Bank Marketing Effectiveness Prediction

Minal Kharbade, Deveshya Gupta

ML Capstone Project-3

ALMABETTER , BANGLORE

ABSTRACT:

- The data from a marketing campaign was run by a Portugal Banking Institution.
- The campaign aim was to increase customer's subscription rates to fixed term deposit products, such as CDs.
- Using knowledge from the course, a number of ML algorithms are implemented to solve the problem.
- How can the banks successfully market these products in the most efficient way possible and with the highest possible rate of success?

PROBLEM STATEMENT:

- The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.
- 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 ('yes') or not ('no') subscribed.
- Classification goal is to predict if the client will subscribe a term deposit.(variable y)

DATA DESCRIPTON:

- 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')
- Education (categorical: 'primary', 'secondary', 'tertiary', 'unkown')
- 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')
- Contact: contact communication type (categorical: 'cellular', 'telephone')
- Month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- 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')

INTRODUCTION:

- Advertising the product has forced businesses to compete for the attention that has ever growing amount of distractions.
- Thus it raises the question that ,How can businesses successfully advertise their products in the most effective way with rate of success?

- Using collected data from a previous bank , marketing campaign itself, and general market conditions will be explored for the future marketing campaigns.
- Which is the most target public for the campaign and where they need to put more effort.
- Based on this data , machine learning models models will predict which clients will subscribe and what banks can do to increase the rate of subscription

STEPS INVOLVED IN PREDICTION:

1. Loading data and data cleaning:

   The whole project is coded using python 3. Packages/libraries used are numpy for the array manipulation, pandas for dataframe operations , and matplotlib and seaborn for visualization. The sklearn libraries were also critical in providing packages for machine learning algorithm .other data structures such as arrays , list and dictionaries are used as needed.

   The dataset was provide by machine learning repository and contained 45,211 clients across 17 different features both categorical(marital status ,job type, education etc.) and numeric(age, number of days since previous contact etc.). the target variable is binary ”yes” (client subscribed) or “no” (client did not subscribed).

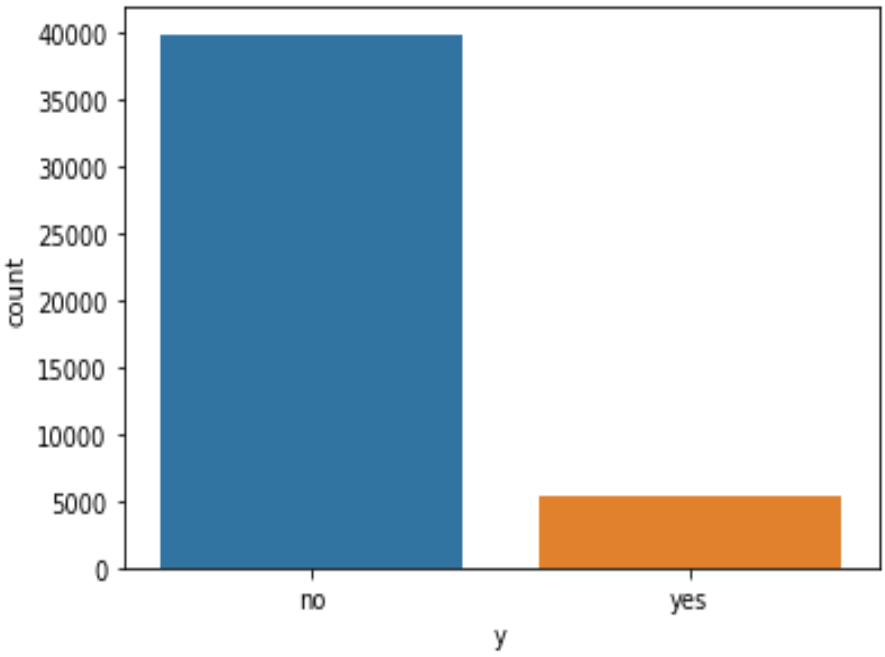
2. Exploratory Data Analysis:

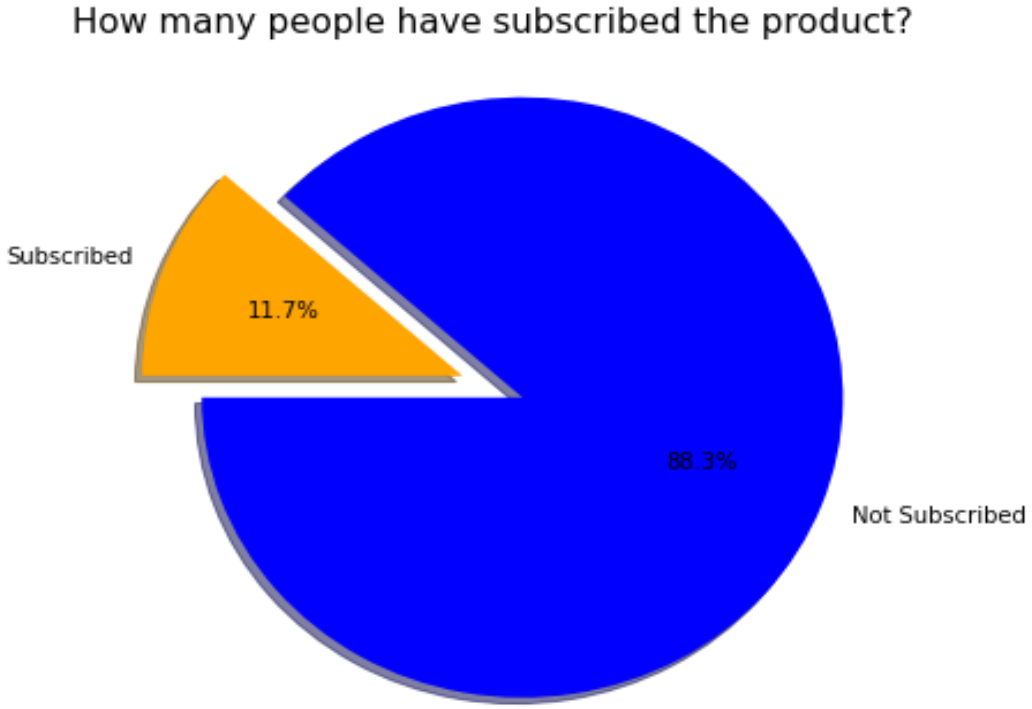
   The next step was to explore the categorical variables such as ‘job type’, ‘marital status’, ‘education’ etc. pots for each were produced that looked their relative frequency as well as normalized relative frequency. These graphs are created using the seaborn packages. For instance cross tabulation between ‘job’ and ‘education’ was used based on the hypothesis that a person’s job will influenced by their education. Thus person’s job is used to predict their education level.

   The python function cross tab was created for this cross-tabulation step. A similar cross tabulation process was carried out for the ‘house owenership’ and ‘loan status’features. Its important to

note that in making these imputations , care was taken to ensure the correlation made sense in the real world.

- Target Variables:

