# Project: Bank Marketing (Campaign) Week 10: Deliverables

Name: Supin Hooda

Email: [hoodasupin@gmail.com](mailto:hoodasupin@gmail.com) Country: Canada

Batch Code: LISUM25 Specialization: Data Science Submission Date: 6th Nov 2023

Submitted to: Data Glacier (Individual project)

# Table of Contents

1. Problem Description
2. Data understanding (Type of data, problems, and approaches to solve the problems)
3. Github Repo link

# Problem Description

Background:

ABC Bank is planning to launch a new term deposit product and is looking to create a predictive model to determine whether a customer is likely to subscribe to this product based on their interactions with the bank and other financial institutions. This predictive model aims to assist the bank in optimizing its marketing efforts and improving the effectiveness of its campaigns.

Objective:

The primary objective of this project is to develop a predictive model that can accurately classify customers into two groups: those who are likely to subscribe to the term deposit ("yes") and those who are not likely to subscribe ("no").

Data Source:

The dataset provided for this project contains various customer attributes and information related to the bank's marketing campaigns. These attributes will be used to build and train the predictive model.

# Data Understanding

Types of Data:

The dataset comprises both numerical and categorical variables. Here is a summary of the variables:

1. Numerical Variables:
   * `age`: Customer's age (numeric)
   * `duration`: Last contact duration, in seconds (numeric)
   * `campaign`: Number of contacts performed during this campaign (numeric)
   * `pdays`: Number of days passed after the client was last contacted from a previous campaign (numeric; 999 means the client was not previously contacted)
   * `previous`: Number of contacts performed before this campaign and for this client (numeric)
   * `emp.var.rate`: Employment variation rate - quarterly indicator (numeric)
   * `cons.price.idx`: Consumer price index - monthly indicator (numeric)
   * `cons.conf.idx`: Consumer confidence index - monthly indicator (numeric)
   * `euribor3m`: Euribor 3-month rate - daily indicator (numeric)
   * `nr.employed`: Number of employees - quarterly indicator (numeric)
2. Categorical Variables:
   * `job`: Type of job (categorical)
   * `marital`: Marital status (categorical)
   * `education`: Education level (categorical)
   * `default`: Has credit in default? (categorical)
   * `housing`: Has a housing loan? (categorical)
   * `loan`: Has a personal loan? (categorical)
   * `contact`: Contact communication type (categorical)
   * `month`: Last contact month of the year (categorical)
   * `day\_of\_week`: Last contact day of the week (categorical)
   * `poutcome`: Outcome of the previous marketing campaign (categorical)
   * `y`: The target variable, whether the client subscribed to a term deposit (binary: 'yes' or 'no')

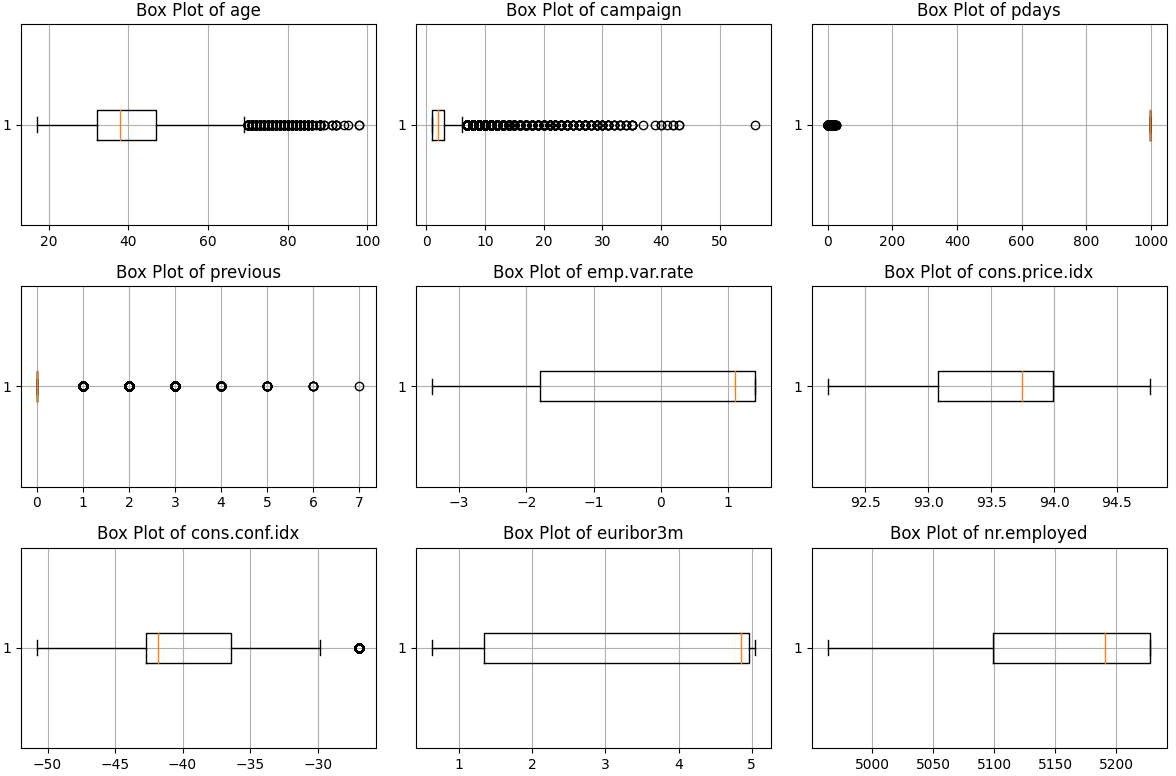
# Problems in the Data

The dataset contains 41,188 rows, with 12 duplicate rows identified and removed using the

`drop\_duplicates` method. Fortunately, there are no missing values in the dataset.

# Outlier Detection

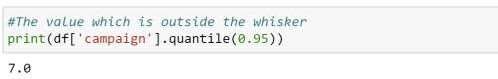
Outliers in the data are detected using box plots. Outliers are data points that deviate significantly from the central distribution of data, impacting key statistics such as mean and mode. Managing outliers during data cleaning is essential to ensure that model performance is not compromised.

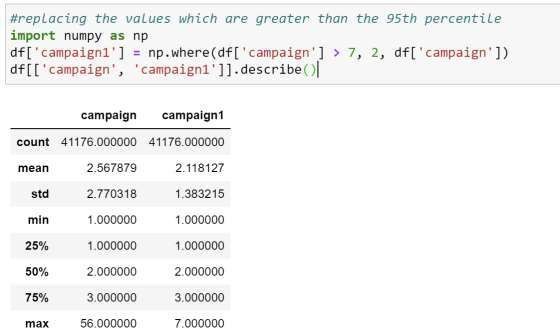
Outliers are observed in the following features: "age," "campaign," "pdays," and "previous." These outliers are identified as data points located outside the box plot whiskers.

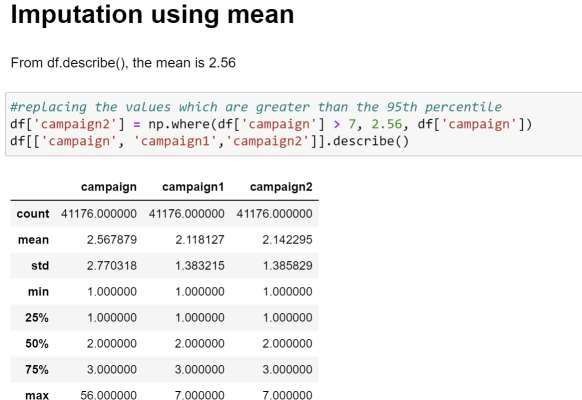
# Approaches to Overcome the Problems

* + - **Age**: The maximum age value of 98 appears to be realistic and is retained in the dataset.
    - **Pdays**: The maximum value of 999 indicates that the client was not previously contacted, and it's present in around 96% of rows. Thus, it's considered unrealistic to drop rows with this value, and they are retained.
    - **Campaign**: The "campaign" feature, which denotes the number of contacts made during the campaign, contains a maximum value of 56, which is considered noise. The portion of records with "campaign" values exceeding 20 is approximately 0.38%. It is proposed to impute these rows with the average of campaign values.
    - **Previous**: The "previous" feature represents the number of contacts made before this campaign. The maximum value of 7 does not appear to be an outlier, and it is retained.

