Data Glacier
Your Deep Learning Partner

Customer Segmentation

EDA & Model Recommendation

Virtual Internship

3rd Oct 2021

Background

- XYZ bank wants to roll out Christmas offers to their customers. But Bank does not want to roll out same offer to all customers instead they want to roll out personalized offer to particular set of customers. If they manually start understanding the category of customer then this will be not efficient and also they will not be able to uncover the hidden pattern in the data (pattern which group certain kind of customer in one category). Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want **more than 5 group** as this will be inefficient for their campaign.

- Objective : Provide actionable insights to help the bank in segregation customers in such a way that the bank can focus on clients who are more likely to agree to the offer

The analysis has been divided into three parts:

- Data Description
- EDA
- Model Recommendation

Data Description

- age (numeric)
- job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")
- marital : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
- education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")
- default: has credit in default? (categorical: "no","yes","unknown")
- housing: has housing loan? (categorical: "no","yes","unknown")
- loan: has personal loan? (categorical: "no","yes","unknown")
- related with the last contact of the current campaign
- contact: contact communication type (categorical: "cellular","telephone")

Data Description

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

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

- 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.

- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

- 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)

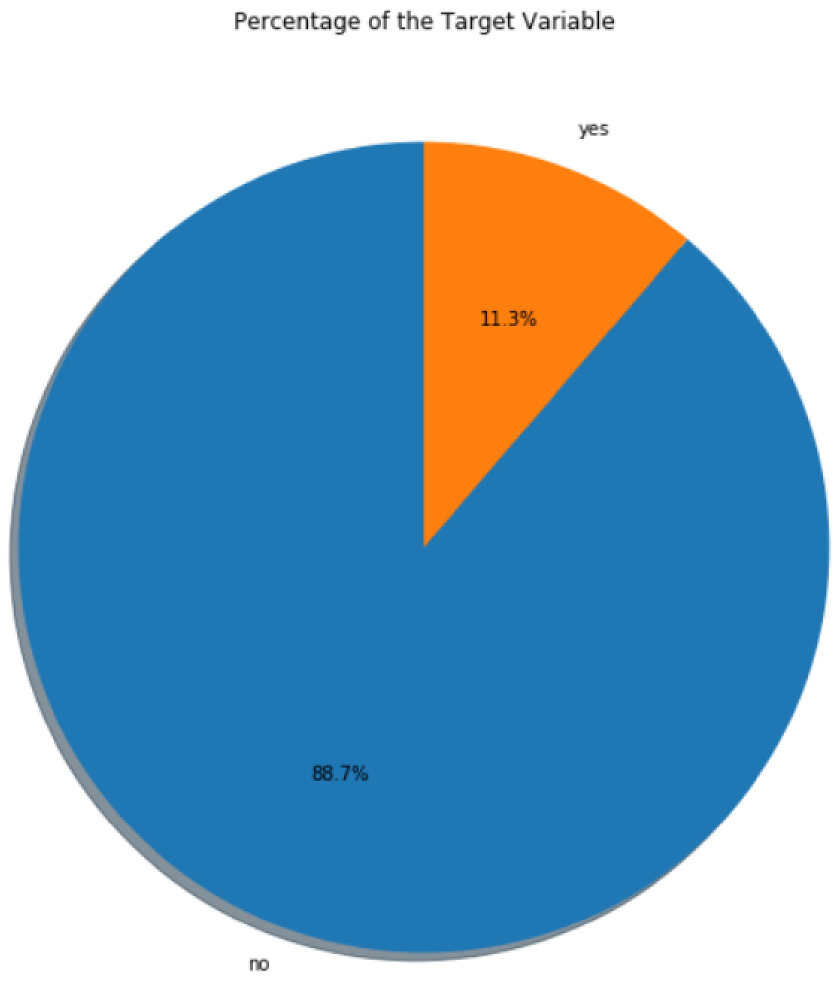
- previous: number of contacts performed before this campaign and for this client (numeric)

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

Data Description

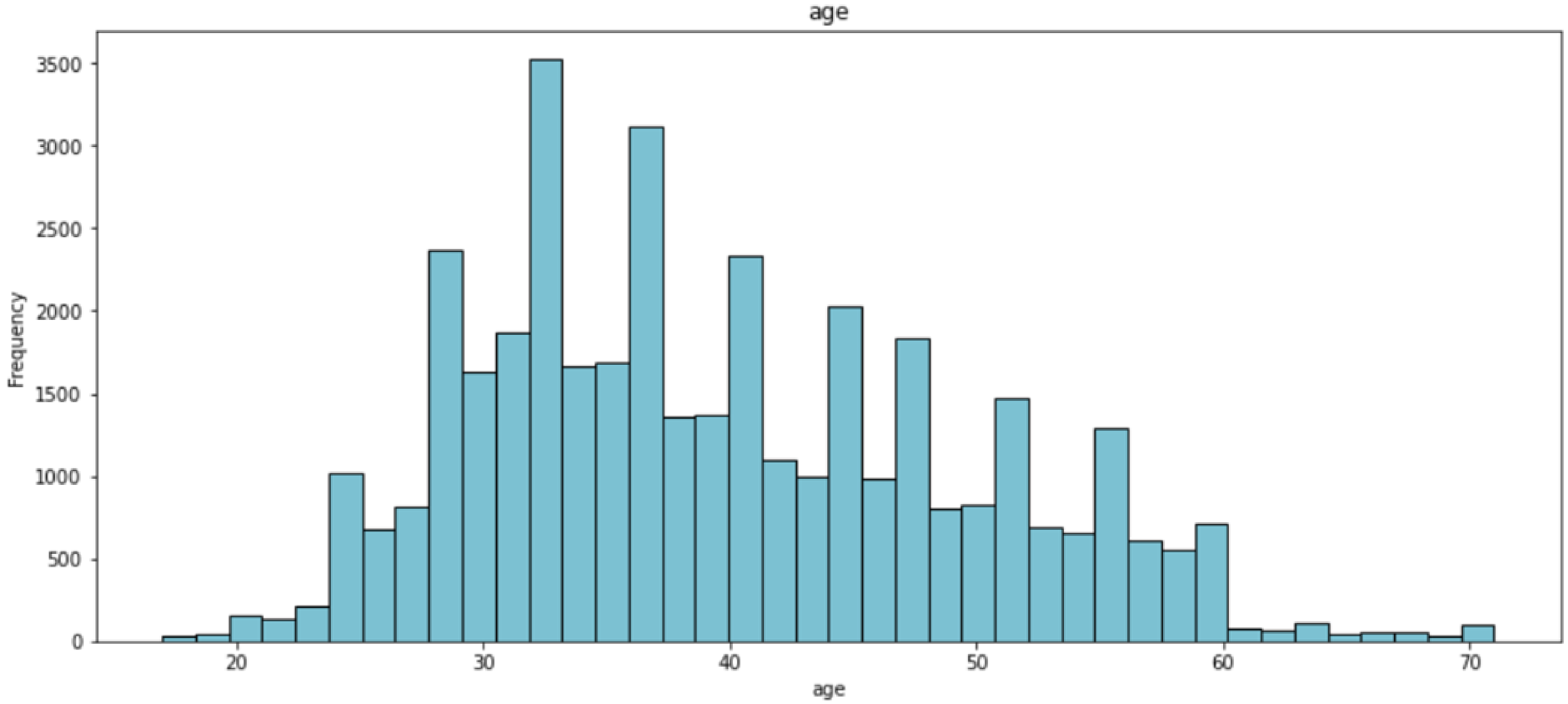
- emp.var.rate (EVR): employment variation rate - quarterly indicator (numeric)
- cons.price.idx (CPI): consumer price index - monthly indicator (numeric)
- cons.conf.idx (CCI): consumer confidence index - monthly indicator (numeric)
- euribor3m: euribor 3 month rate - daily indicator (numeric)
- nr.employed: number of employees - quarterly indicator (numeric)
- y - has the client subscribed a term deposit? (binary: "yes","no")