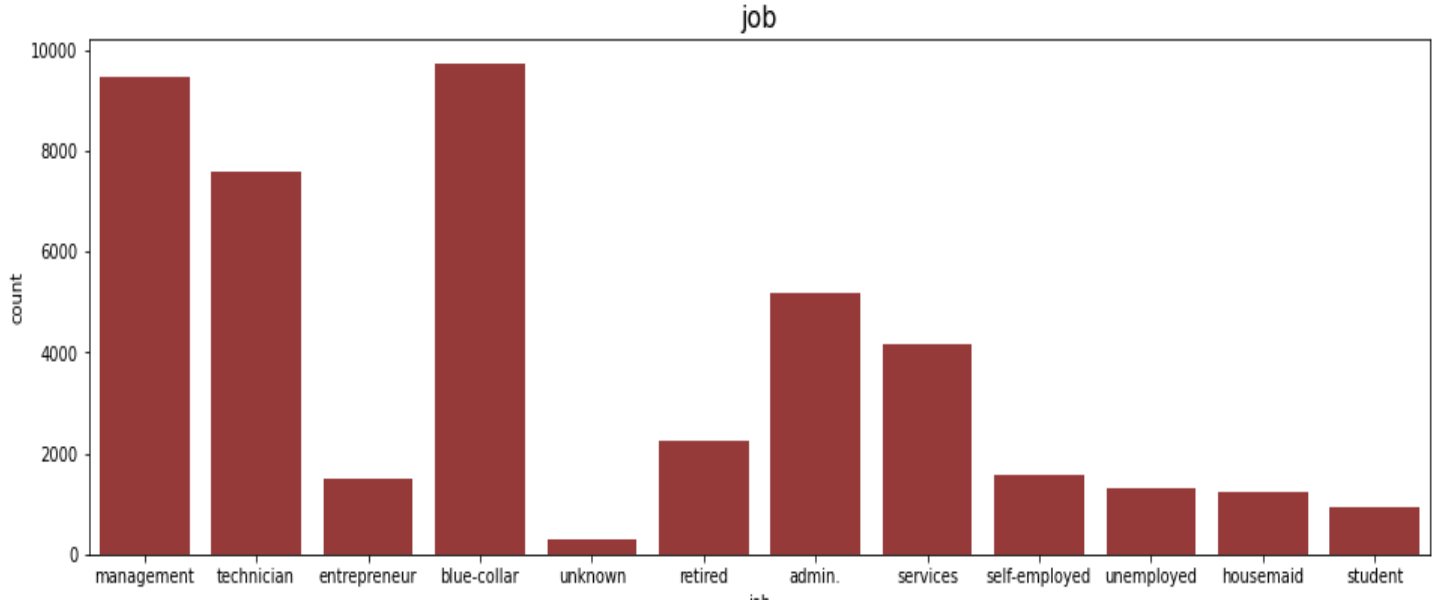


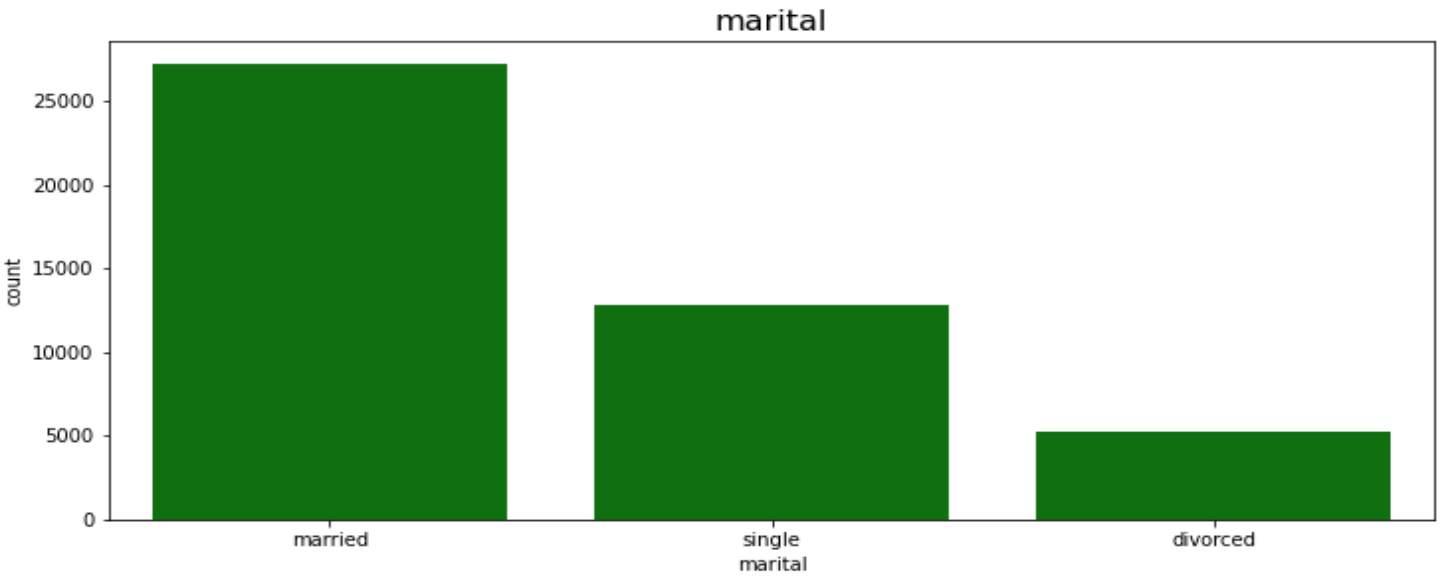


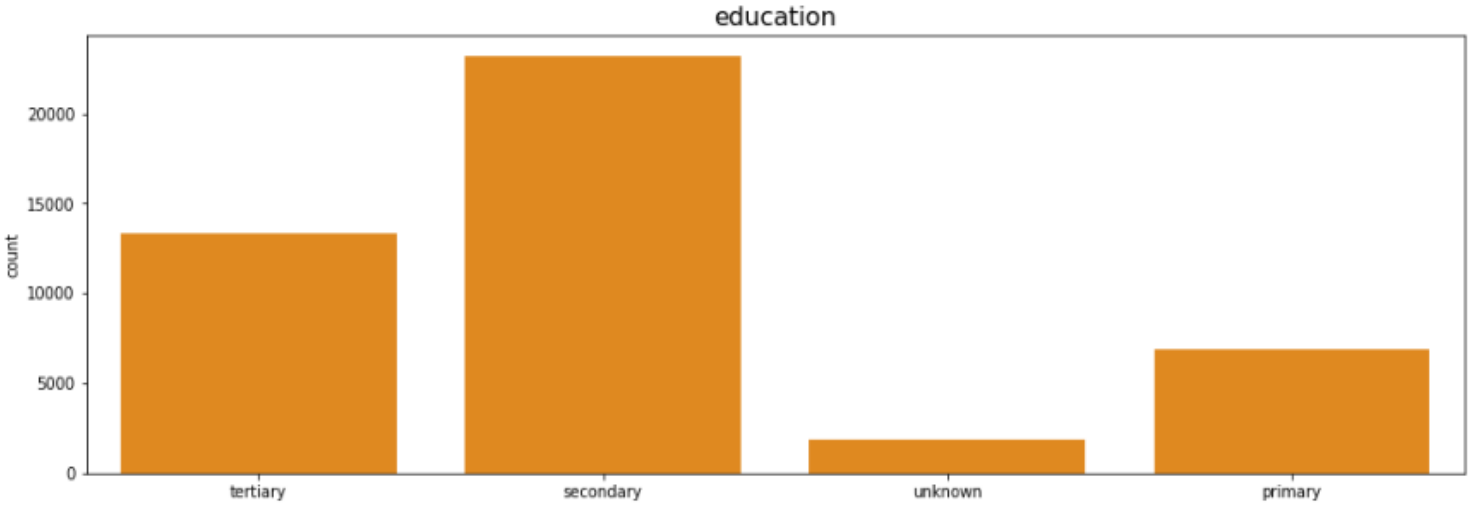
Only 11.7% peoples have subscribed to our product.

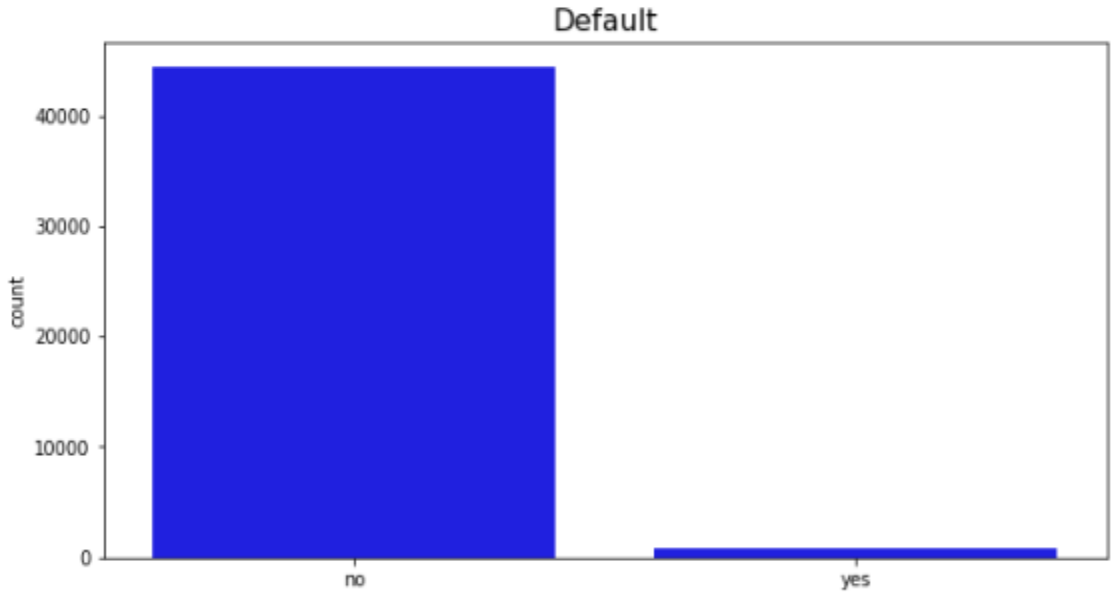
- **Bar Graph representation of each variable:**

