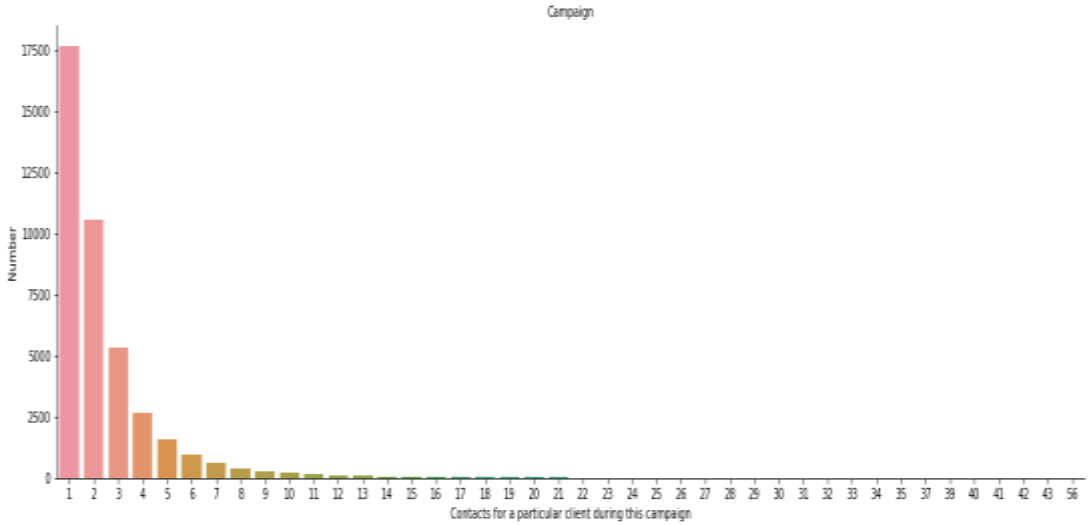
PREVIOUS

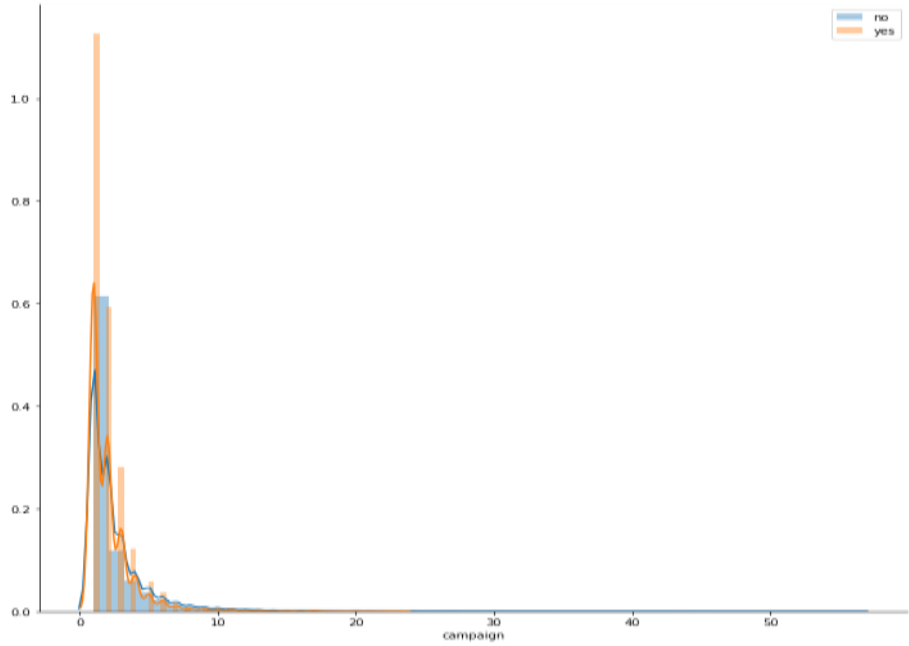




From the plot majority of the clients the bank contacted in this campaign had not been contacted in the last campaign.
It is also clear that previous will be very useful in prediction.