Target Variable



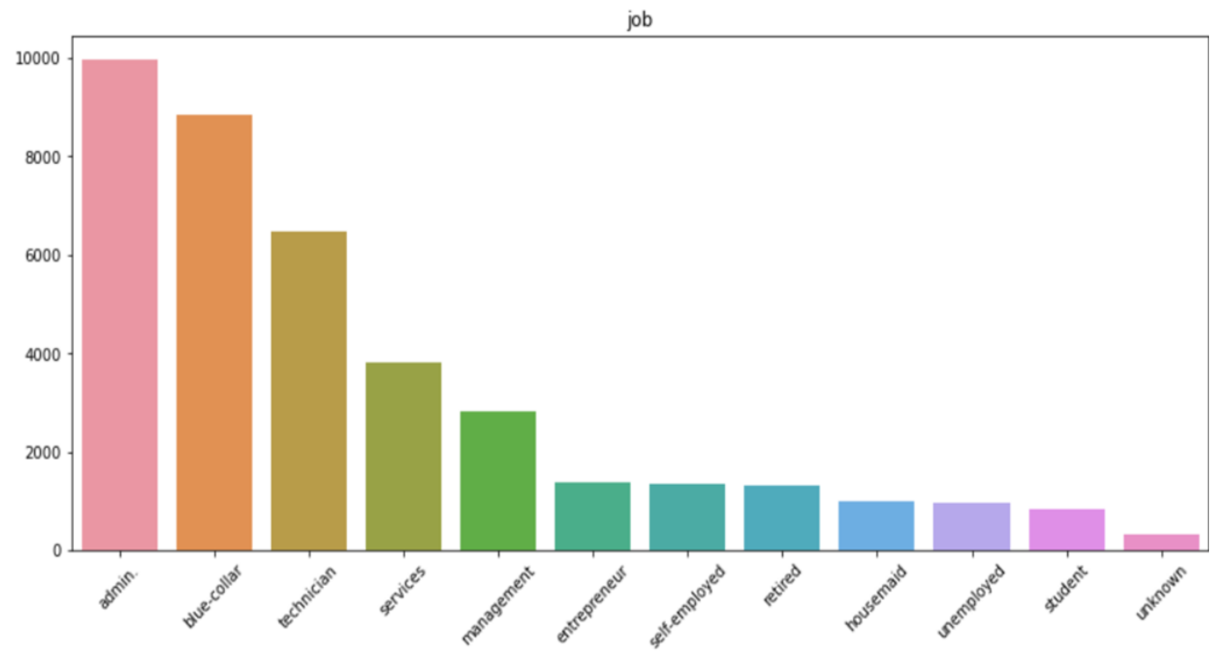
- The target column, i.e. ‘y’ tells whether the client had subscribed to term deposit or not
- From the above piechart, we notice that 11.3% of clients have subscribed a term deposit. In other words, 11.3% of the campaign calls are successful.

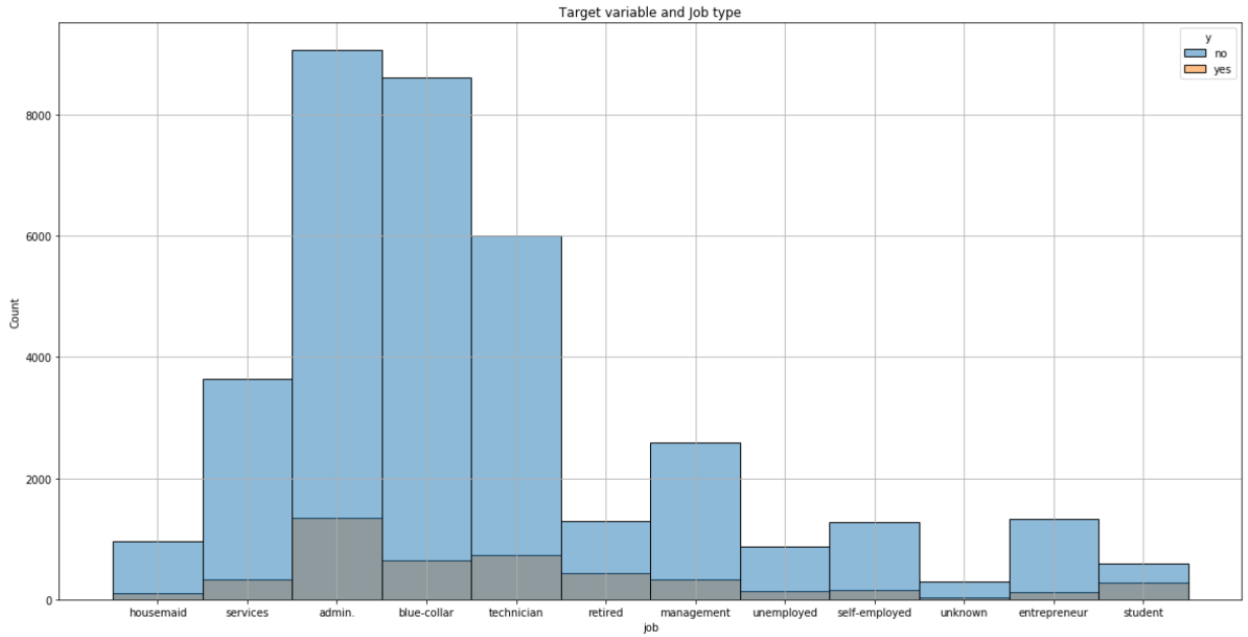
Age range of Clients



- 97.5% of the clients fall between ages 20 and 60

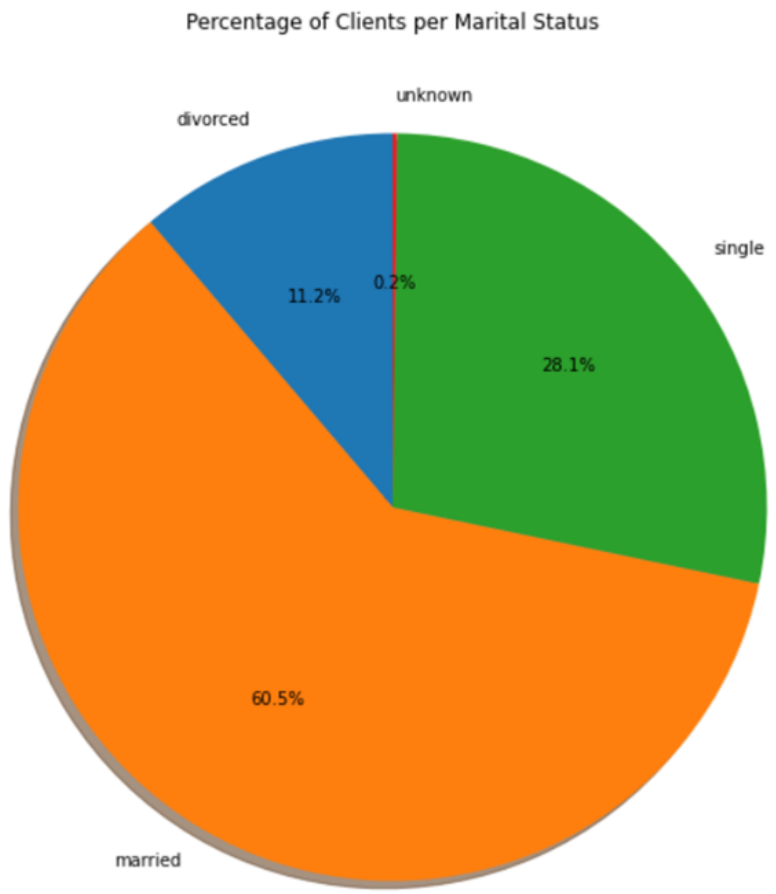
Job Description and its effect on Target Variable

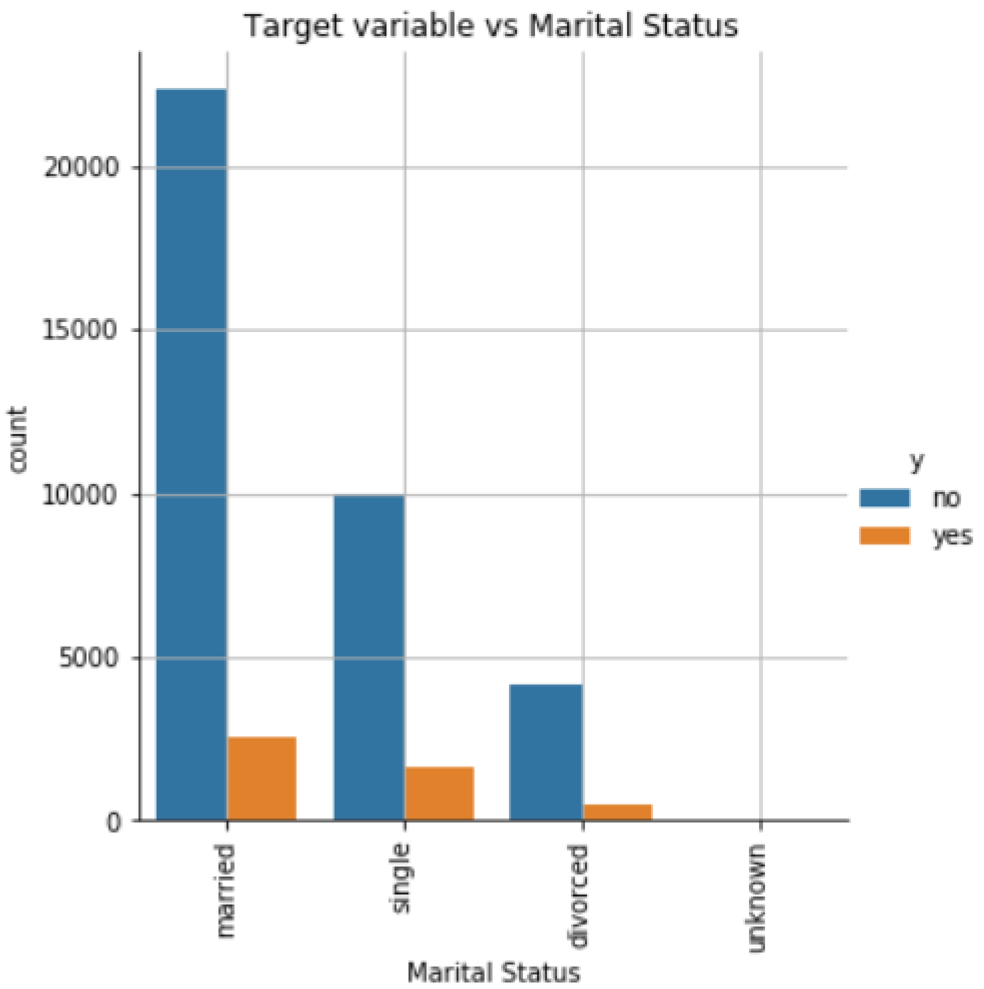




- The clients best 4 job types are admin, blue-collar, technician, and services.
- Customers from admin,blue-collar and technician job types open a deposit account