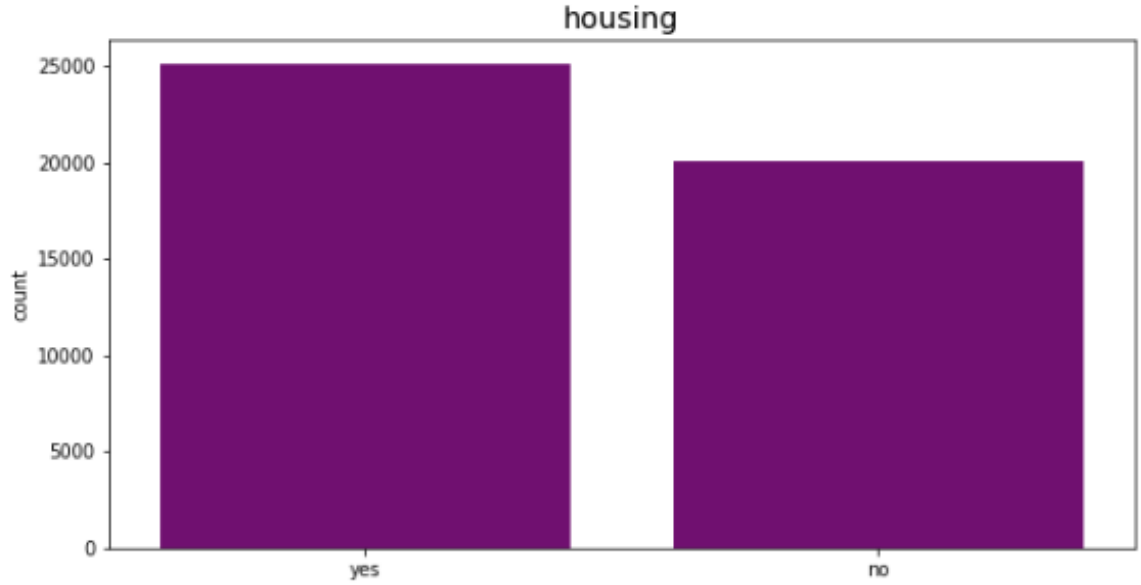
  Categorical Features Explorations

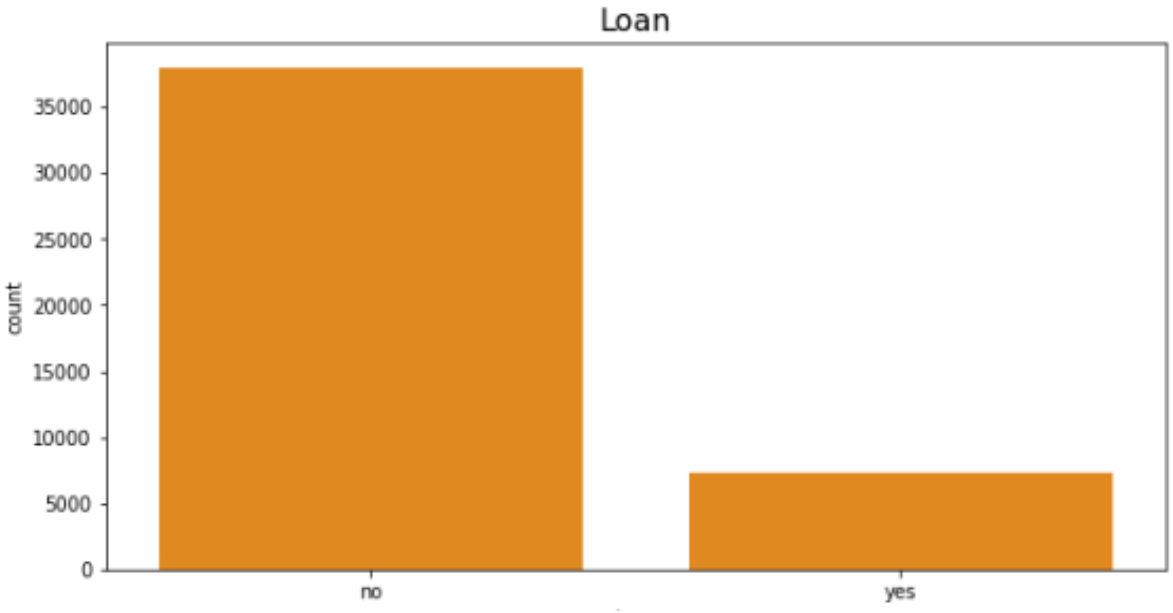


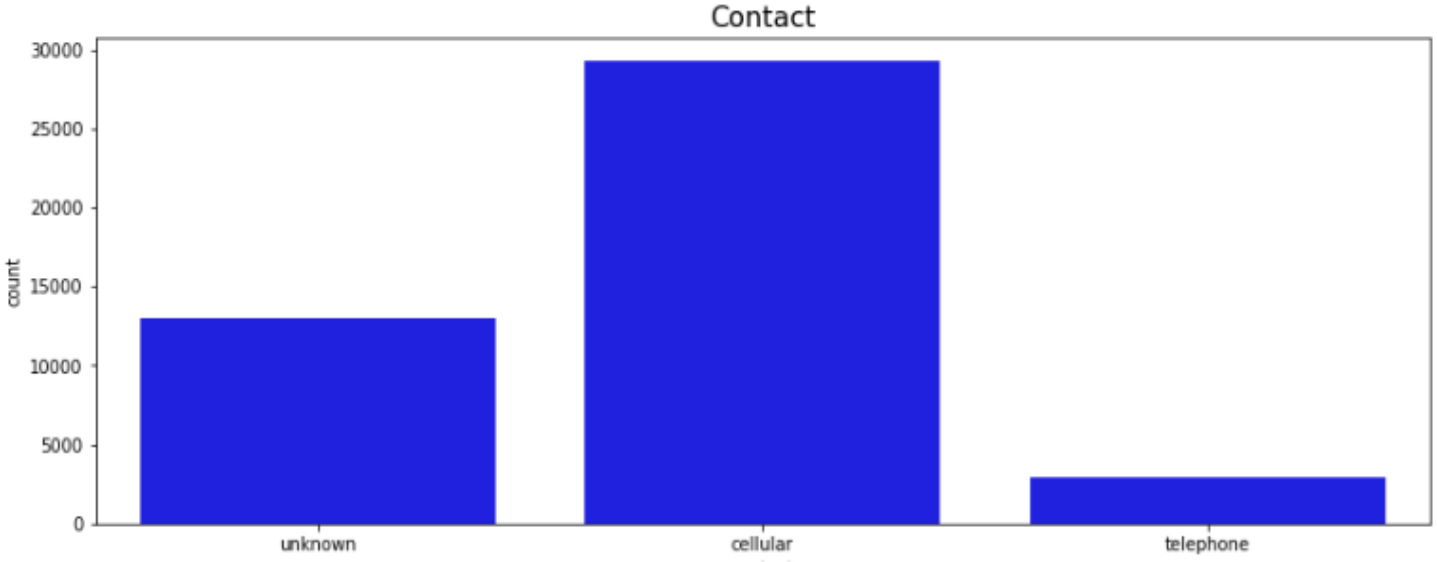


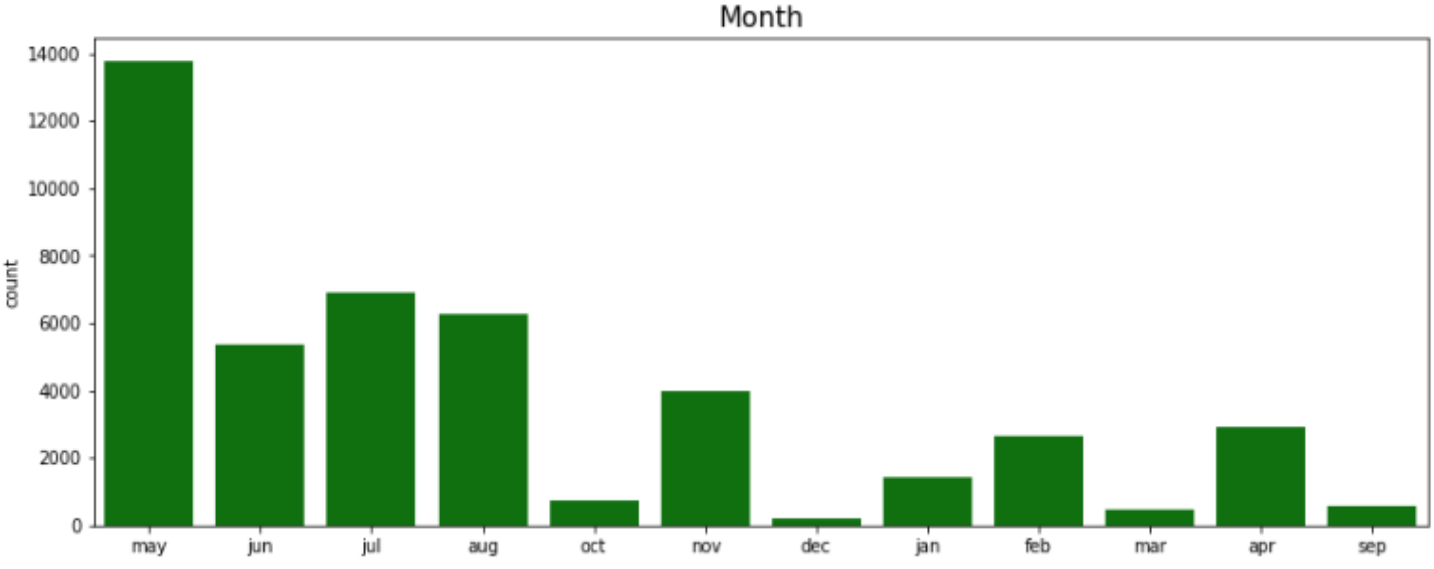

education
count
20000
15000
10000
5000
0
tertiary
secondary
unknown
primary

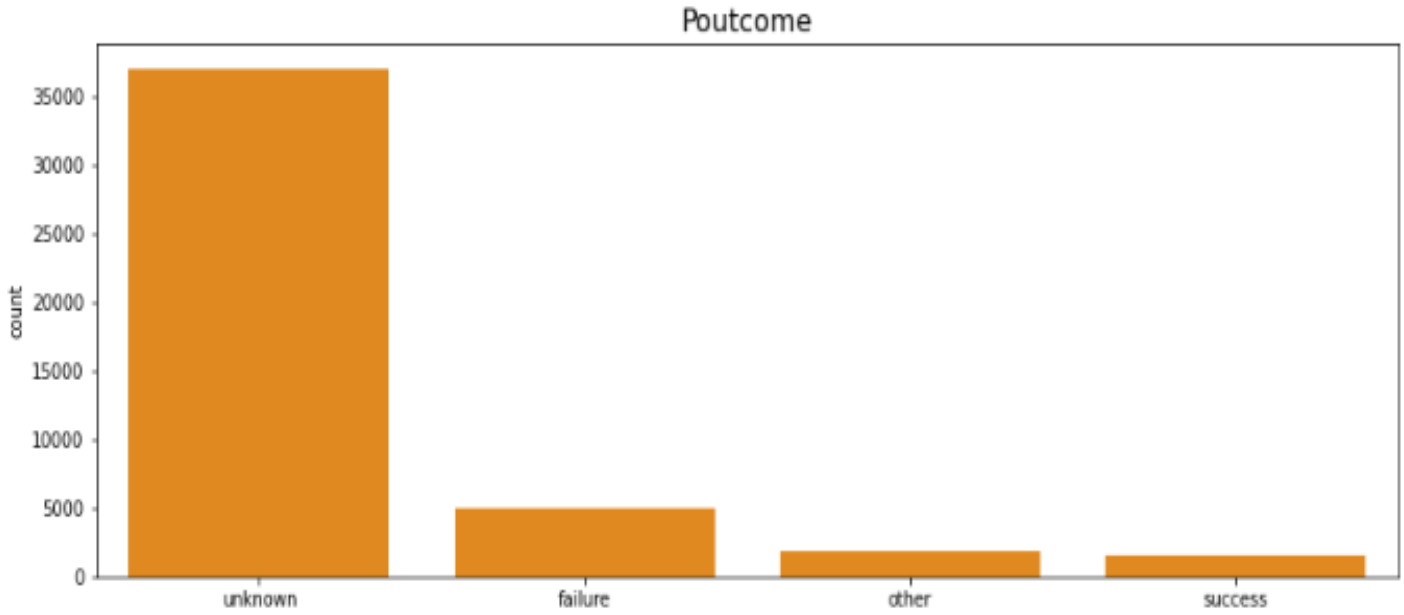

Default
count
40000
30000
20000
10000
0
no
yes

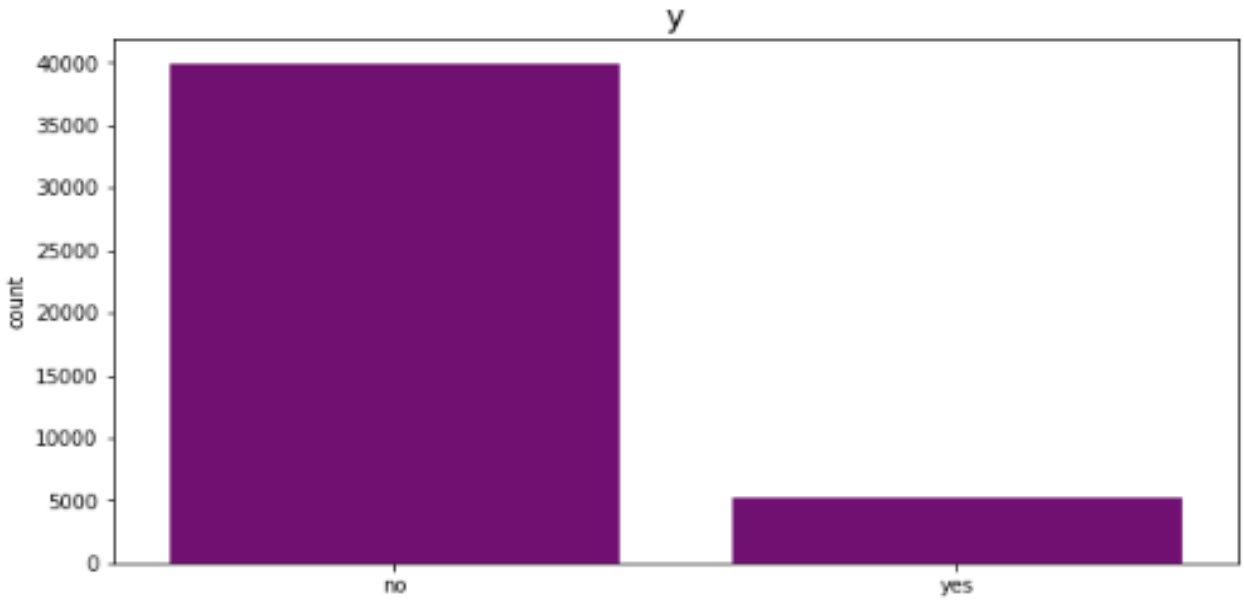