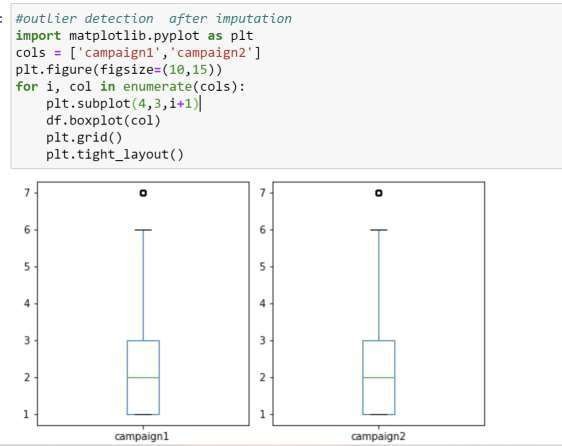
# Treating Outliers

1. **Treating outliers**



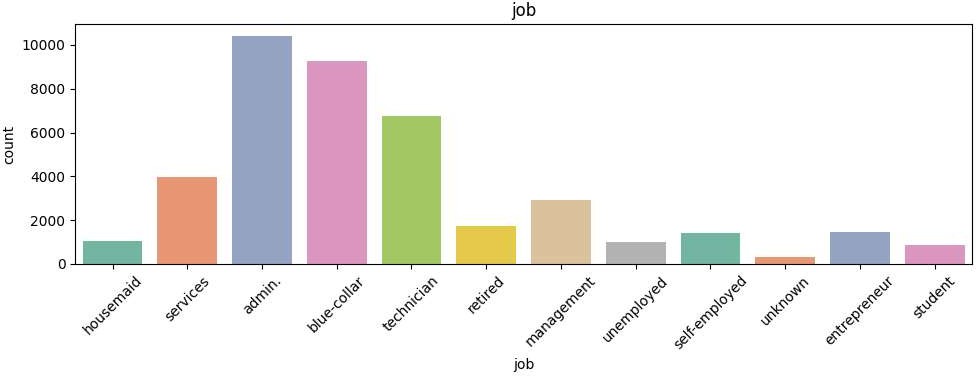


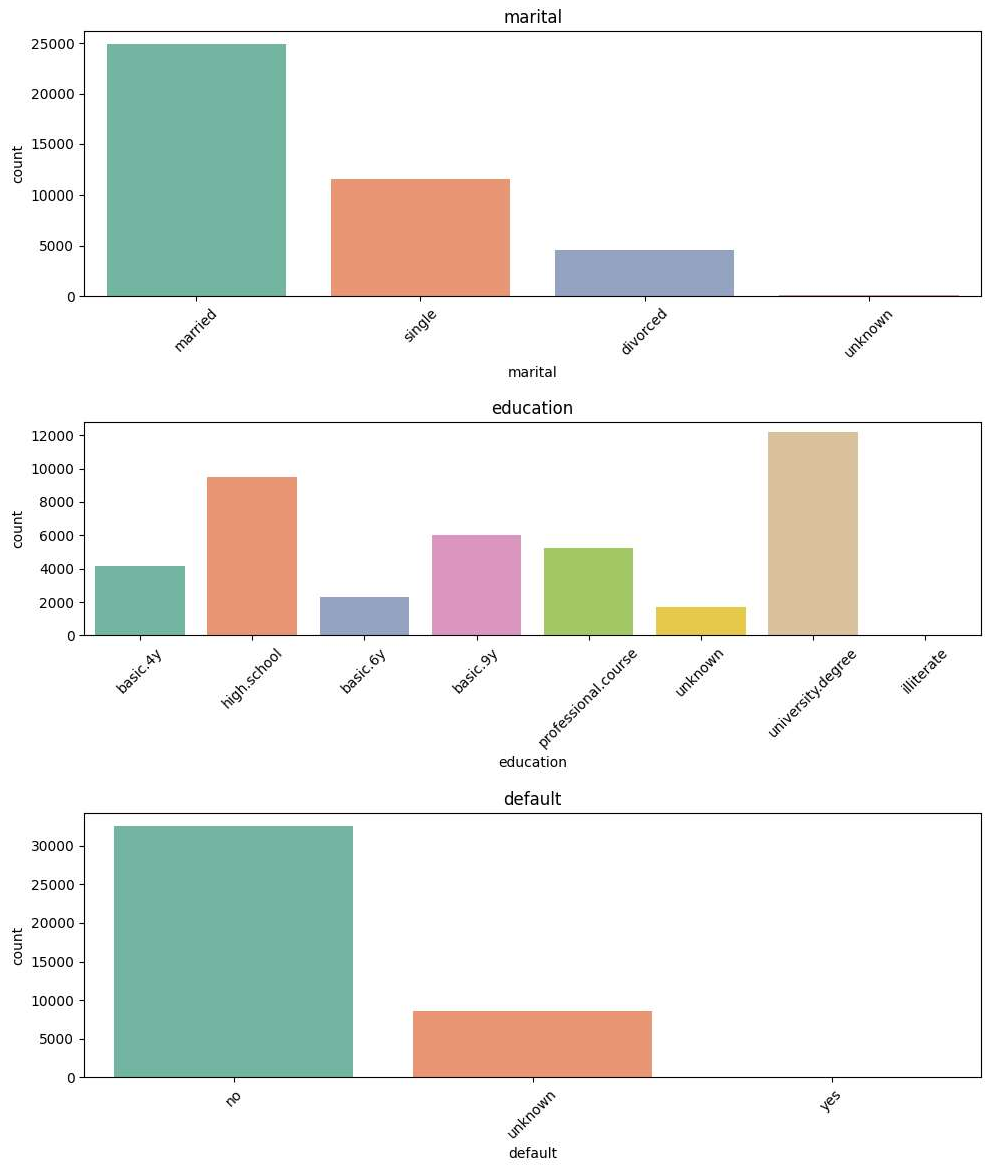
Statistics of the dataset after both median and mean imputation remainsmore or less the same.