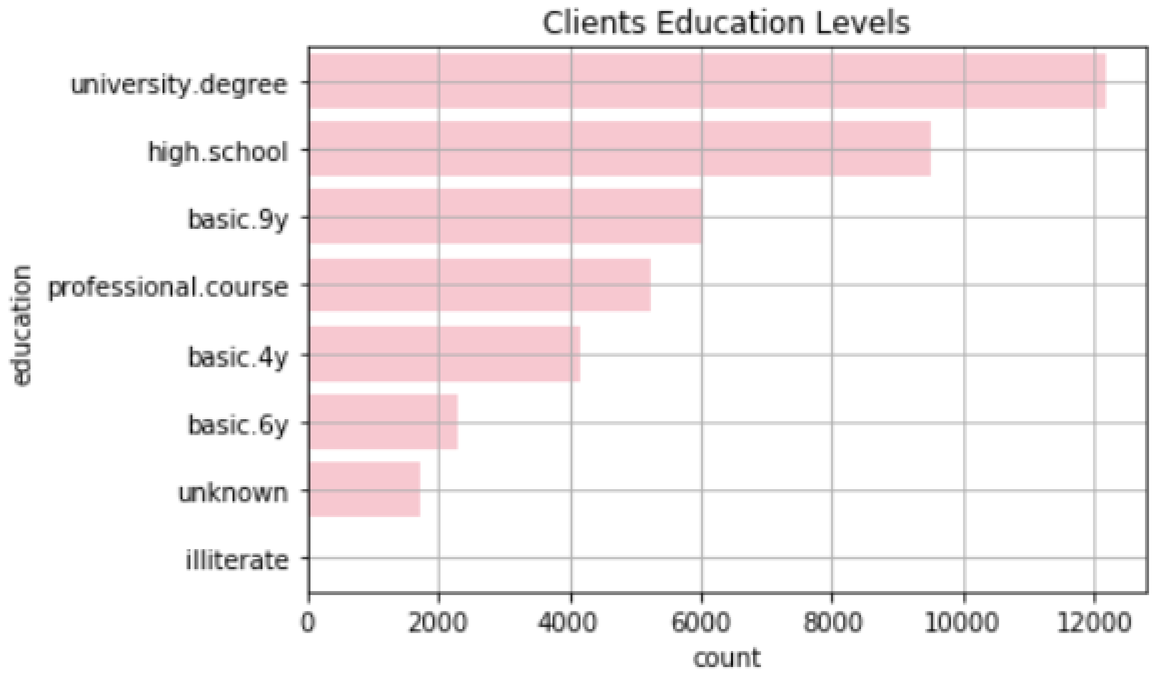
Customers Marital Status and its Effect

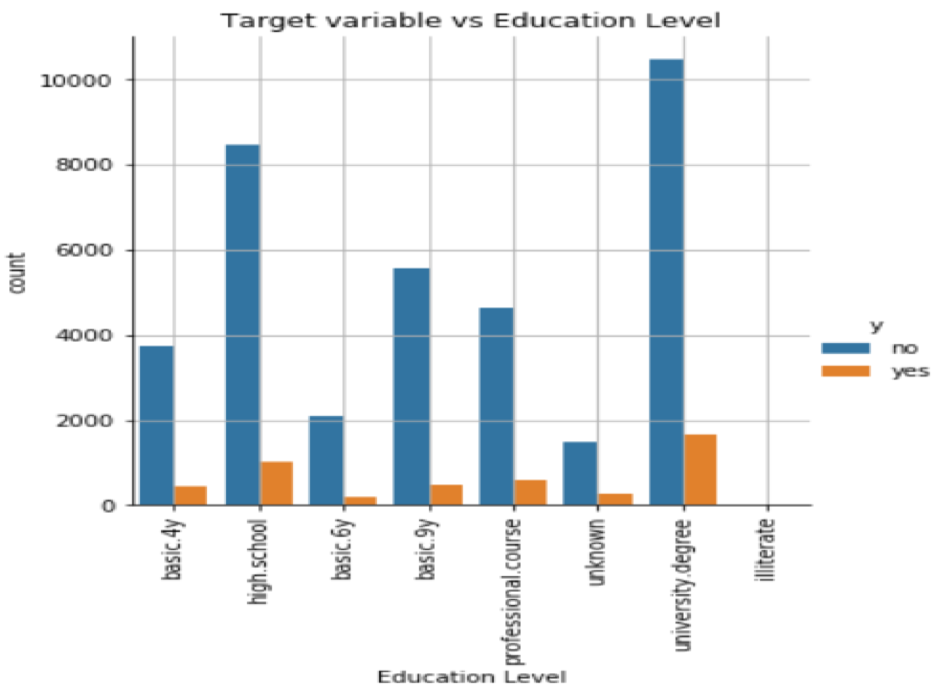




- 60.5% of the clients are married Customers who are married more open to a deposit account compared to other marital status

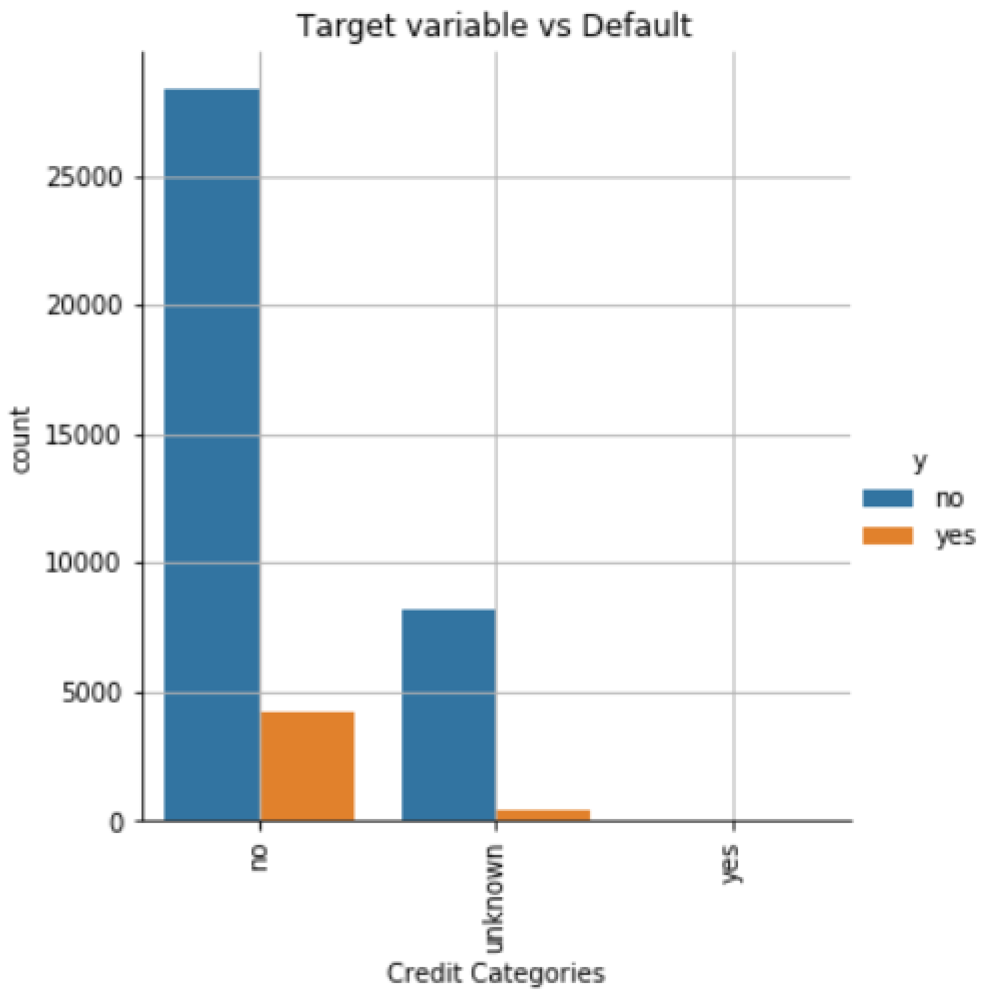
Customers Education Level and its Effect