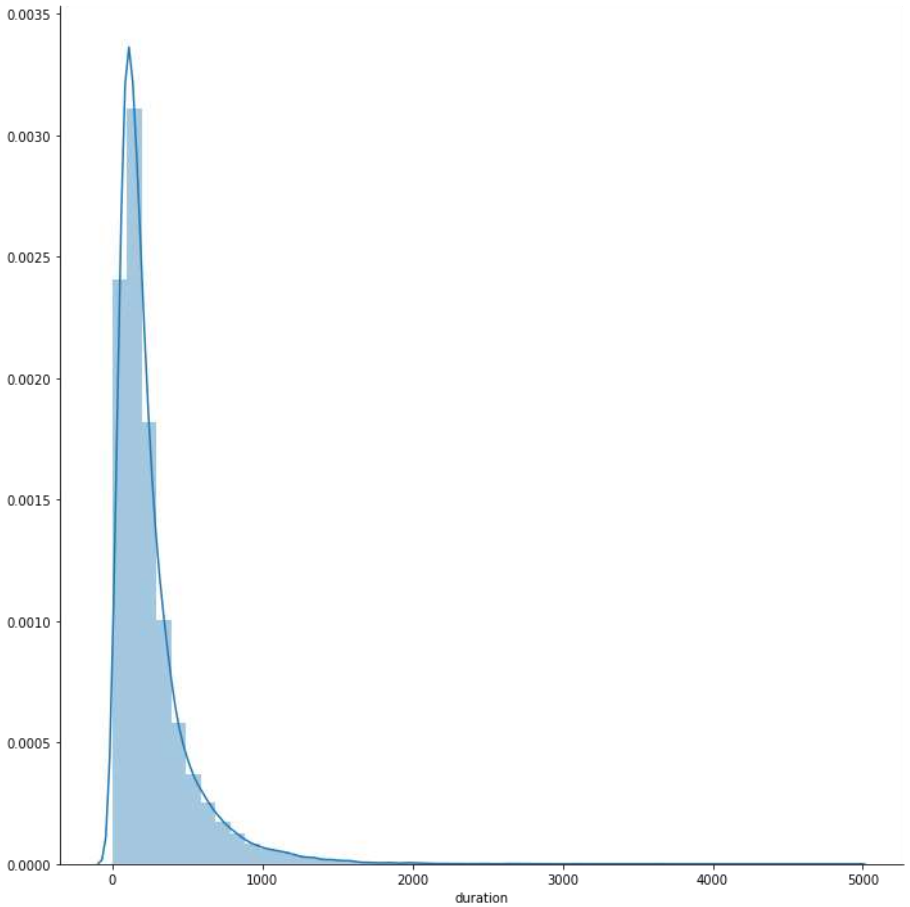
CAMPAIGN

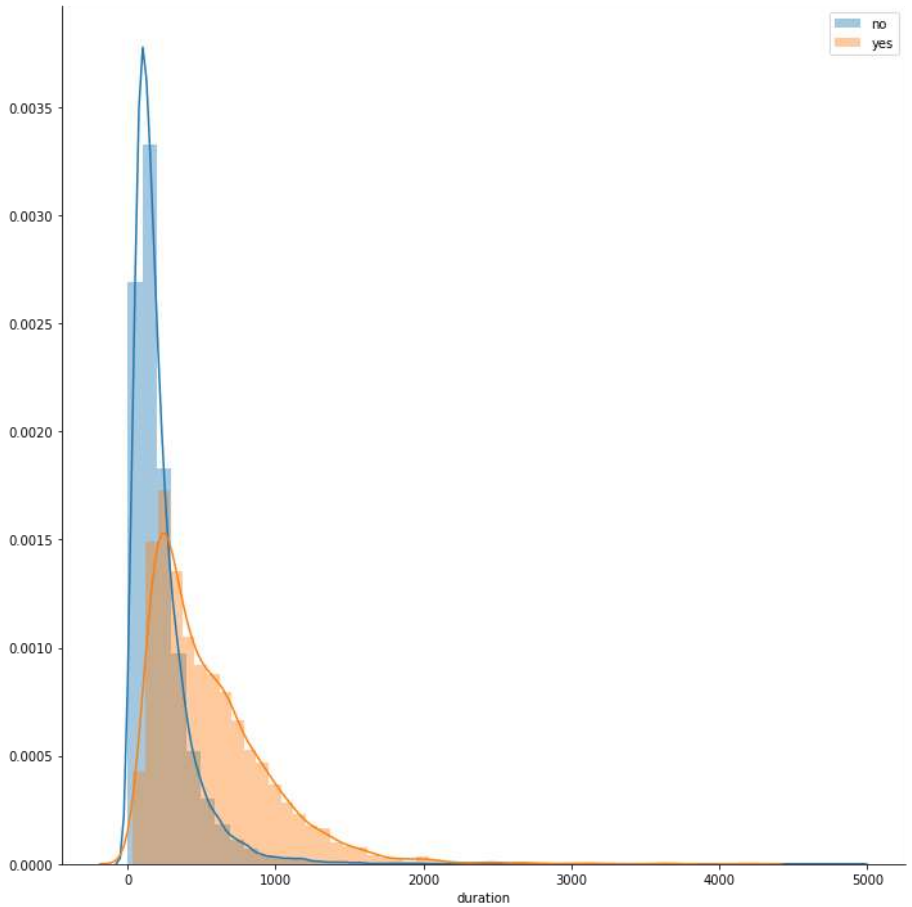