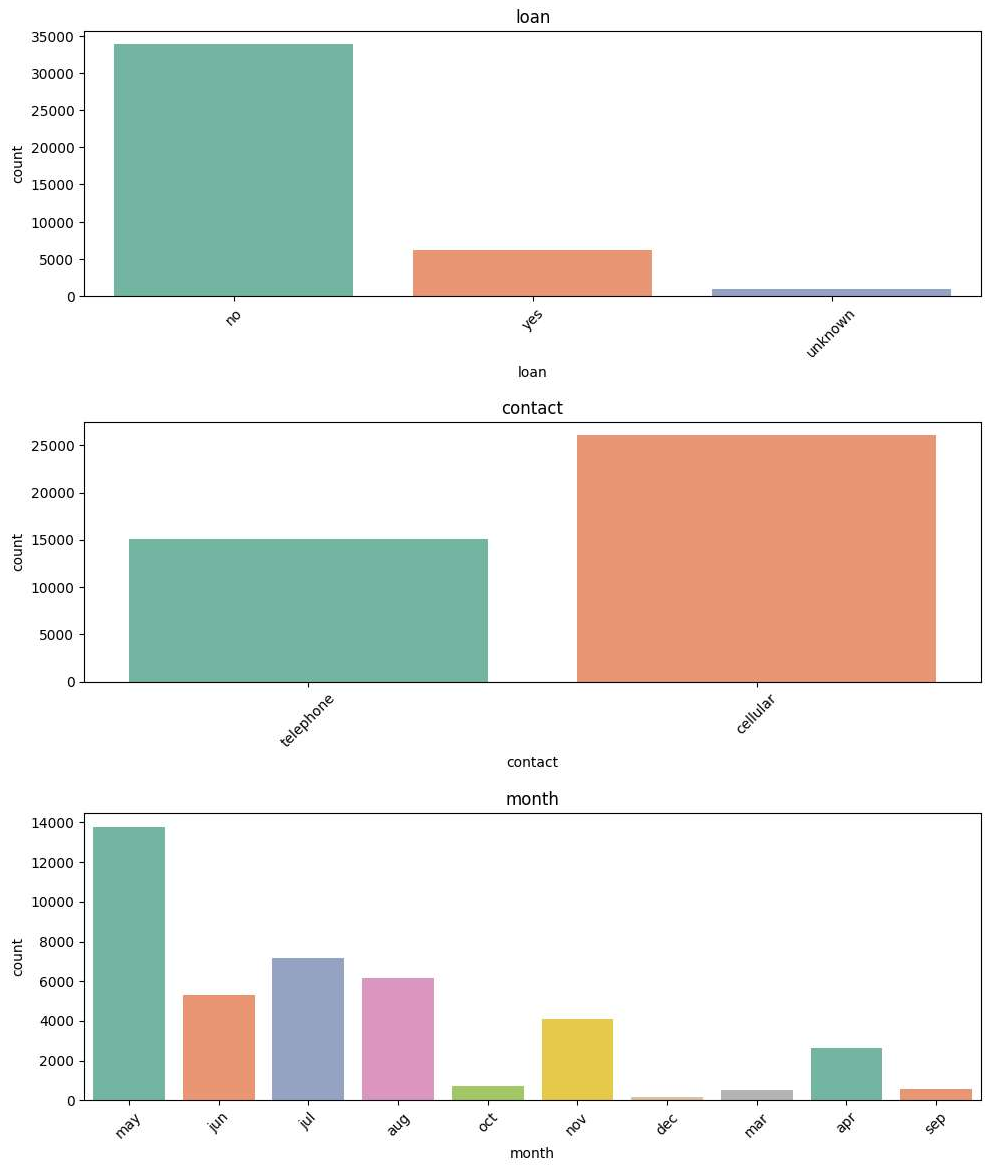


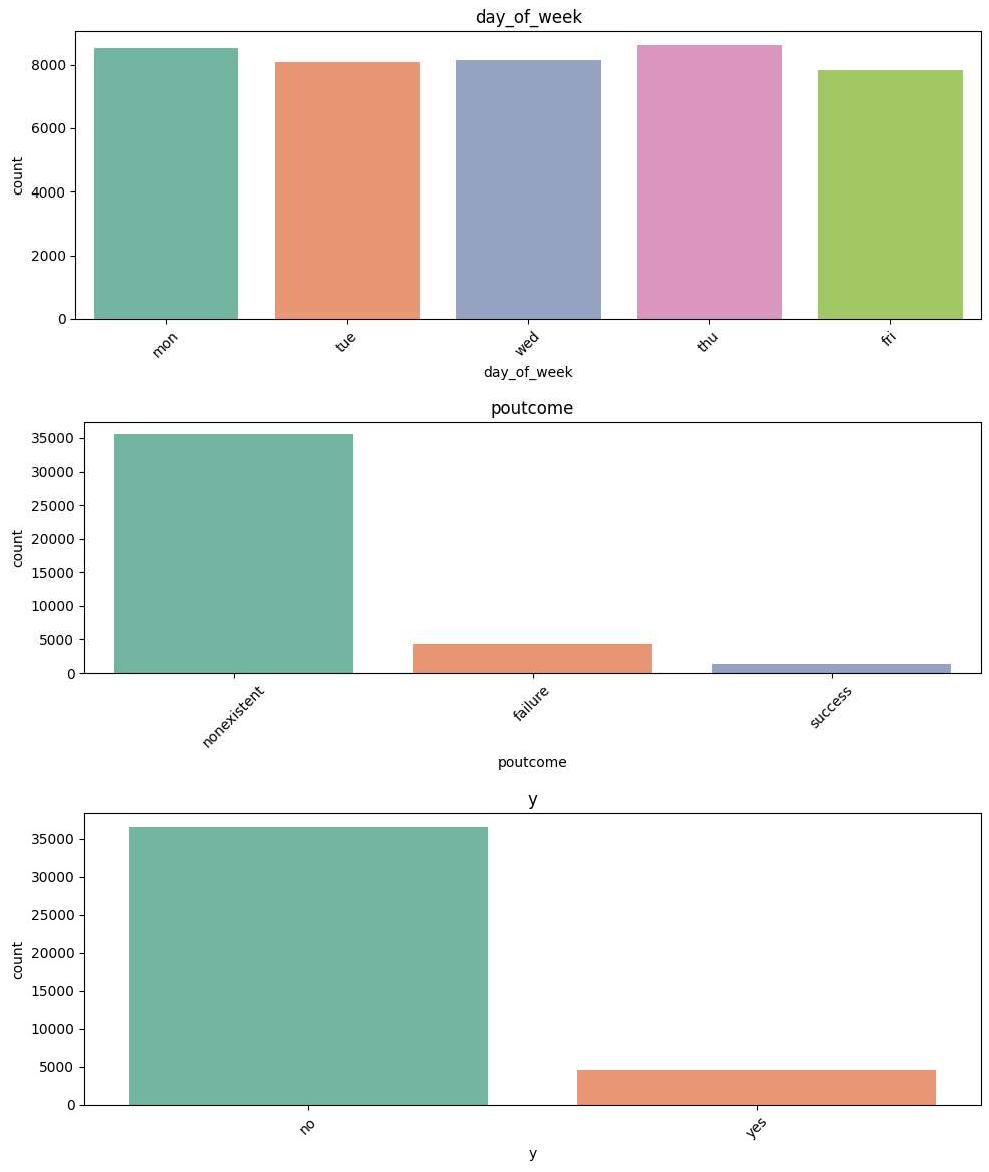
# EDA and recommendation

 Exploring categorical values

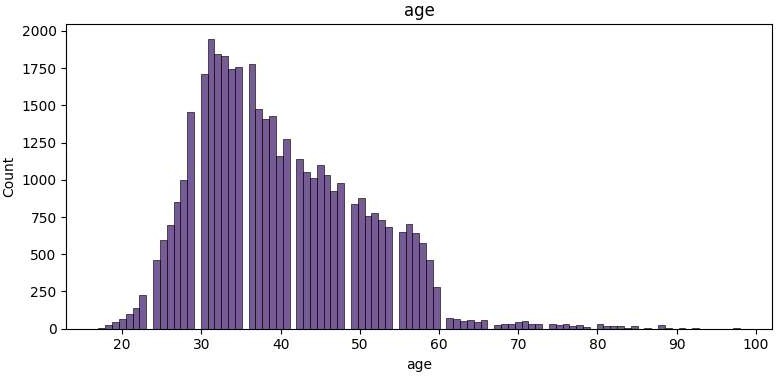


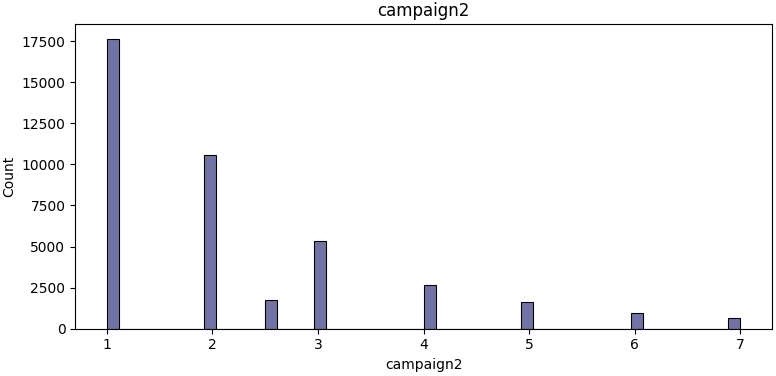