housing
count
25000
20000
15000
10000
5000
0
yes
no


Loan
count
35000
30000
25000
20000
15000
10000
5000
0
no
yes

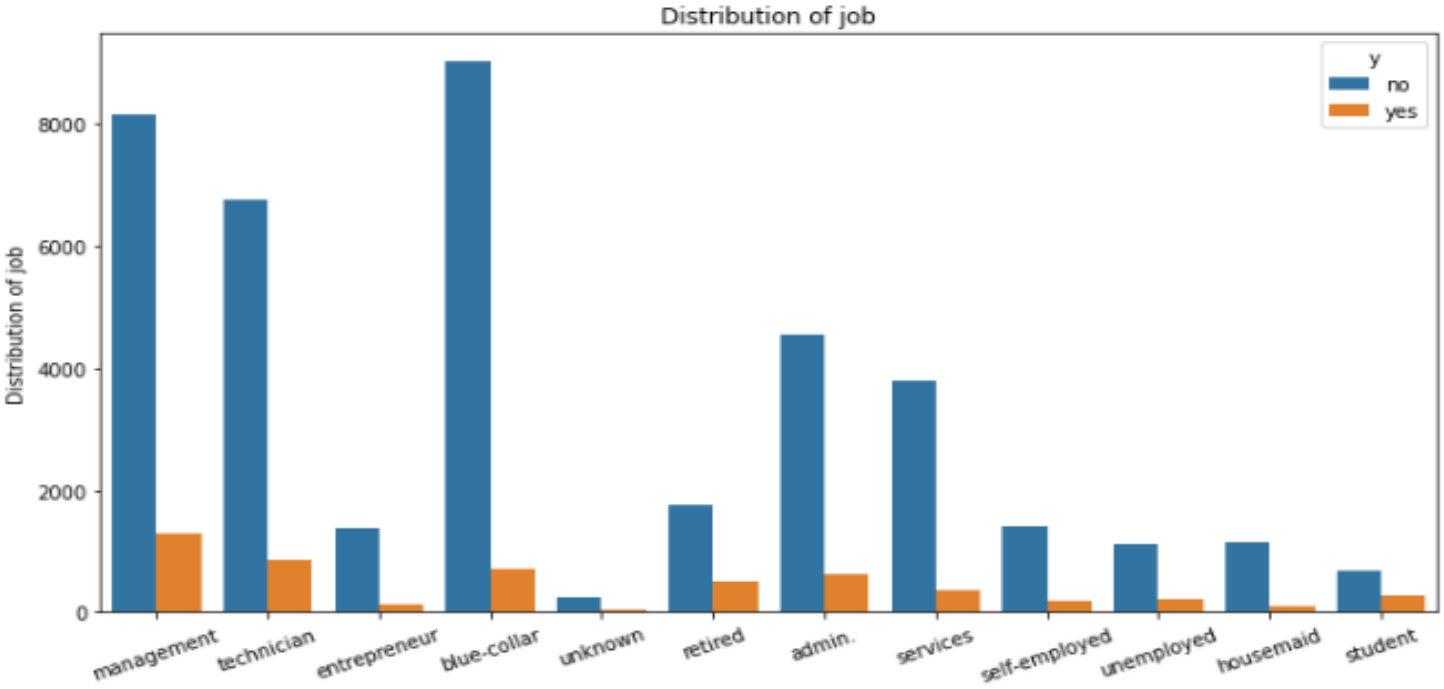

Contact
count
30000
25000
20000
15000
10000
5000
0
unknown
cellular
telephone

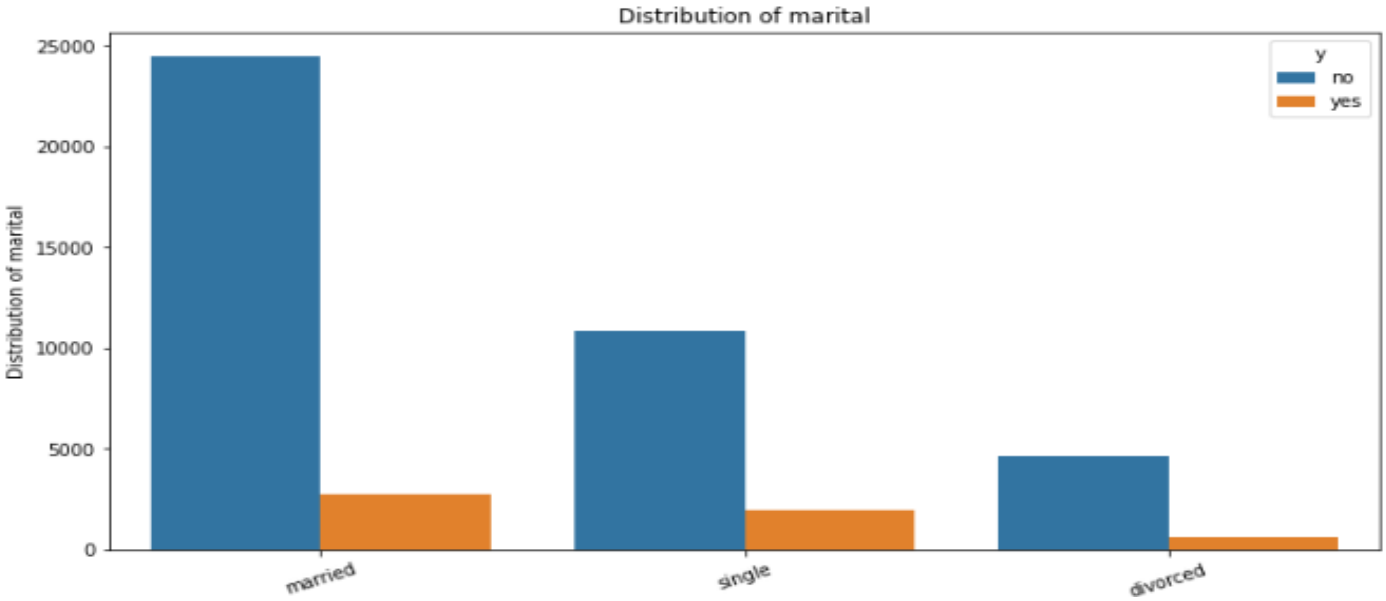

Month
count
14000
12000
10000
8000
6000
4000
2000
0
may
jun
jul
aug
oct
nov
dec
jan
feb
mar
apr
sep