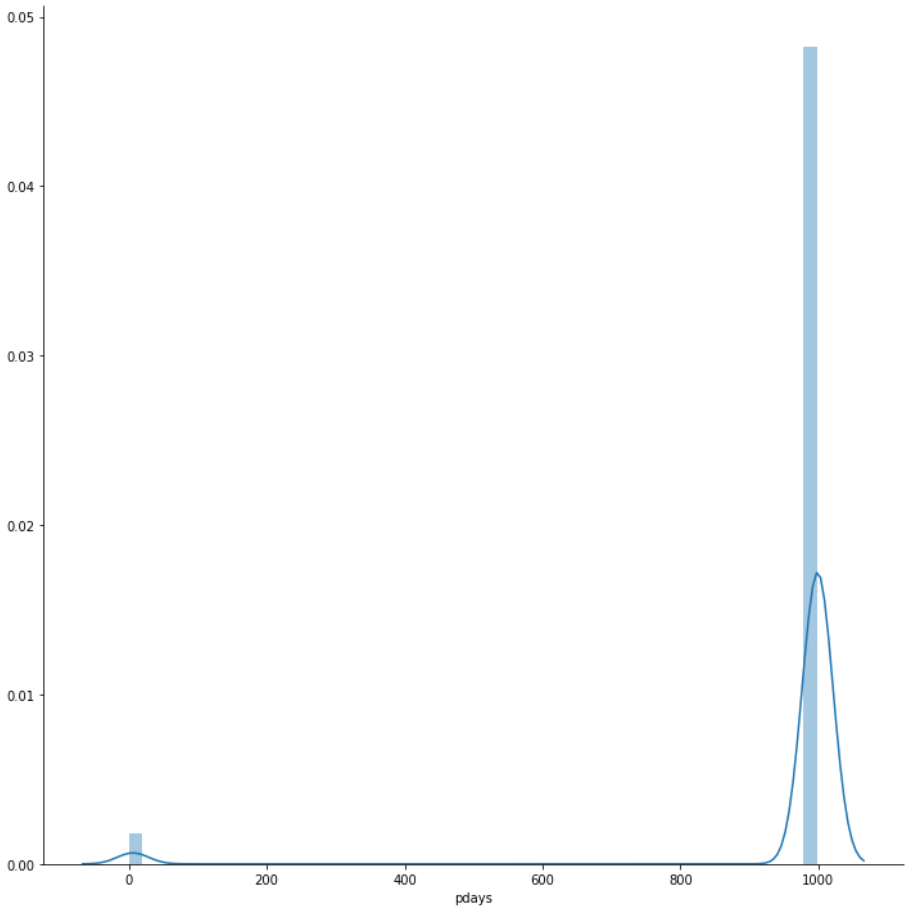
Majority of the clients were only contacted once during this campaign.