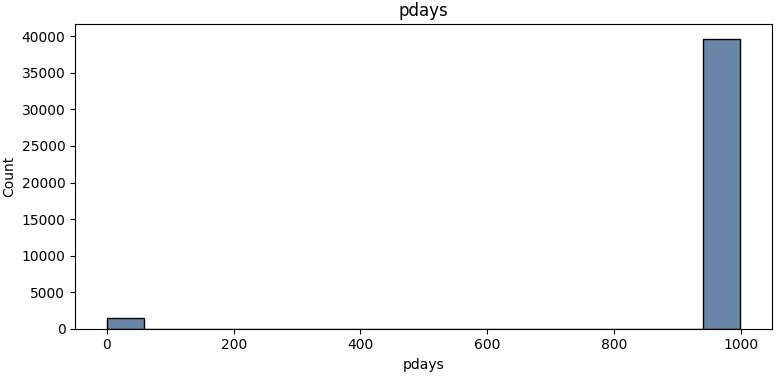


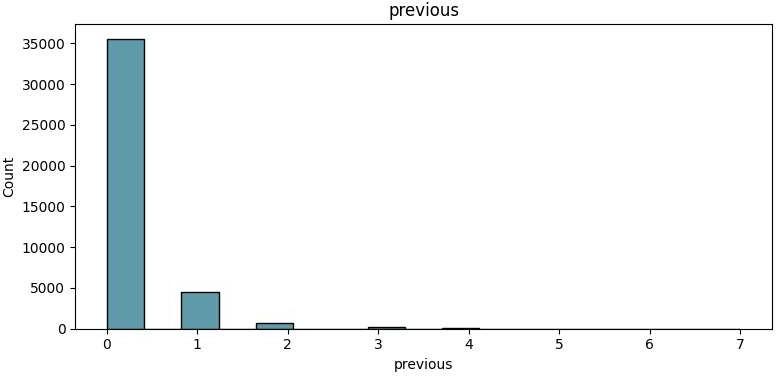


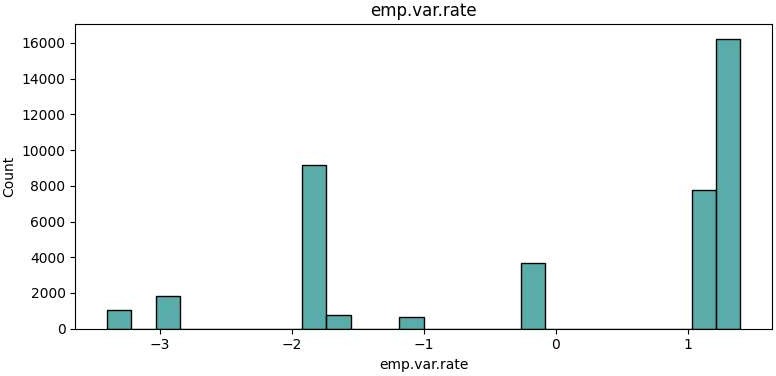
 Exploring numerical features

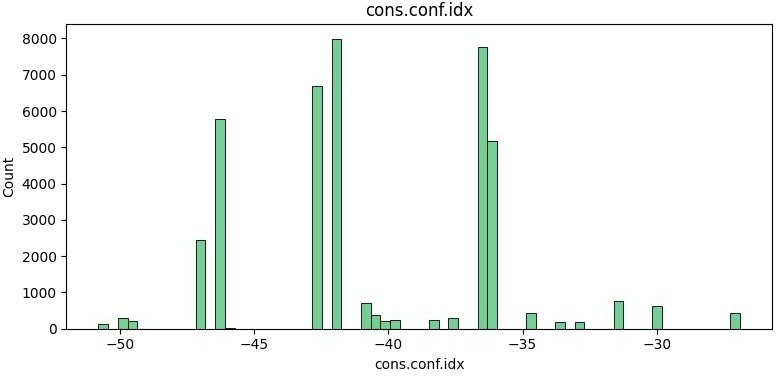
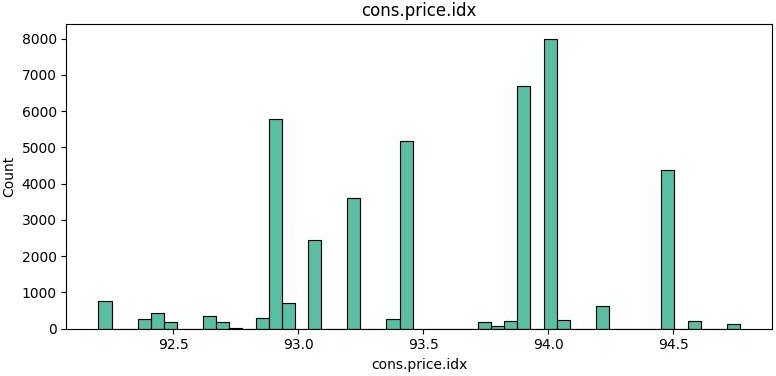