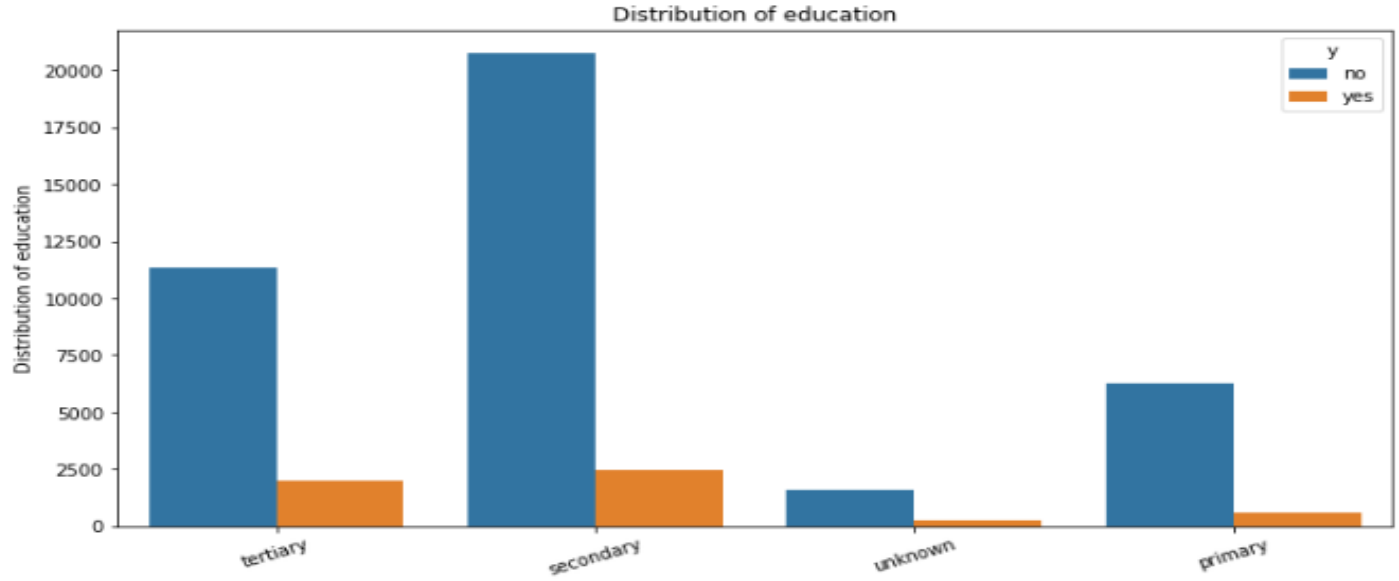


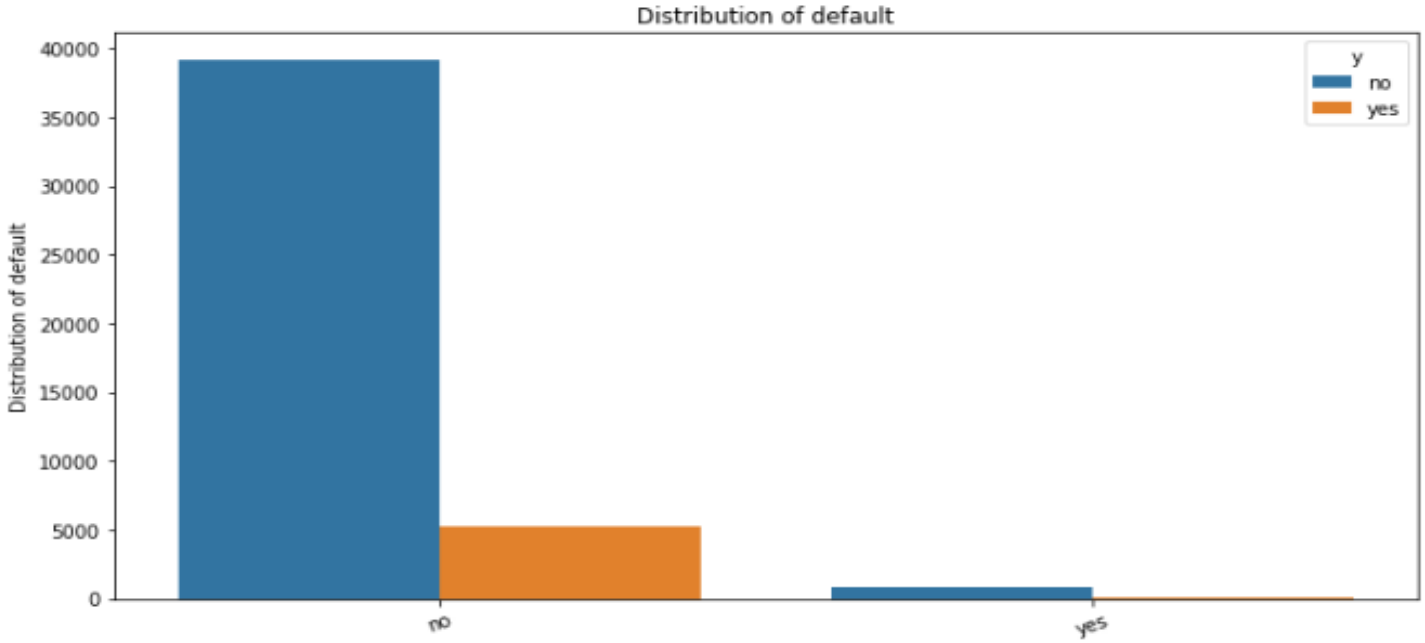


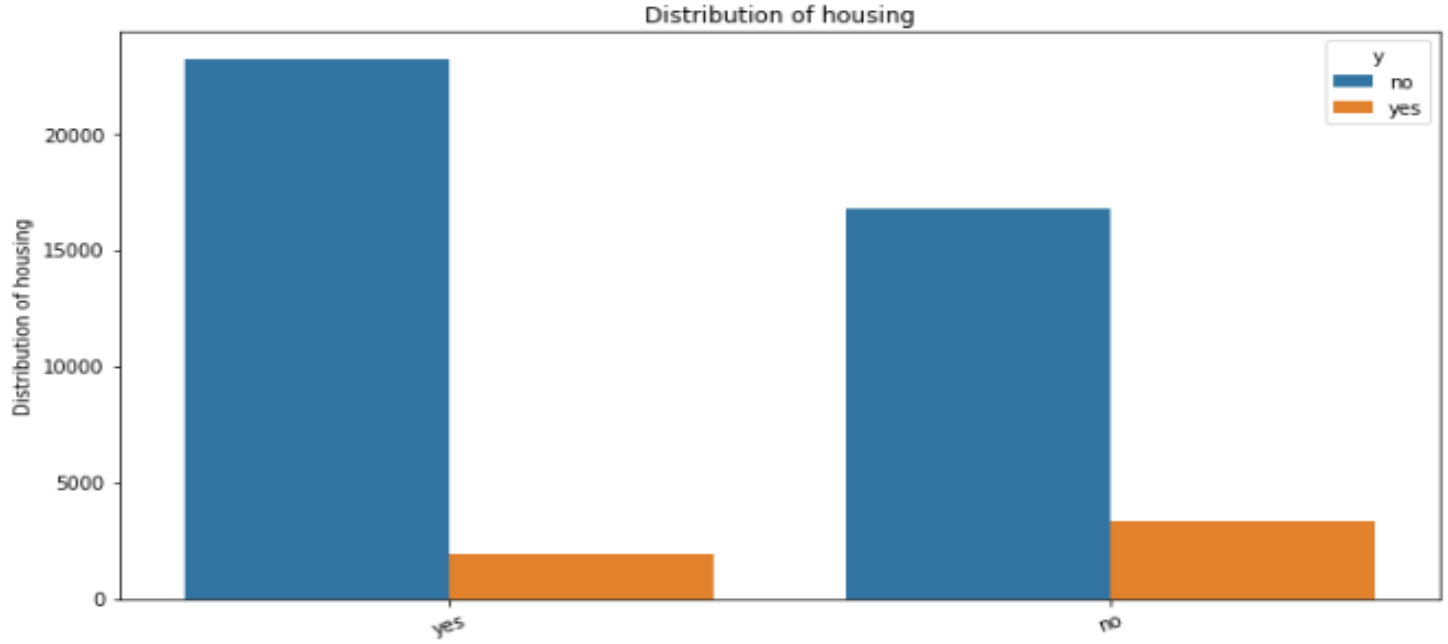
- **Countplot Distribution of Categorical Variables:**

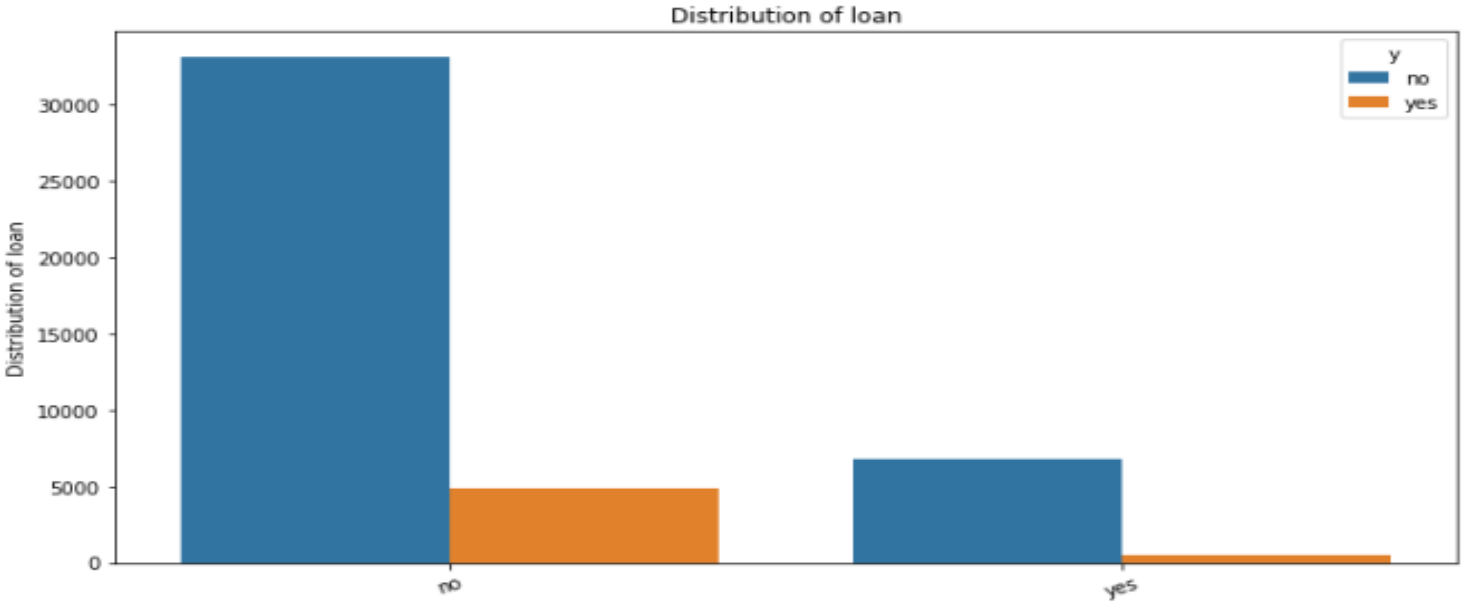

Distribution of job
y
no
yes
Distribution of job
0
2000
4000