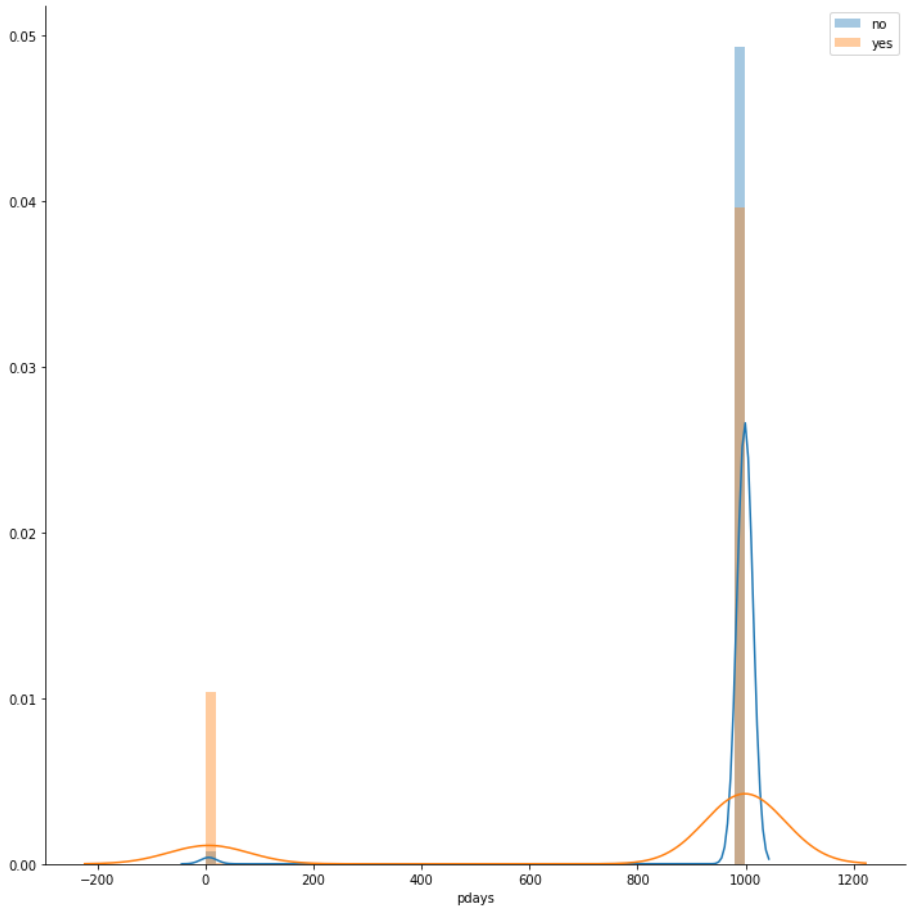
DURATION


0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
3000
4000
5000
duration

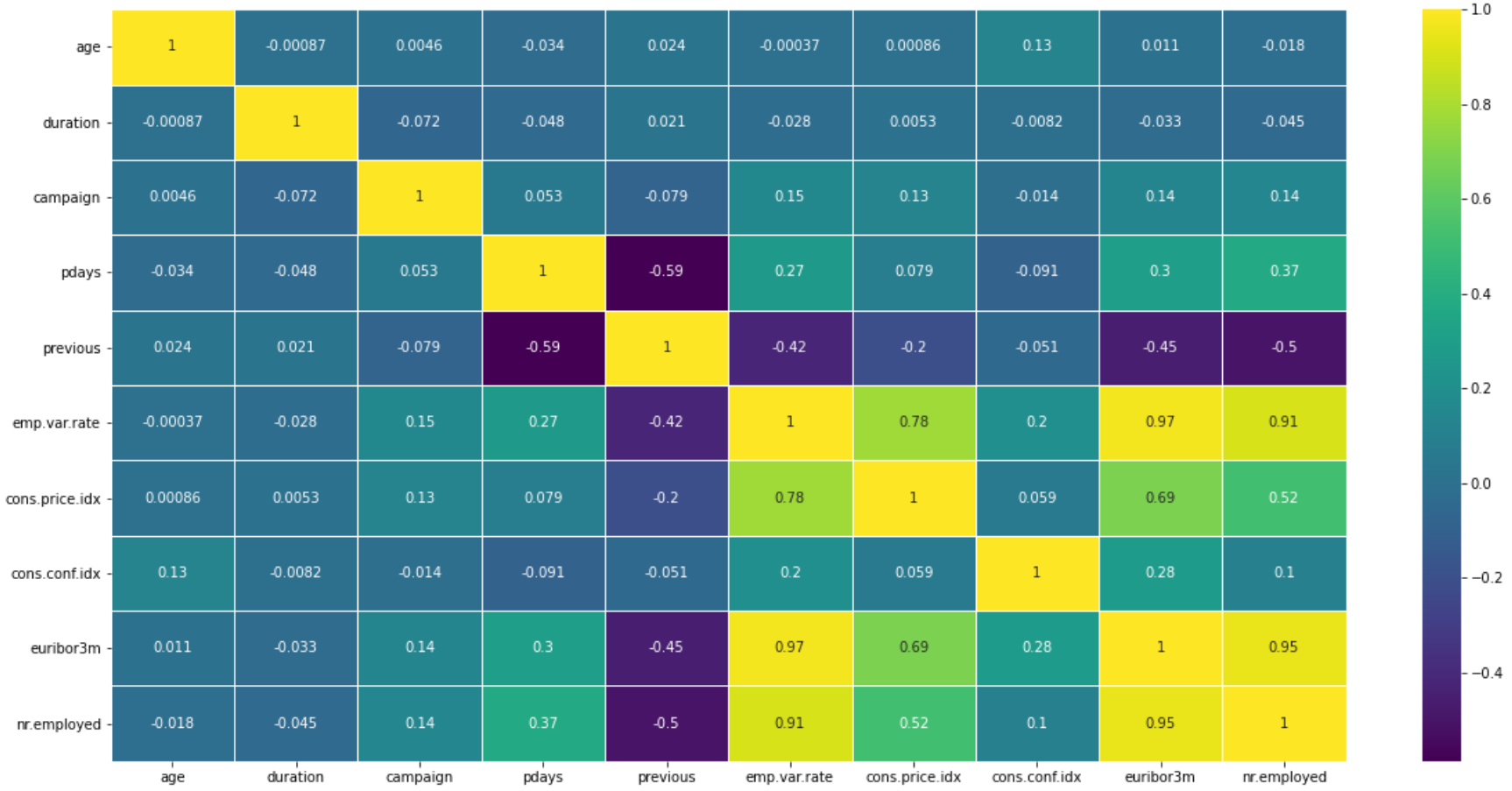

no
yes
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
3000
4000
5000
duration

PDAYS


0.05
0.04
0.03
0.02
0.01
0.00
0
200
400
600
800
1000
pdays


no
yes
0.05
0.04
0.03
0.02
0.01
0.00
−200
0
200
400
600
800
1000
1200
pdays

CORRELATION MATRIX



Emp.var. Rate, euribor3m, nr.employed and cons.price.idx have very high correlation

RECOMMENDED MODELS

- Logistic Regression
- XGBoost Models
- Multilayer Perceptron
- Random Forest

Thank You