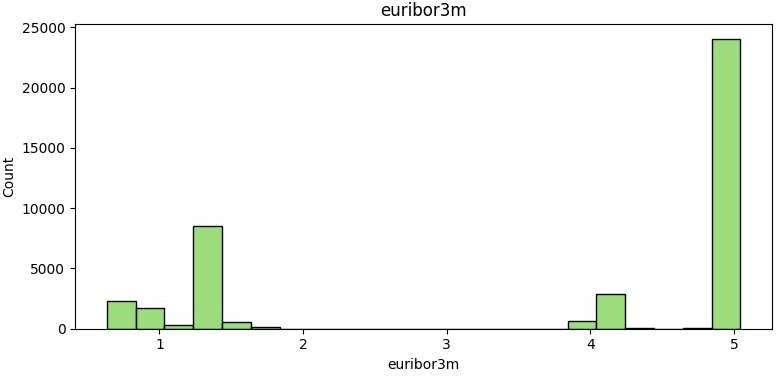


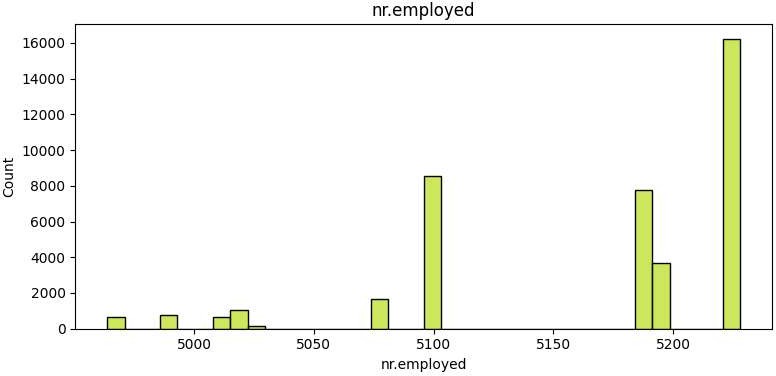




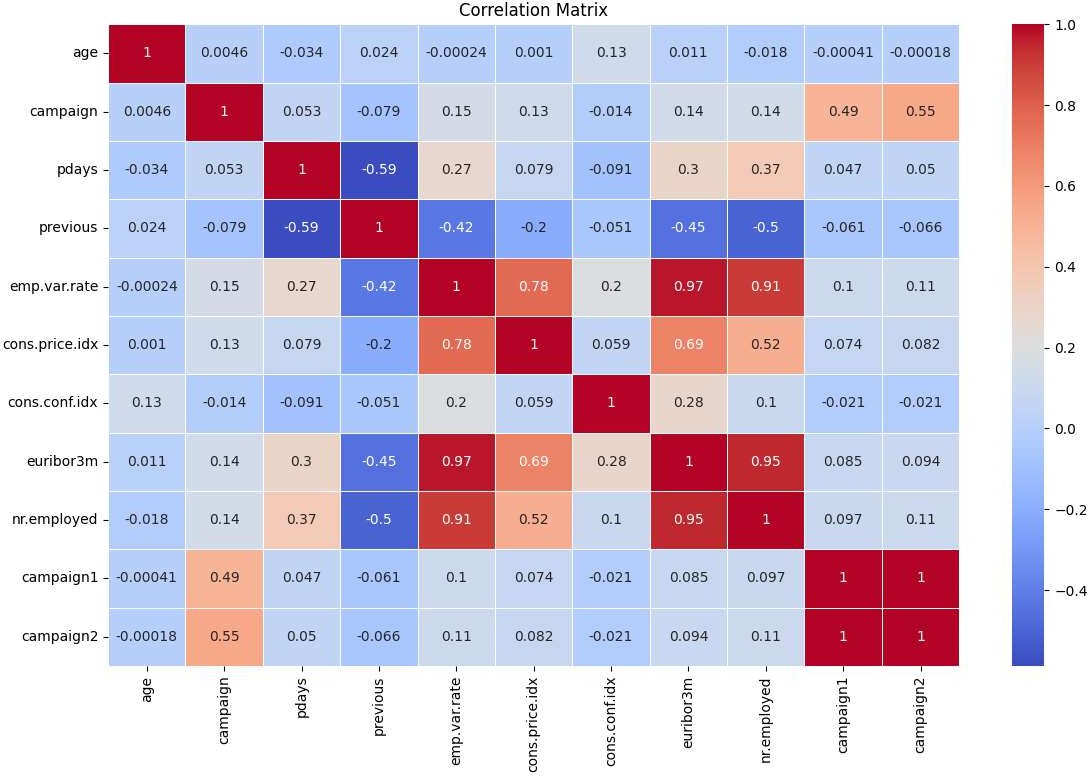




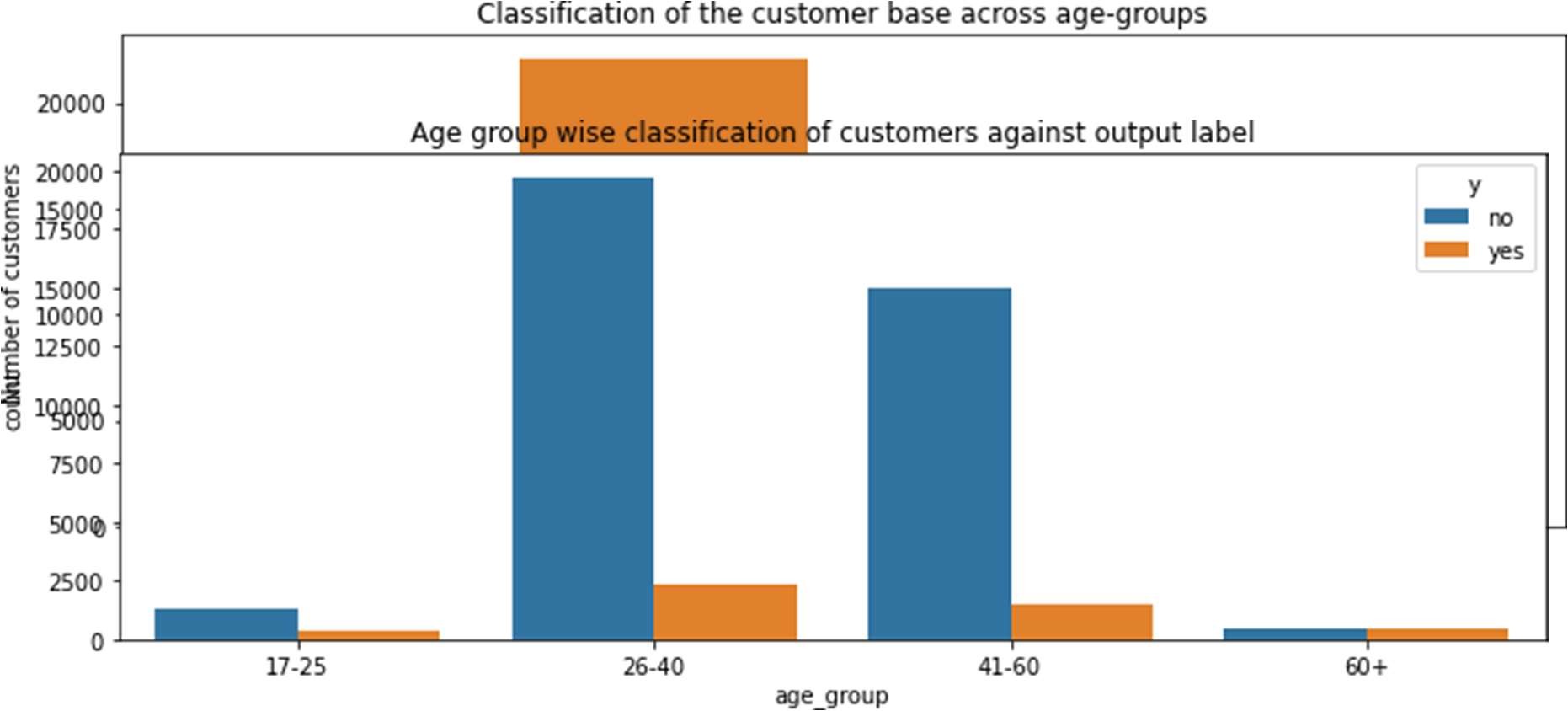




 Correlation matrix



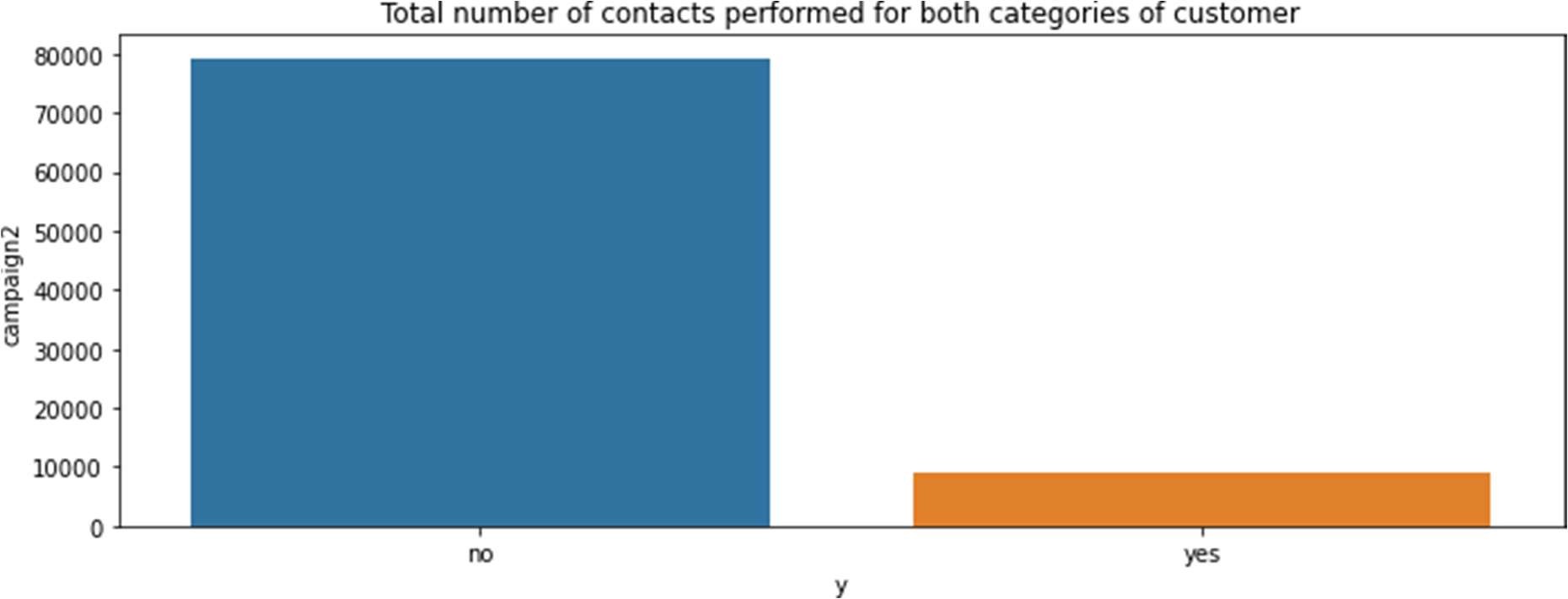
 Classification of the customer base across age-groups