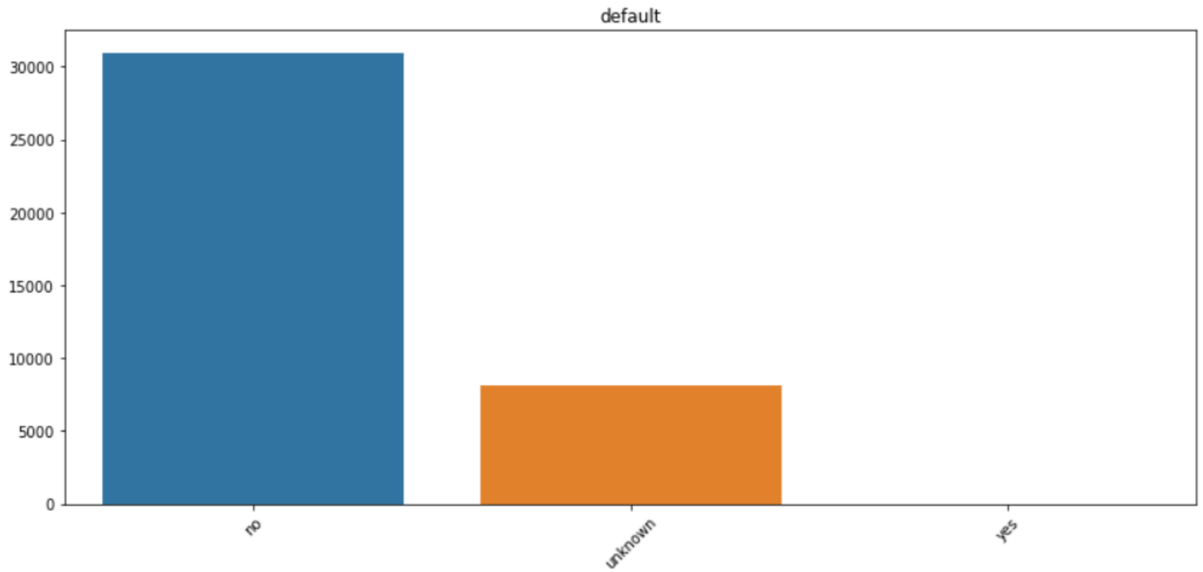




- The majority of clients have completed the university degree followed by the high school degree
- While 1730 clients have unknown education level
- Customers with university degree education level are more likely to sign up for schemes

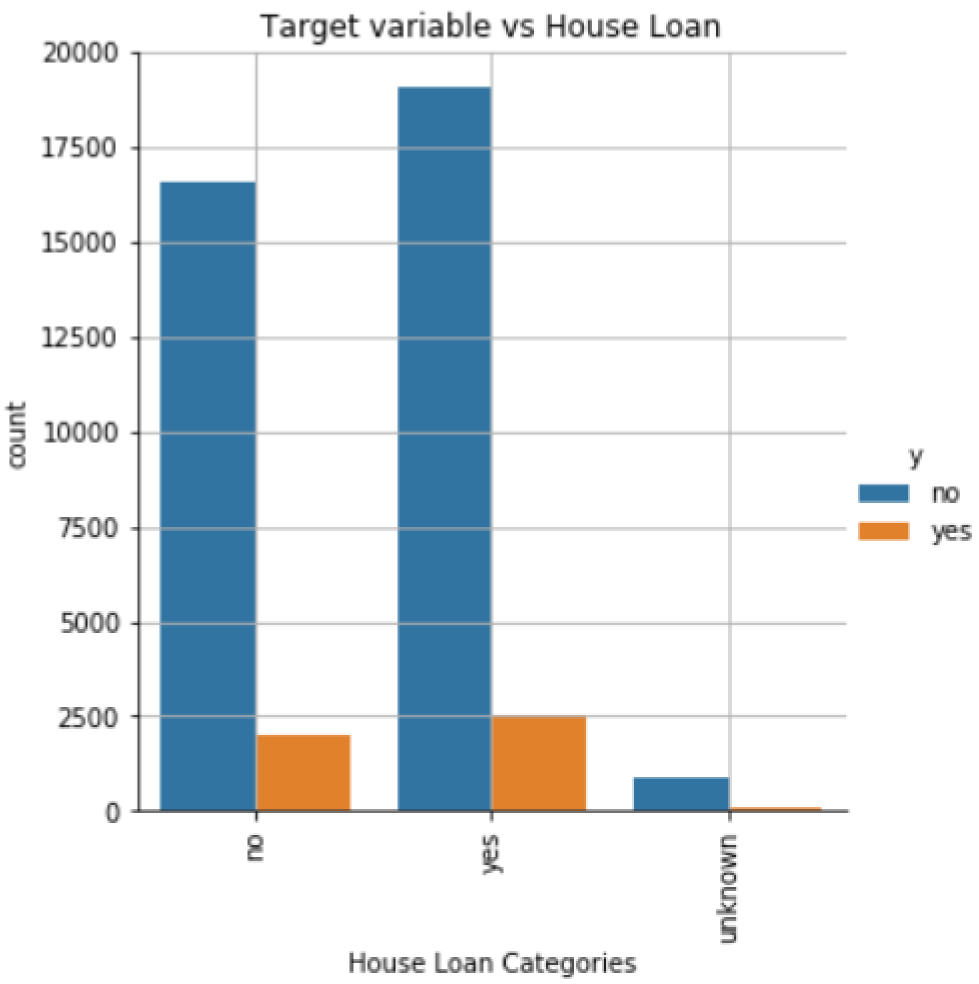
Customers Default Credit Category and its Effect

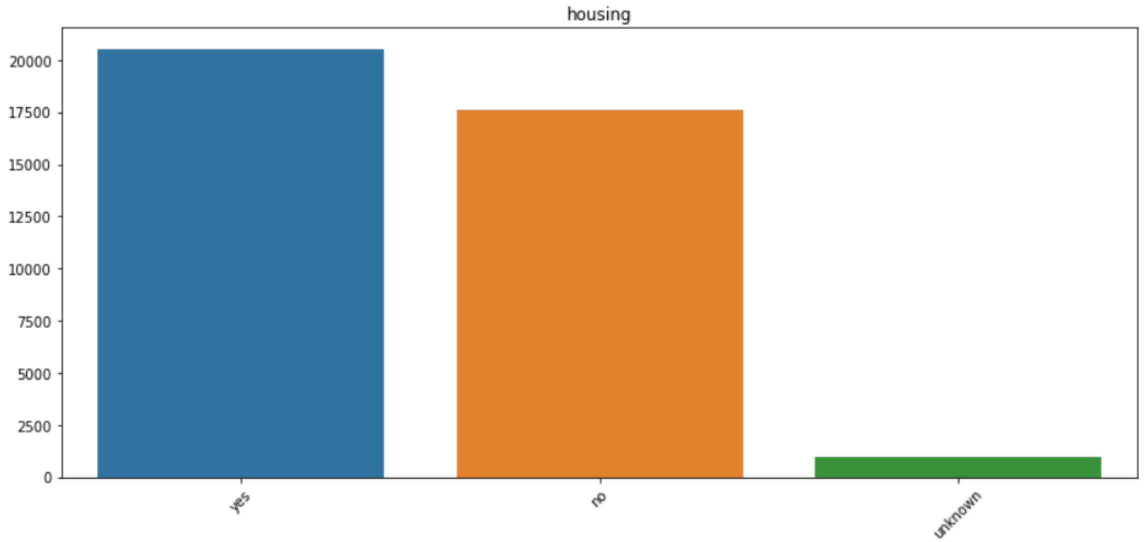




- Customers who have no credits in default open more deposit account compared to other categories