6000
8000
management
technician
entrepreneur
blue-collar
unknown
retired
admin.
services
self-employed
unemployed
housemaid
student

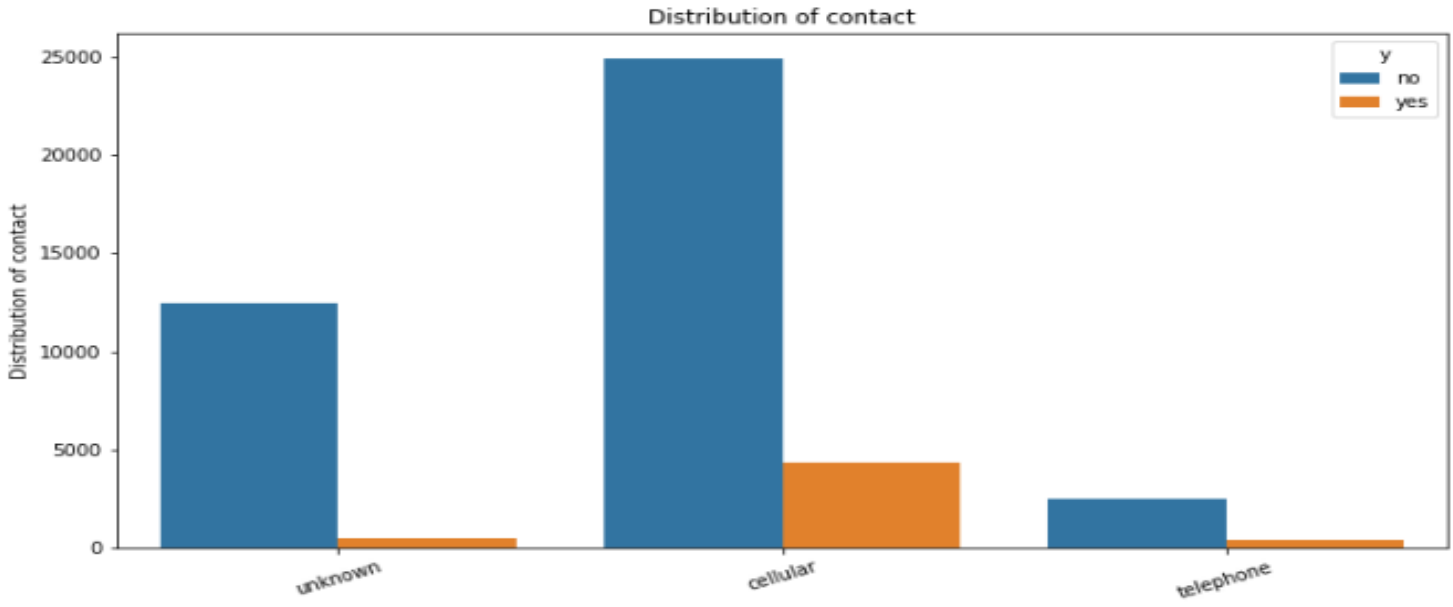

Distribution of marital
y
no
yes
Distribution of marital
0
5000
10000
15000
20000
25000
married
single
divorced