 Looking at relation between different age groups and the output labely

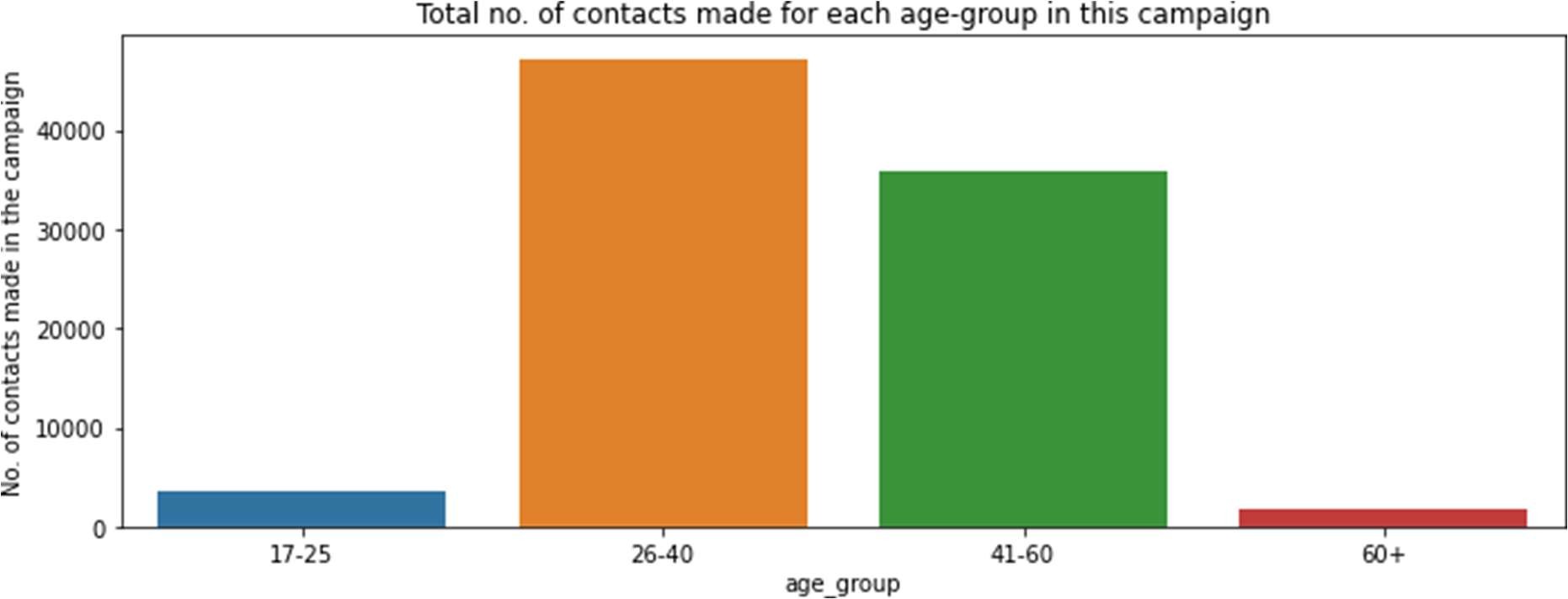
**Observation:** In the age-groups of 26-40 and 41-60 yrs, majority of thepeople are not subscribed to the term deposit plan

 Looking at relation between Number of contacts made to the customer (campaign) and the output label y



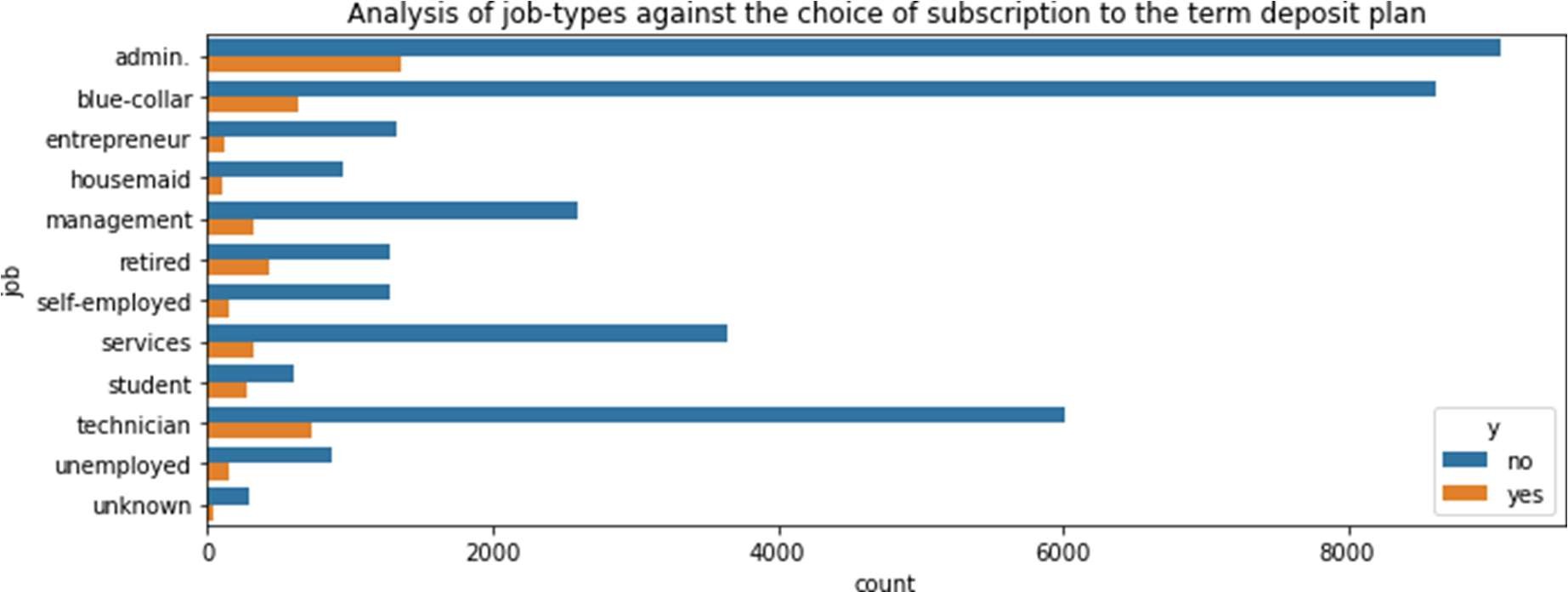
**Observation**: When a greater number of contacts is made to the customer,they haven't subscribed to the term deposit plan

 Looking at relation between 'age\_group' and 'campaign' that is numberof contacts performed for each age group



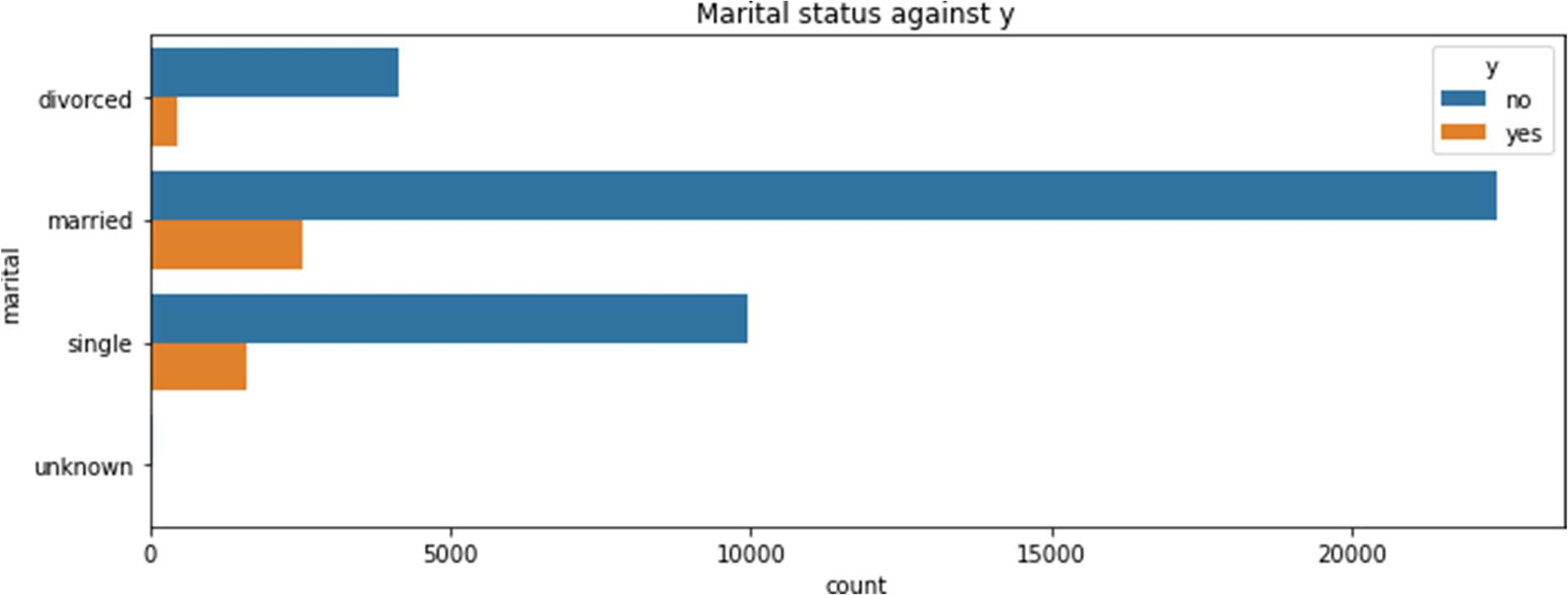
**Observation:** The 26-40 and 41-60 age-groups witness majority of thecontacts made in this campaign. These two age-groups seem to the target groups for the bank.