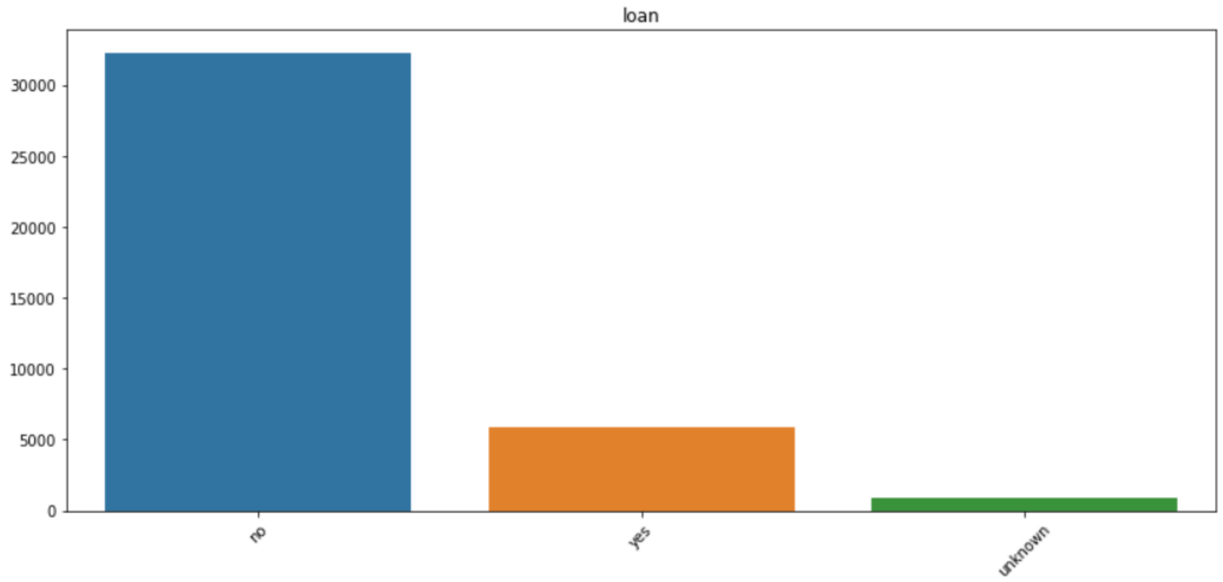
Customers Housing Loan Status and its Effect





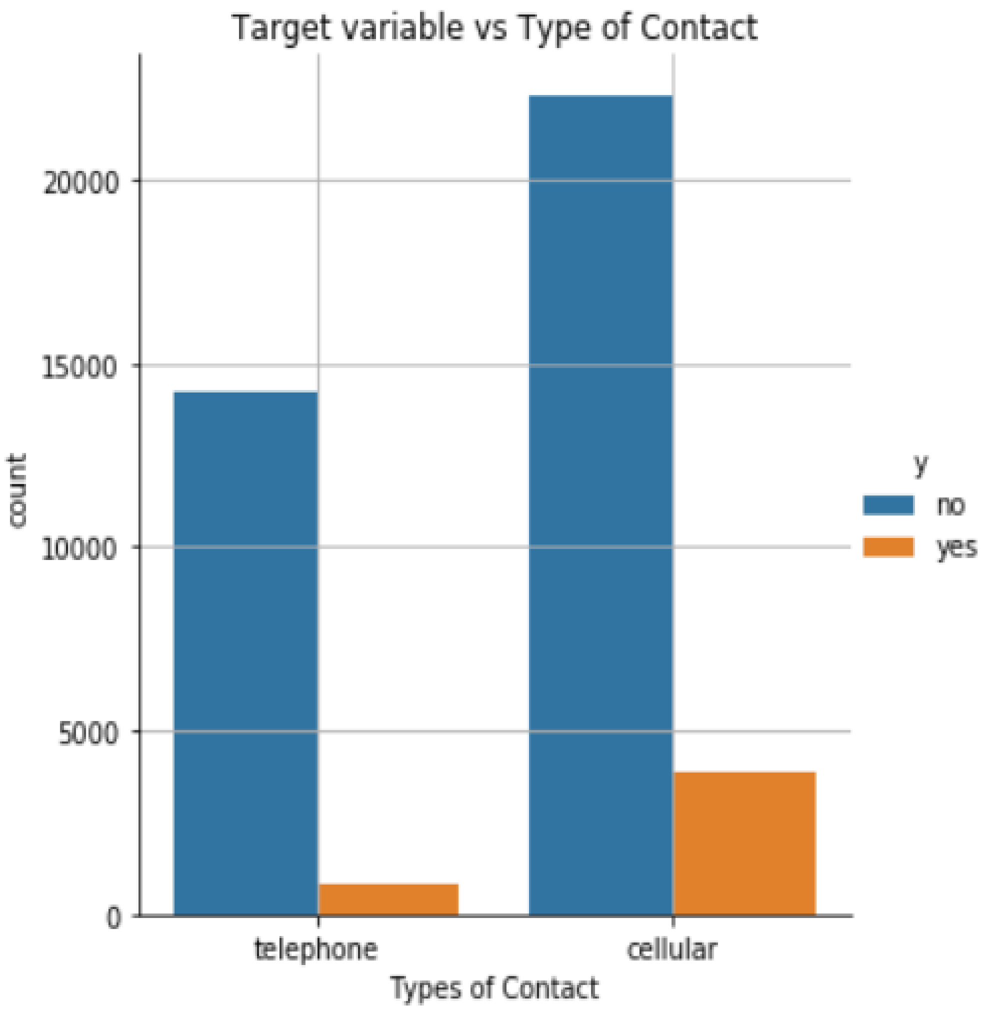
- Customers who have house loan open more deposit account compared to others categories.

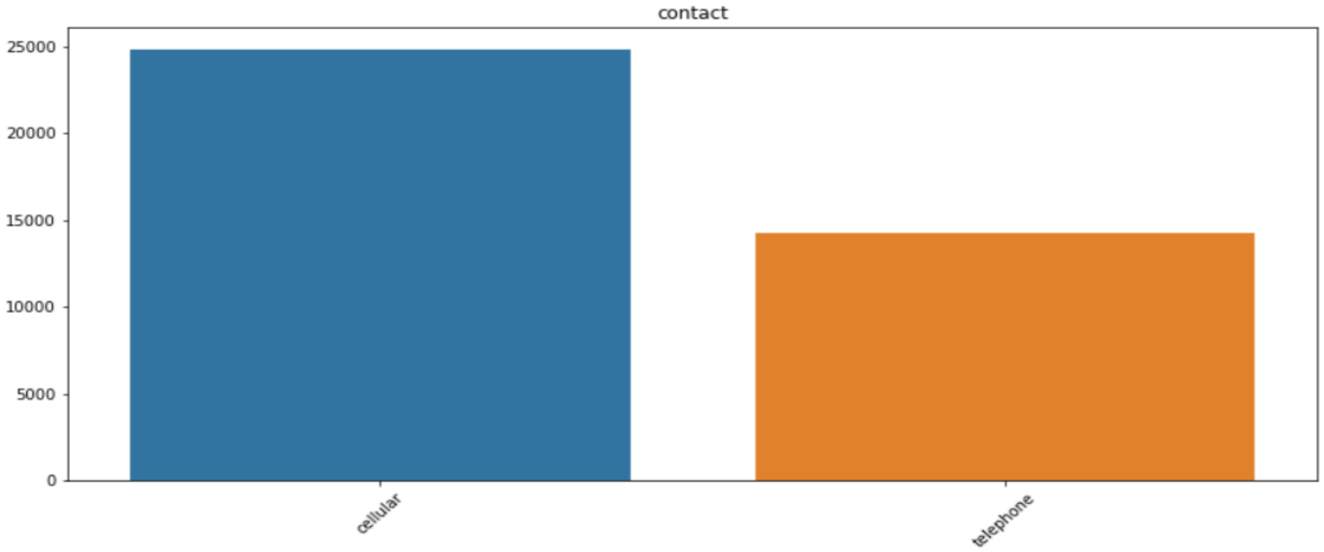
Customers Personal Loan Status and its Effect



- Customers who have no personal loan open more deposit accounts compared to others categories.