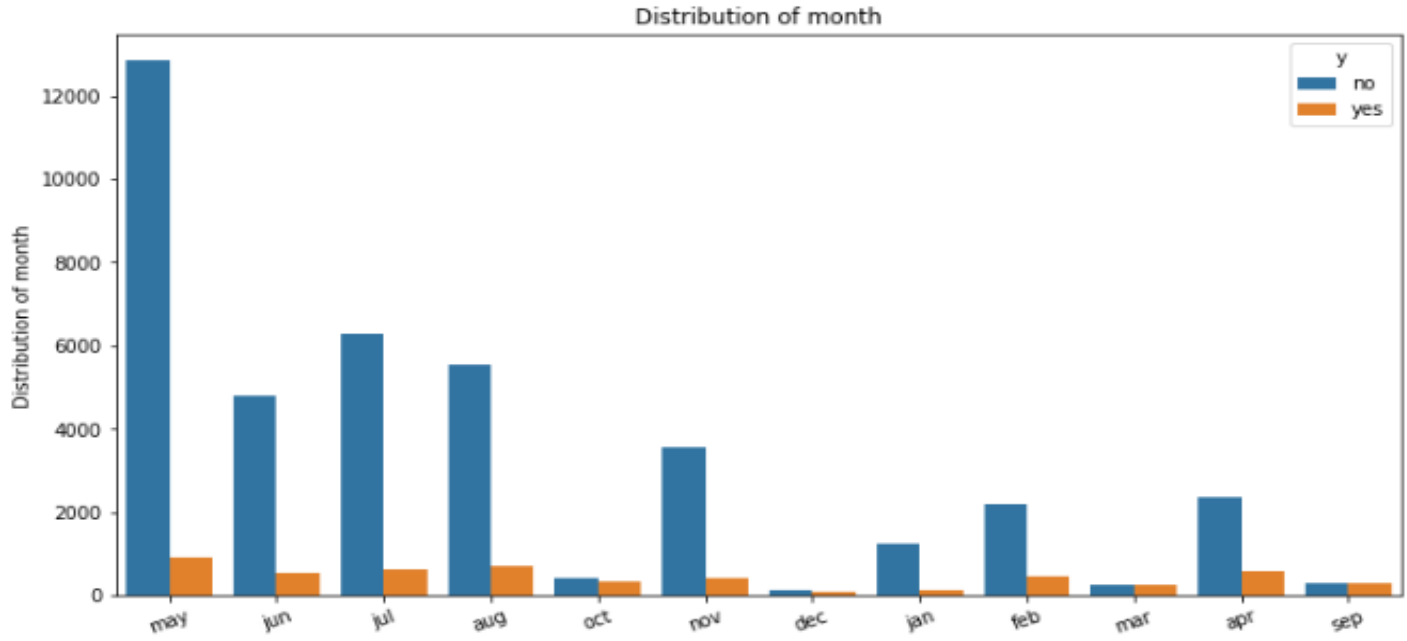

Distribution of education
y
no
yes
Distribution of education
20000
17500
15000
12500
10000
7500
5000
2500
0
tertiary
secondary
unknown
primary

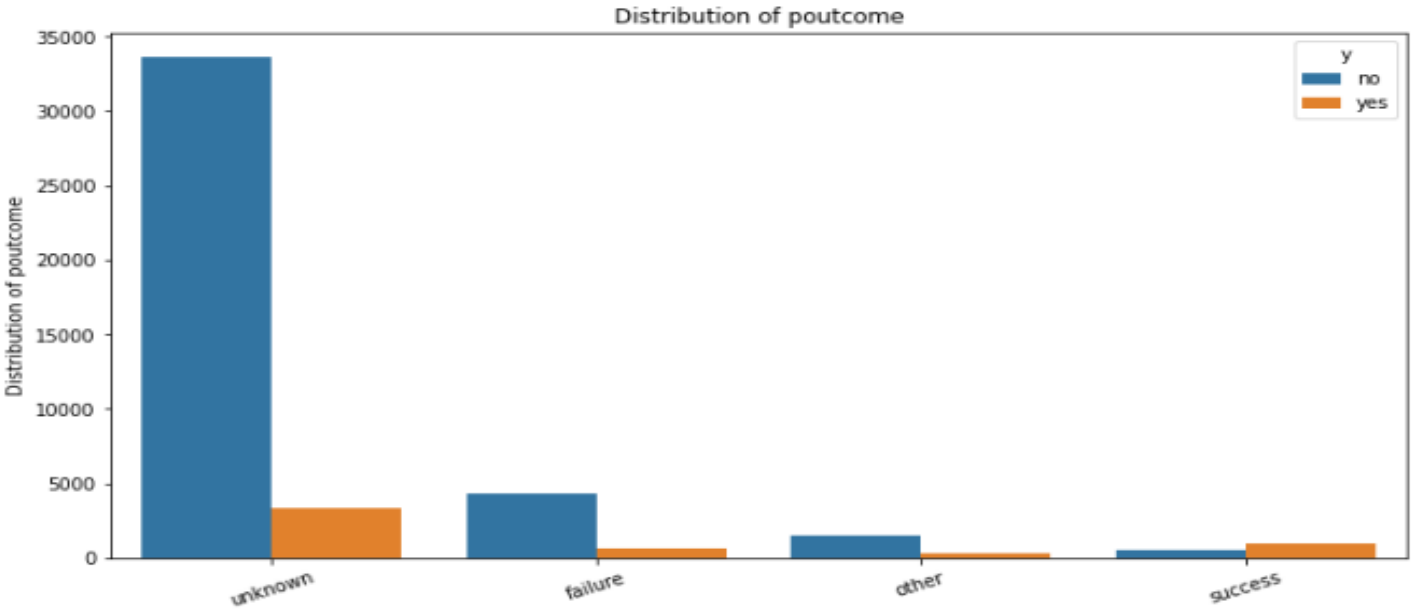

Distribution of default
y
no
yes
Distribution of default
40000
35000
30000
25000
20000
15000
10000
5000
0
no
yes

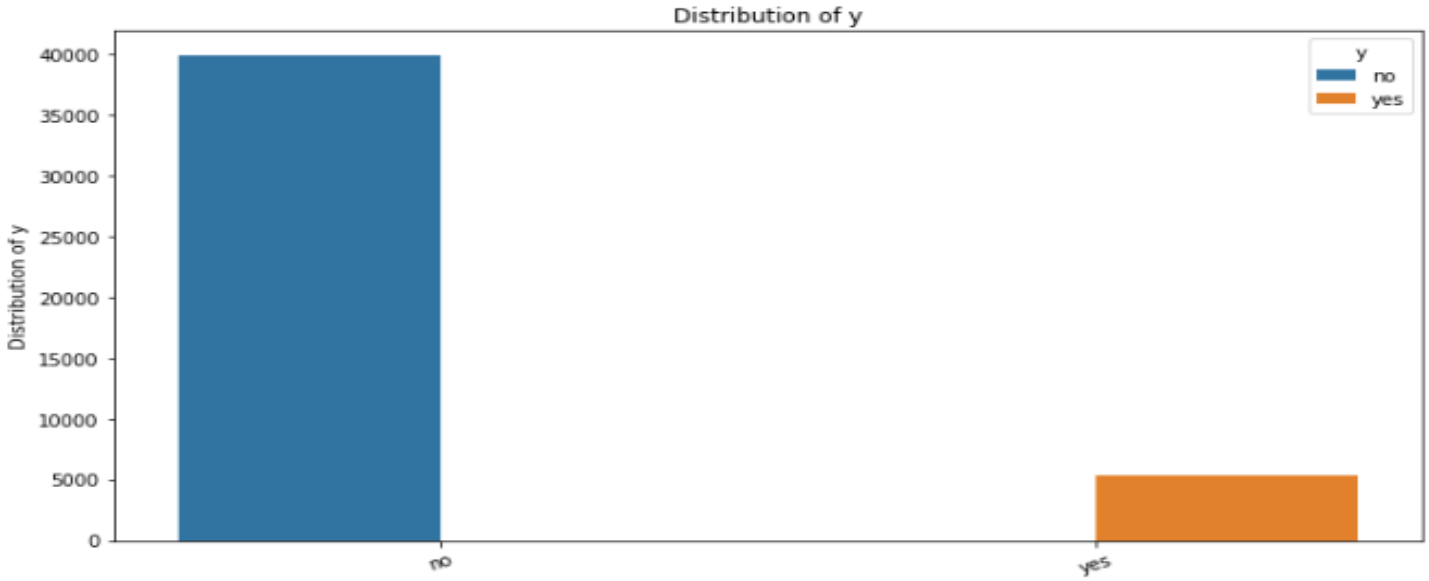

Distribution of housing
y
no
yes
Distribution of housing
20000
15000
10000
5000
0
yes
no


Distribution of loan
y
no
yes
Distribution of loan
30000
25000
20000
15000
10000
5000
0
no
yes


Distribution of contact
y
no
yes
Distribution of contact
25000
20000
15000
10000
5000
0
unknown
cellular
telephone


Distribution of month
y
no
yes
Distribution of month
12000
10000
8000
6000
4000
2000
0
may
jun
jul
aug
oct
nov
dec
jan
feb
mar
apr
sep





- Correlation Analysis:

The heatmap is created using spearman correlation , which measures the degree to which the ranking of each variable align , thus minimizing the effect of outliers. Once this is measured , those variables are expected to be significant during the modeling stage.

3. <u>Model Building:</u>

The dataset is divided into training data and test data with intention of using the training data to find the parameters of the particular model being used and then applying this to the test data to determine the model's performance and to draw conclusion about its predictive capability. This can be done with a sklearn. Cross validation. Train test split function called by specifying split ratio.