 Looking at relation between job and the output label y



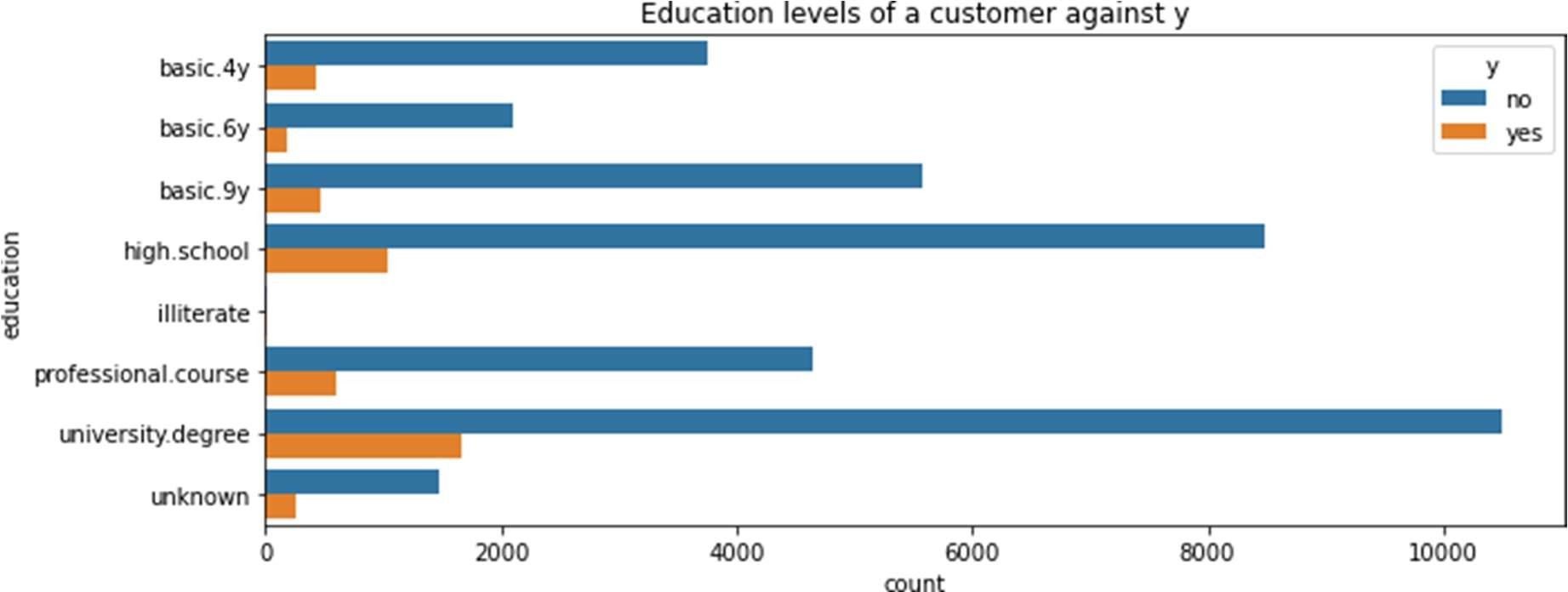
**Observation:** Looking at the jobs, 'admin', 'blue-collar' and 'technician' are the prominent jobs and most of the customers in these jobs have rejected the term deposit plan.

 Analysing marital status and the output label



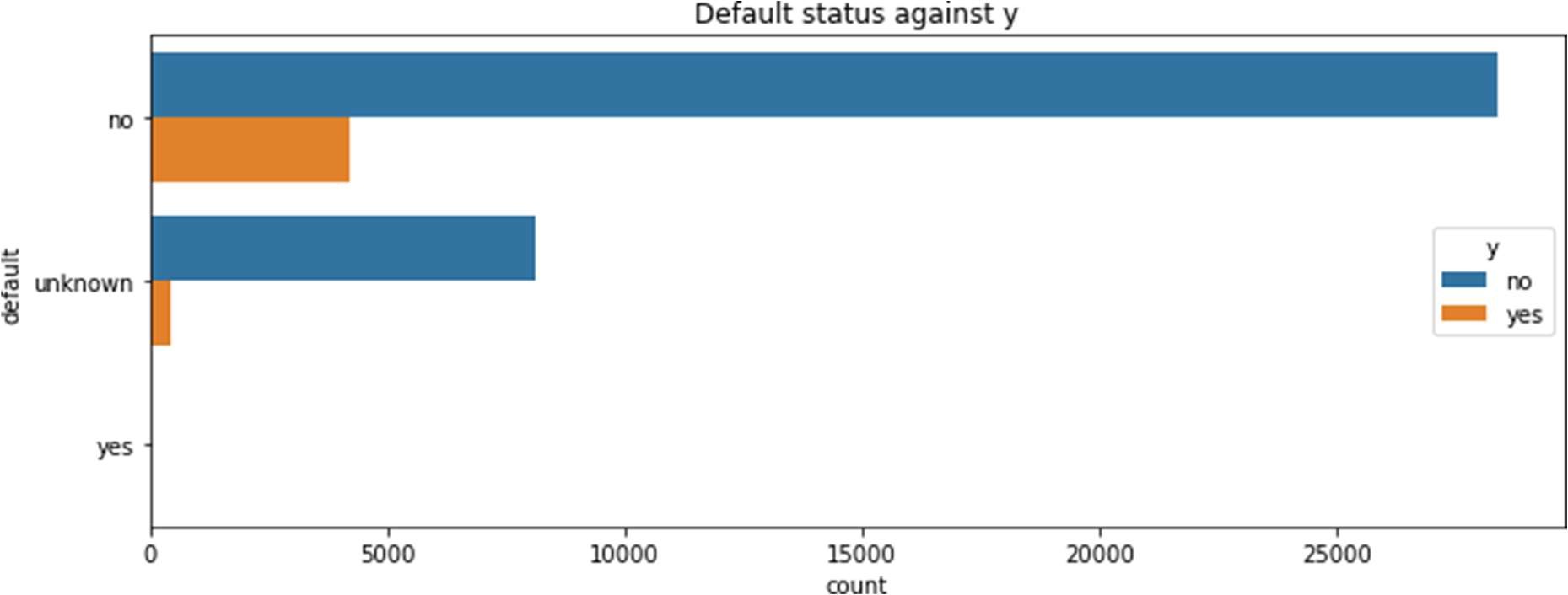
**Observation:** married and single customers are the majority of the customer base and comparatively married customers have taken theterm deposit

 Analysing the different education levels of a customer against thechoice of subscription