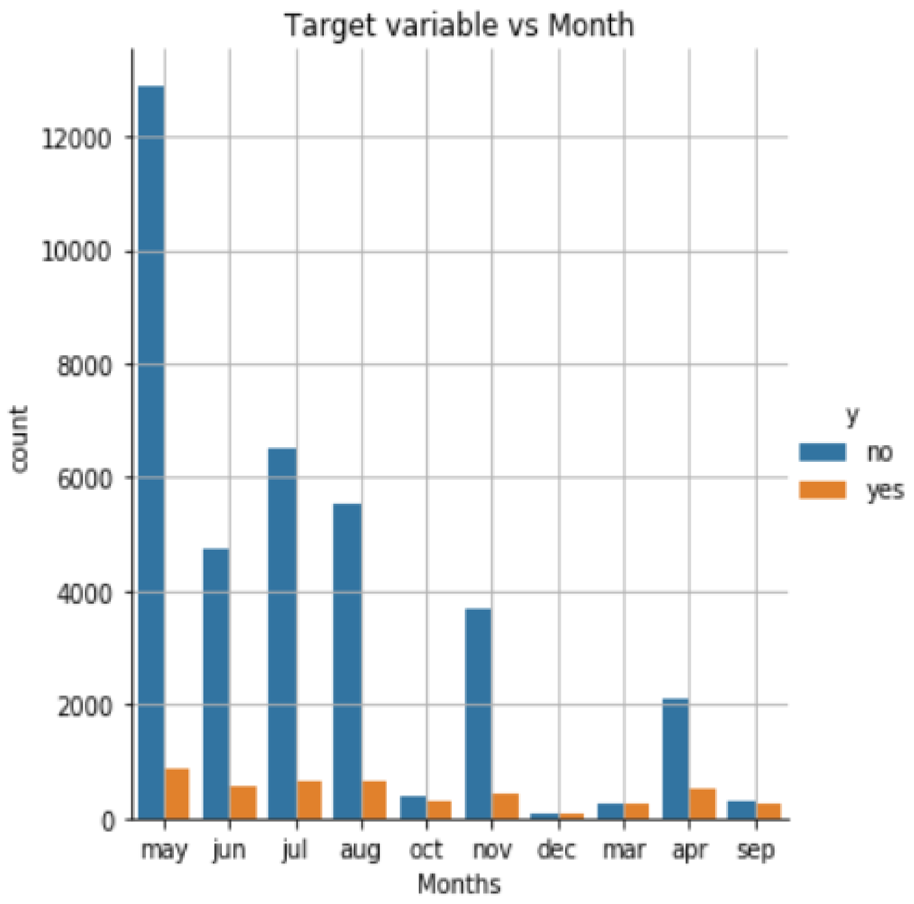
Customers Contact and its Effect

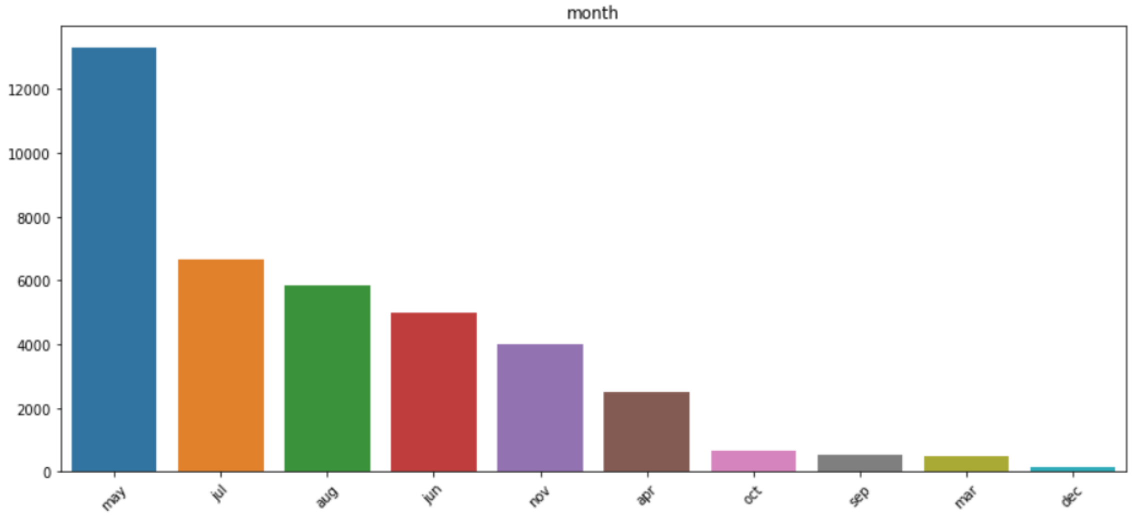




- The majority of contacts is of cellular type. And the clients that open a deposit accounts are almost all contacted by cellular.

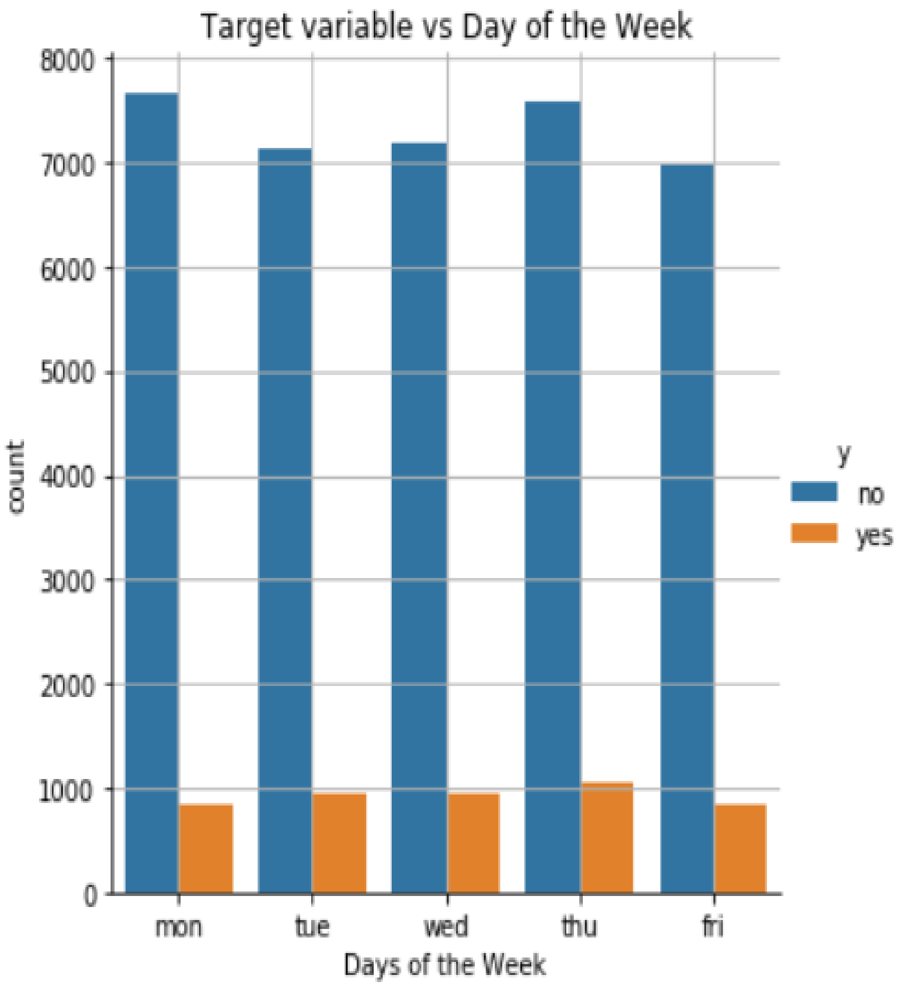
Month and its Effect

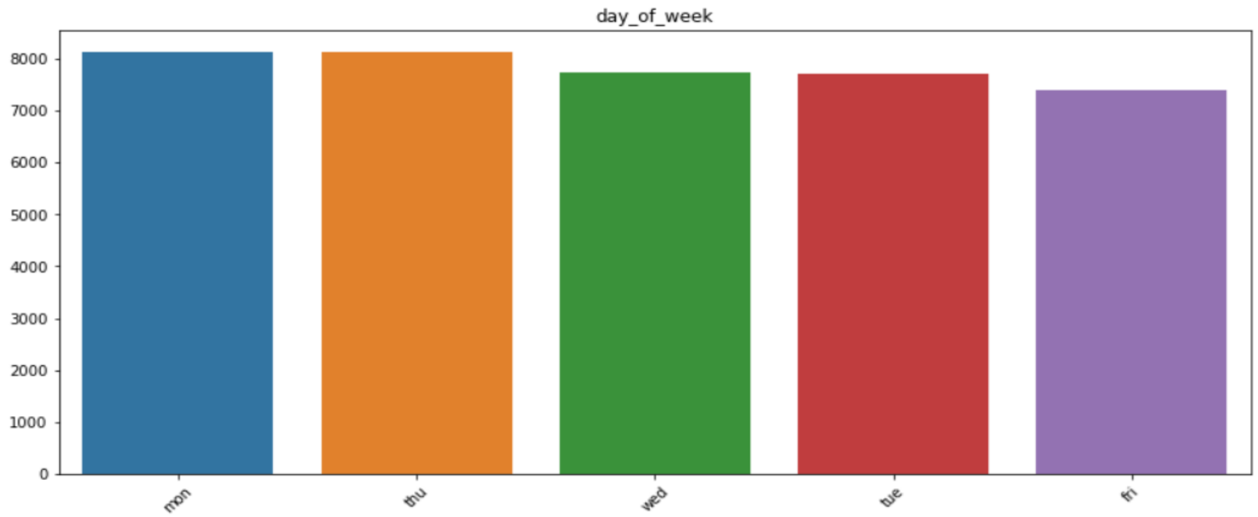




We can notice that the majoriy of calls were in the month of May. Moreover, the majority of clients that open a deposit account were contacted in the month of May. But comparing the number of call to the number of people signing up, the rate is quite poor for May.