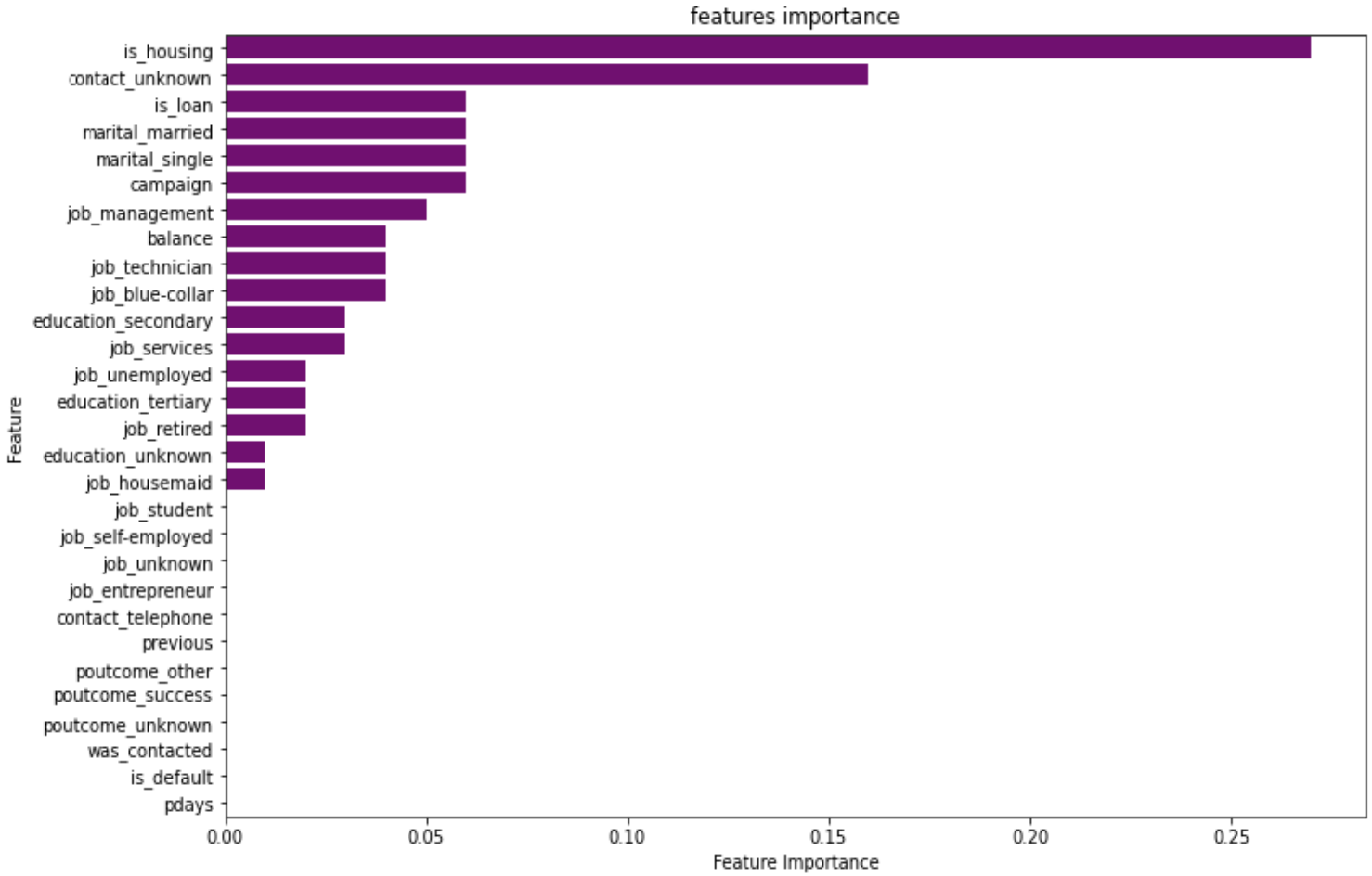
- **<u>Logistic Regression:</u>**

Python provides the package sklearn, linear model , logistic regression for logistic regression. Logistic regression is a"supervised machine learning " algorithm that can be used to model the probability of a certain class or event. It is used when the data i9s linearly separable and the outcome is binary or dichotomous in

nature. That means logistic regression is usually used for binary classification problems.

- **Decision Tree:**

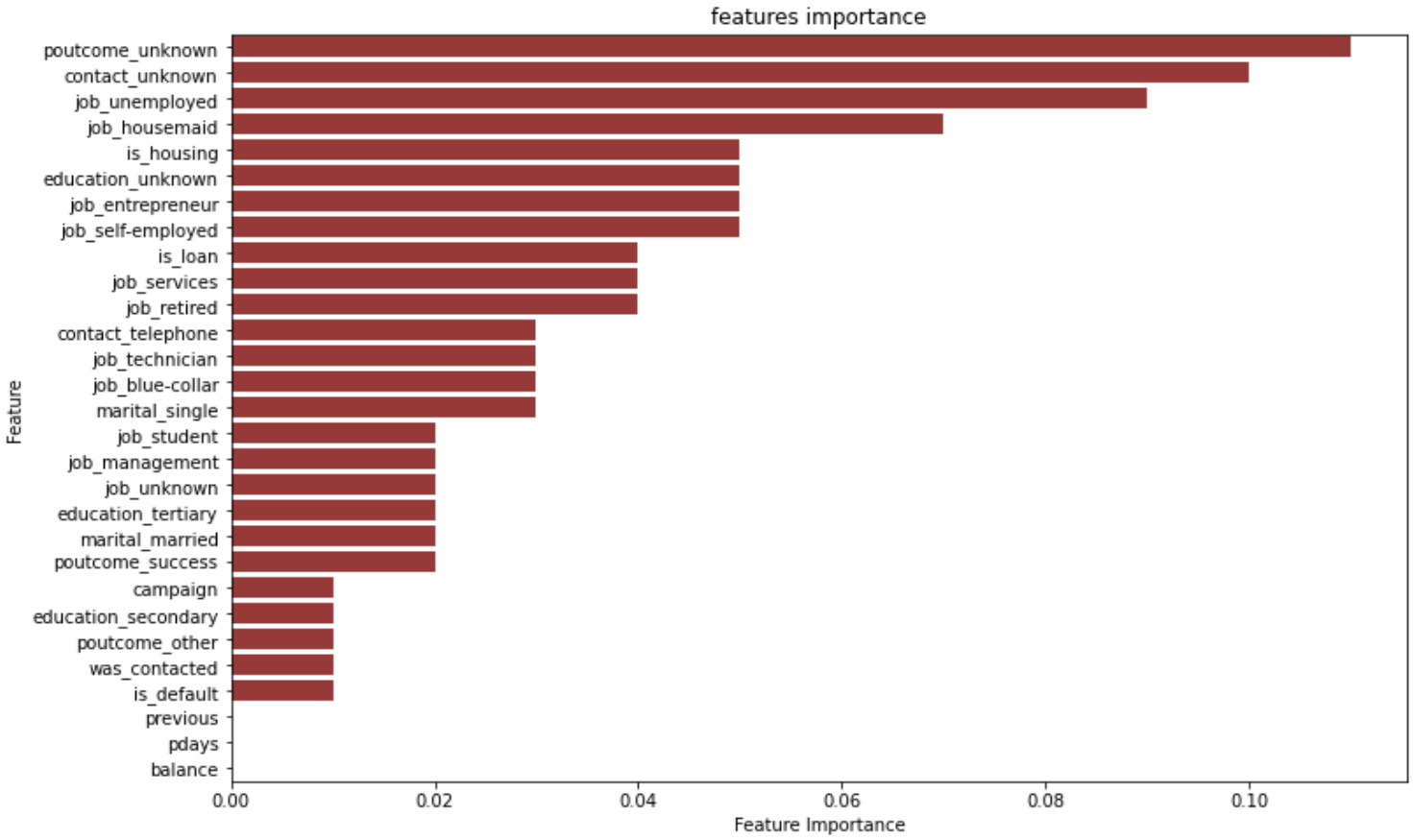
Python provides the package sklearn tree decision tree classifier for the decision tree classifier. Decision trees are a simply yet effective method for classification. Using a tree structure , this algorithm splits the data set based on one feature at every node until all the data in the leaf belongs to the same class.



- **XGBoost classifier:**

XGBoost stands for the “Extreme Gradient Boosting”. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning Algorithms under the Gradient boosting framework. It provides a

parallel tree boosting to solve many data science problems in a fast and accurate way.



● K-Nearest Neighbors(KNN):

K- Nearest Neighbors is a machine learning technique and algorithm that can be used for both regression and classification tasks. K- Nearest Neighbors examines the lables of a chosen number of data points surrounding a target data point, in order to make a prediction about the class that the data point falls into. K-Nearest Neighbors is a conceptually simple yet very powerful algorithm, and for those reasons, it's one of the most popular machine learning algorithms. Let's take a deep dive into KNN algorithm and see exactly how it works.

KNN is a supervised learning algorithm , meaning that the examples in the dataset must have labels assigned to their classes must be known. There are two other important things to know about KNN. First KNN is a non-parametric algorithm. This means no

assumption about the dataset are made when the model is used. Rather, the model is constructed entirely from the provided data. Second there is no splitting of the dataset into training and test sets when using KNN.KNN makes no generalizations between a training and testing set, so all the training data is also used when the model is asked to make predictions.

CONCLUSION:

- In age category , most of the customers are in the age range of 30-40.
- In balance category , above 1000$ is like to subscribe a term deposit.
- They need to more focus in married people because approx. 60% married people are present in subscription list.
- Job profile blue collar are most targeted for this campaign.
- The model can help to classify the customers on the basis on which they deposit or not.
- The model helps to target the right customer rather than wasting time on wrong customer.
- Comparing to all algorithms XGboost algorithm has best accuracy score and ROC-AUC score . So it is concluded as optimal model.