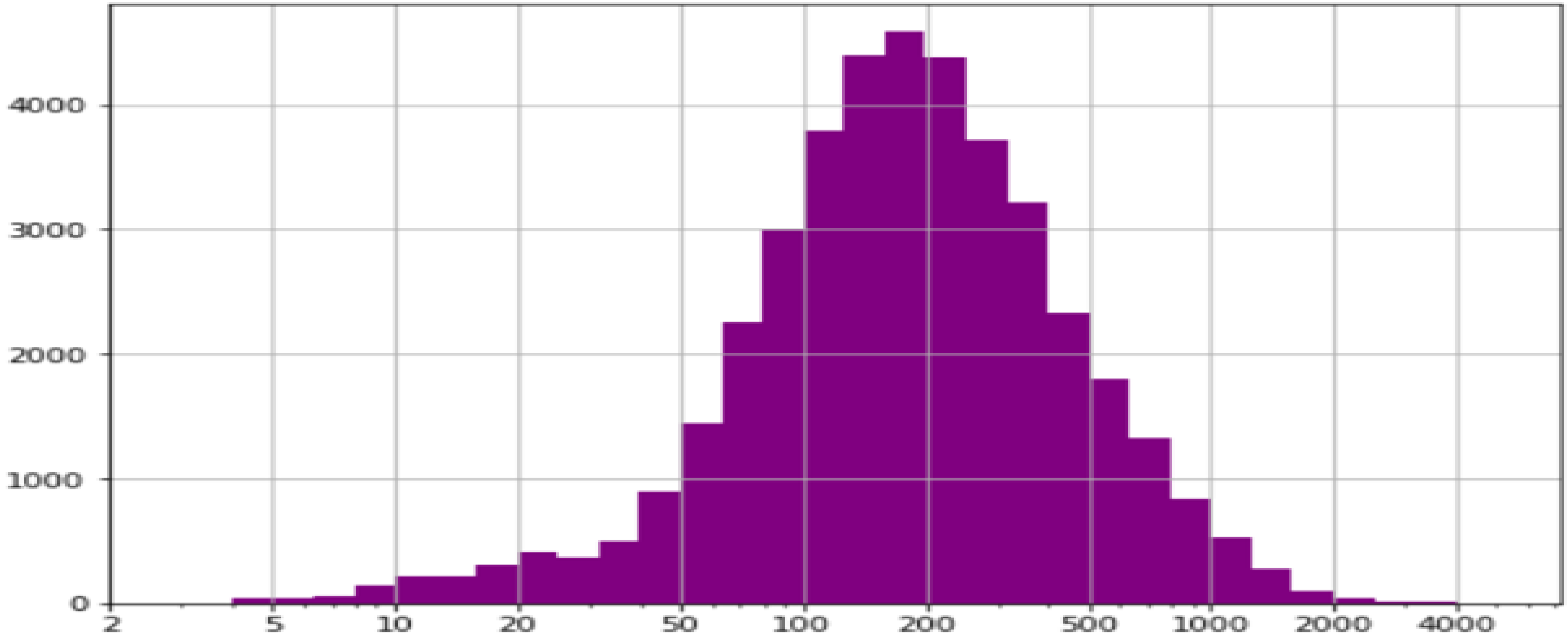
Days of week and its Effect





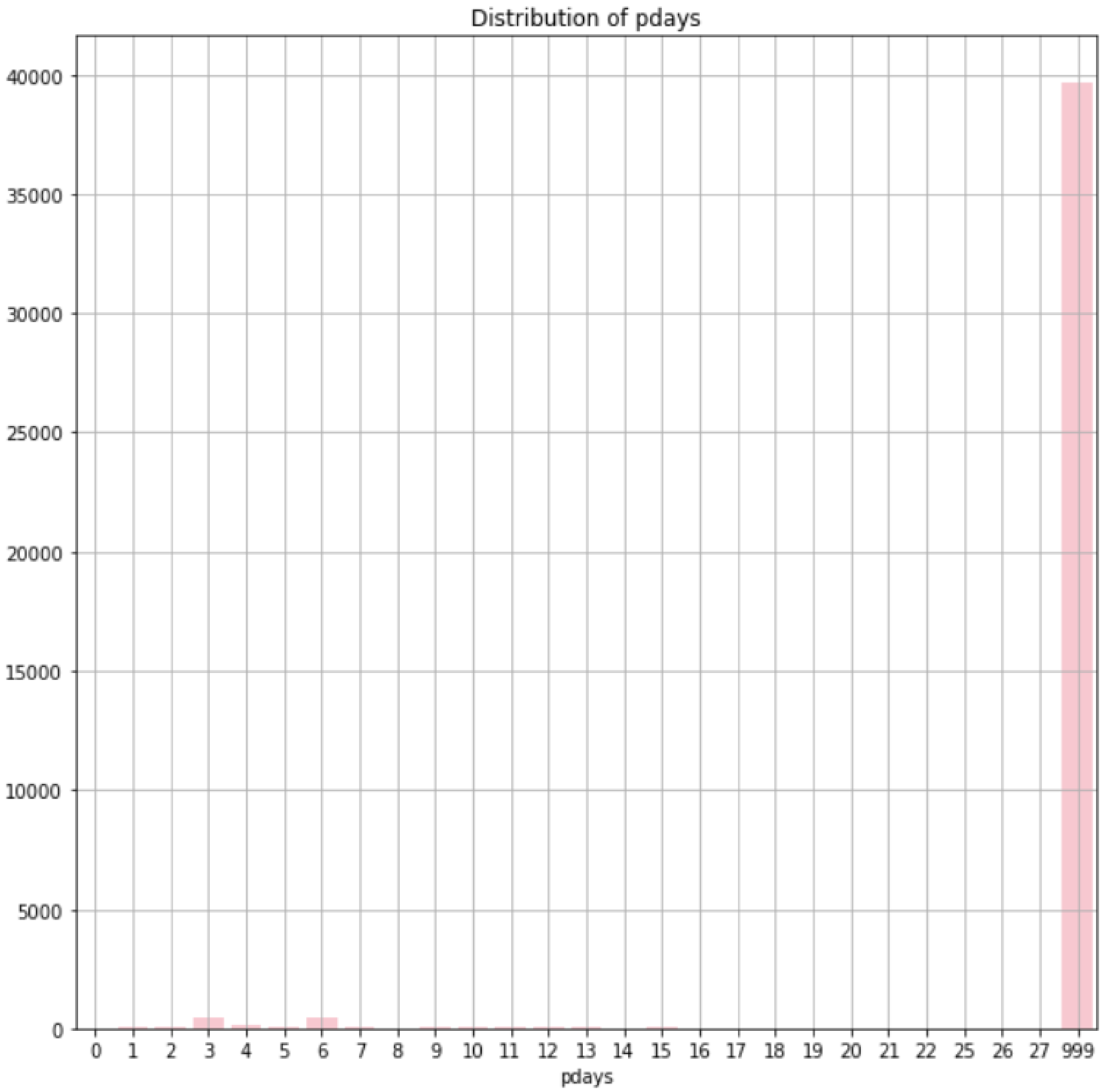
- We can notice that the majority of calls were in Thursday of every week..

Duration of Phone Call



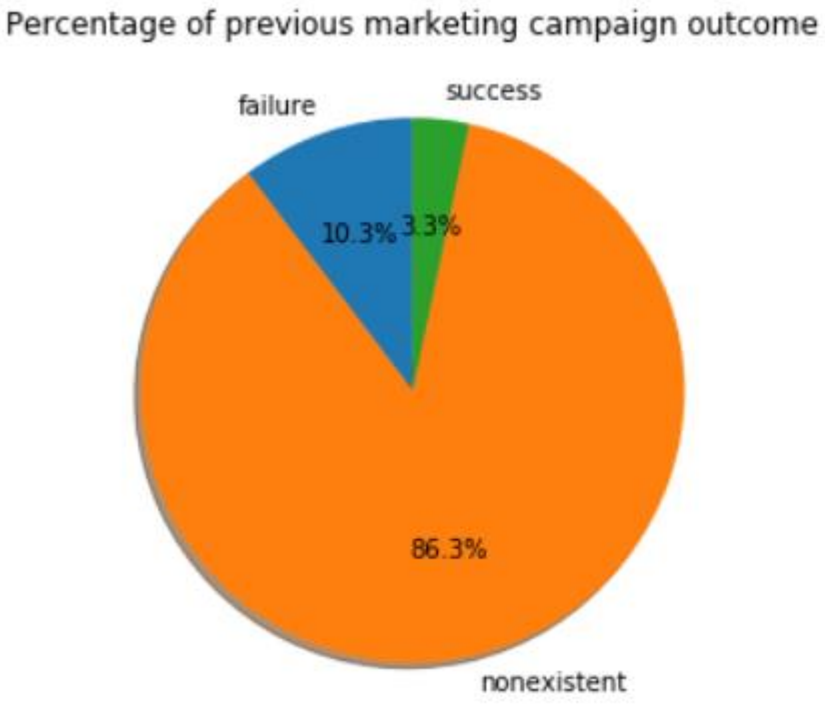
- The majority of calls duration lies in the range between 50 and 800 secs

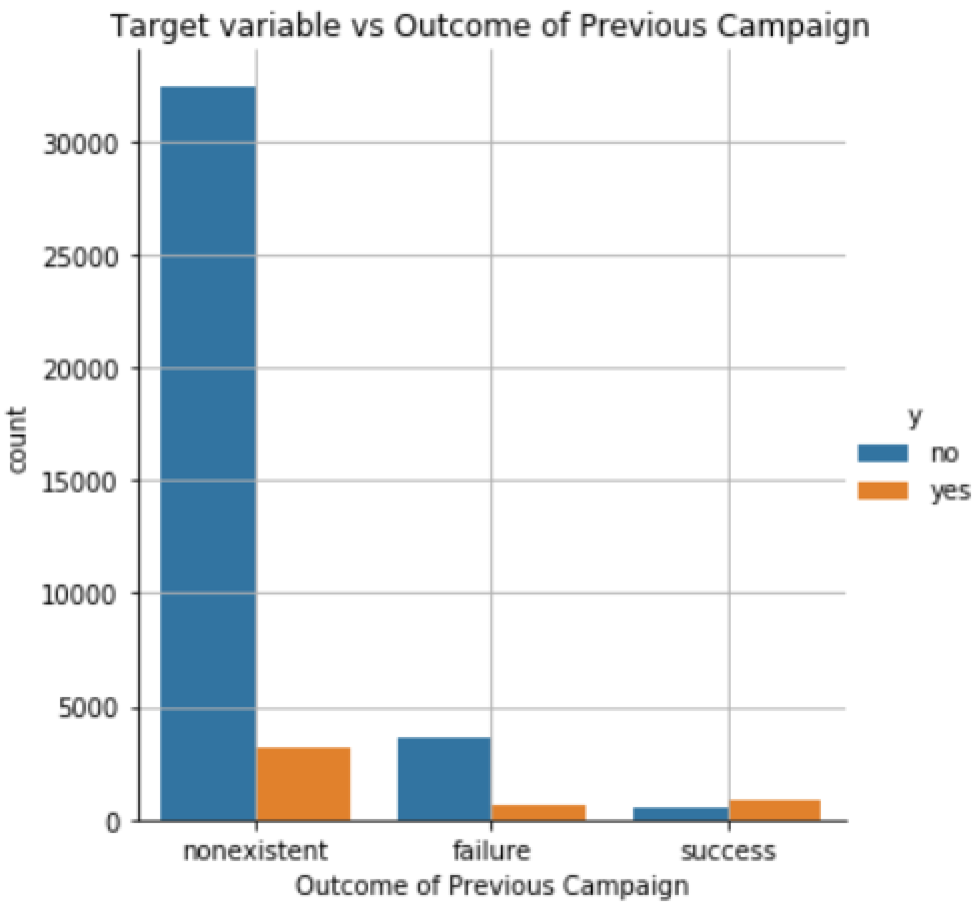
Gap between contacted days



- 999 means they weren’t contacted before
- We can notice that the majority of clients are not previously contacted.
- The data has outliers at 3 days and 6 days.

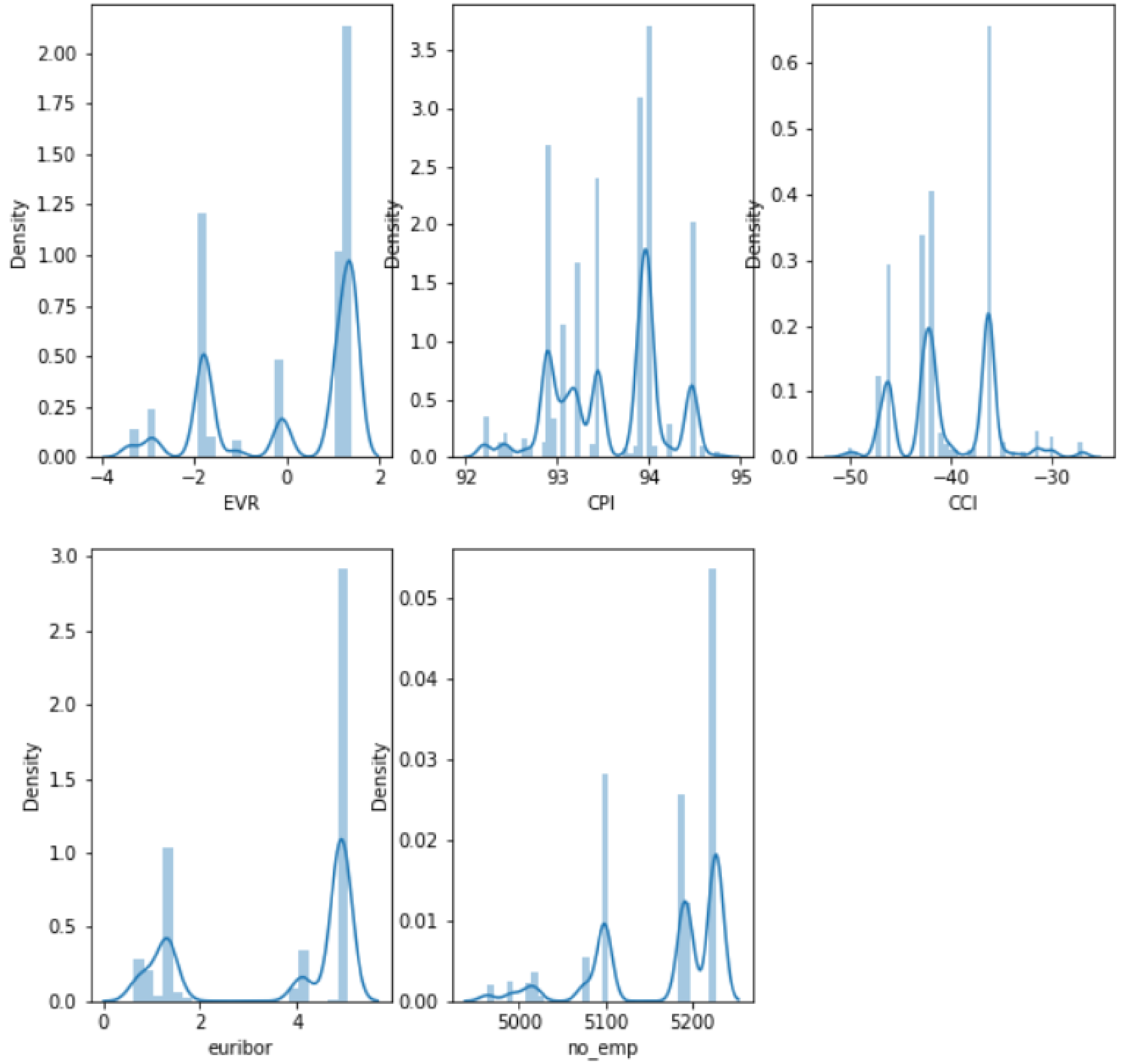
Previous Campaign Outcomes





- Majority of clients open deposit accounts in campaign of non previous campaign outcome.

EVR, CPI, CCI, euribor and no_emp



- We can see there is a high employee variation rate which signifies that they have made the campaign when there were high shifts in job due to conditions of economy
- The Consumer price index is also good which shows the leads where having good price to pay for goods and services may be that could be the reason to stimulate these leads into making a deposit and plant the idea of savings
- Consumer confidence index is pretty low as they don't have much confidence on the fluctuating economy
- The 3 month Euribor interest rate is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of 3 months. In our case the interest rates are high for lending their loans
- The number of employees were also at peak which can increase their income index that could be the reason the campaign targeted the leads who were employed to make a deposit

Model Recommendations

In order to identify the models that would work best with the data at our disposal, we ran cross validation on our data split and used accuracy as the metric to judge the models. We ran the following classification models to identify which would work best:

- Logistic Regression: 0.884
- Decision Tree: 0.640
- K Nearest Neighbours: 0.876
- Support Vector Classifier: 0.918
- Bagging ensemble on K Nearest Neighbour: 0.859
- Random Forest: 0.706
- Naive Bayes:0.850

As we can clearly see from the values presented above, the best algorithm for prediction is the Support Vector Classifier followed by Logistic Regression and K Nearest Neighbours algorithms. Decision Trees and Random Forest algorithms are quite poor performers with our dataset.

Final Recommendation

Here are the few recommendations for the bank than can help improve the deposit rate:

- We can clearly identify duration playing an important attribute in defining the outcome of our dataset. It is absolute that the more the leads are interested in starting a deposit will have higher number of calls and the call duration will be higher than the average.
- We have also figured out that job, education, and marital status also acts as a crucial deciding factor and influences the outcome a lot.
- Classify job roles based on corporate tiers and approach all tier 1 employees within few days after the campaign commences.
- Listen to the leads and extract more information to deliver the best deposit plan, which can increase the duration of calls and that can lead to a deposit.
- Approaching the leads during the start of new bank period (May-July) will be a good choice as many have shown positive results from data history.
- Tune the campaign according to the national econometrics, don't channelize the expenses on campaign when the national economy is performing poor.
- Models such as Support Vector Classifier, Logistic Regression and K Nearest Neighbours seem to be the best algorithms for modelling.

Thank You

Data Glacier
Your Deep Learning Partner