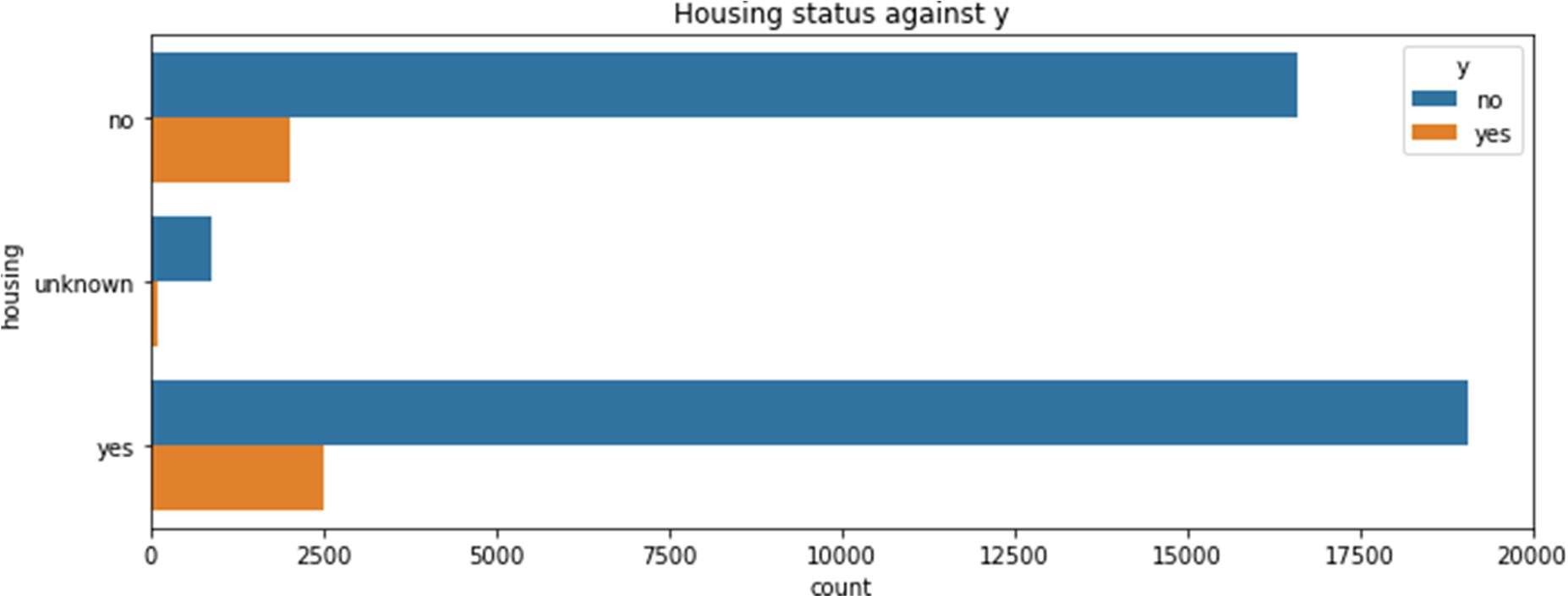


**Observation:** Customers with university degree have subscribed tothe term deposit more

 Analysing the default status against the choice of subscription

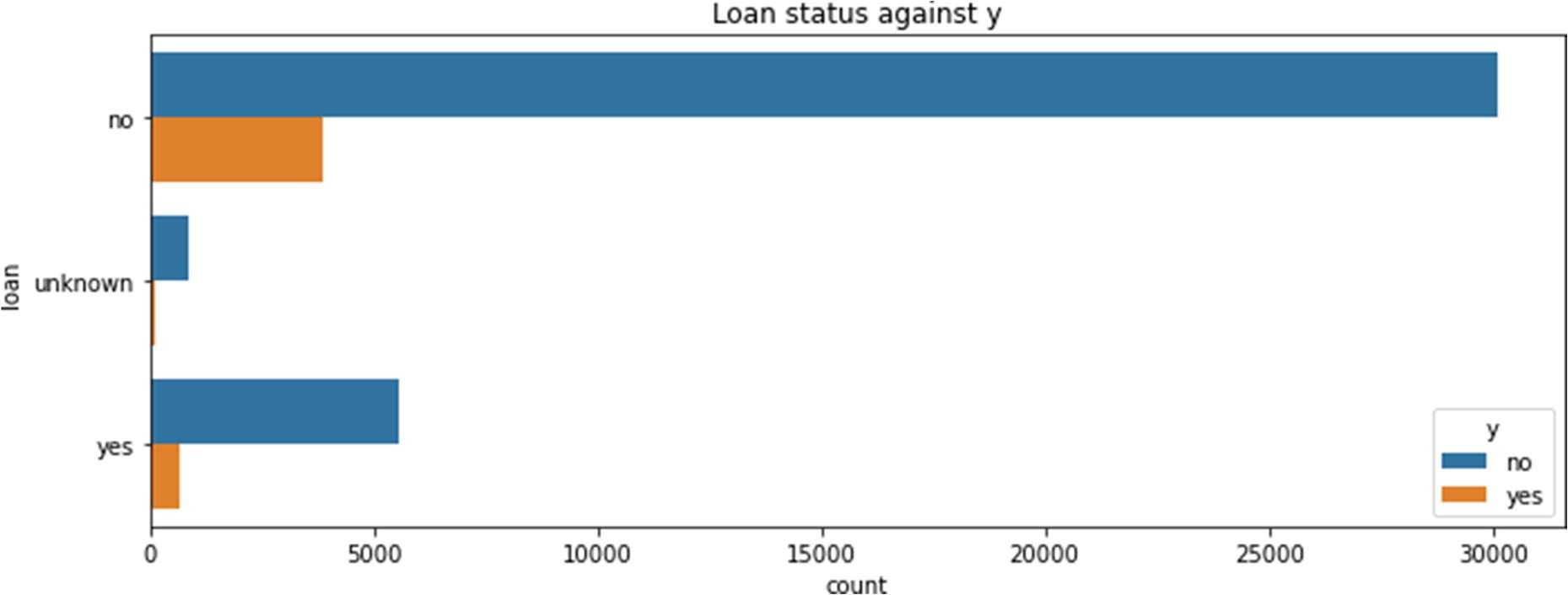


 Analysing housing status and y



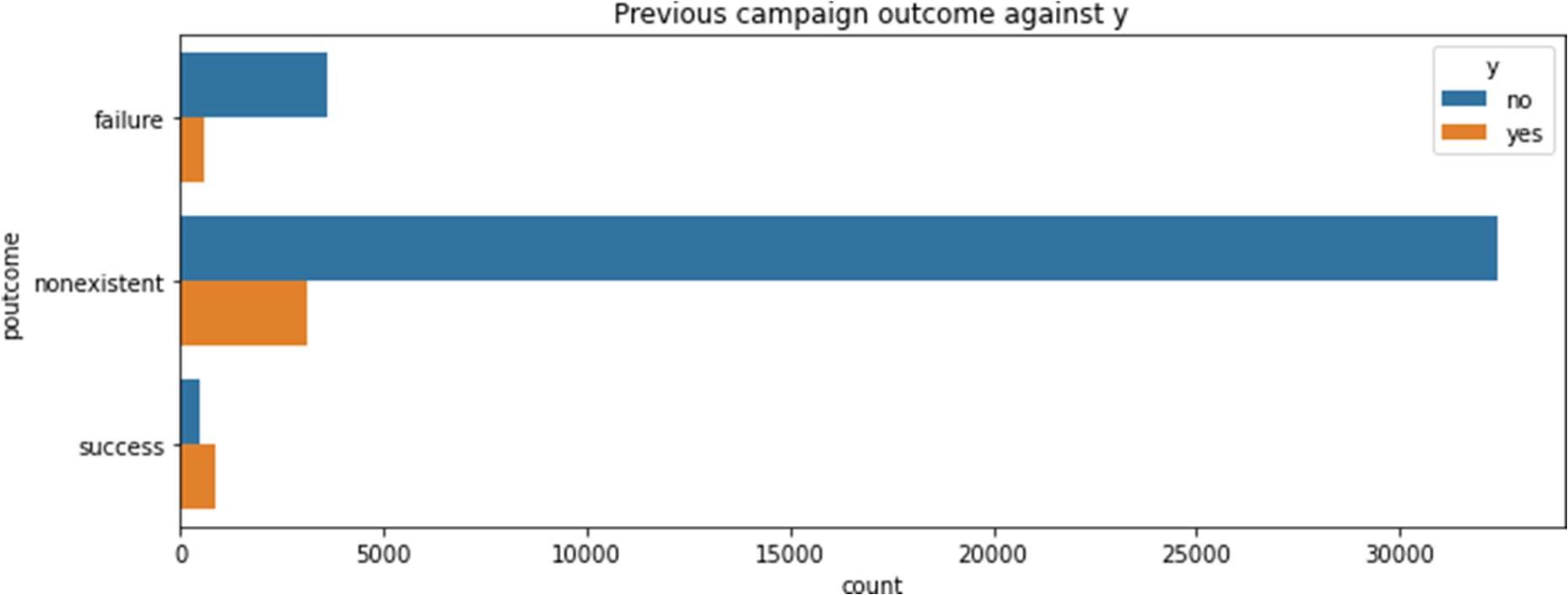
**Observation:** Number of customers who have subscribed to the termdeposit is comparatively more for those with housing loan

 Analysing loan status and y



**Observation:** Number of customers who have subscribed to the term deposit is comparatively less for those with personal loan

 Analysing poutcome and y



**Observation:** The success rate of previous marketing campaign has resulted in a greater number of people subscribing to the term deposit

# Model Building

In the context of machine learning, where models rely on numerical input features, the initial step involves converting categorical features into numeric representations. This conversion is followed by feature scaling across all features to expedite convergence to the global minimum during gradient descent. The categorical features encompassed in this process include 'job,' 'marital,' 'education,' 'default,' 'housing,' 'loan,' 'contact,' 'month,' 'day\_of\_week,' 'poutcome,' 'campaign2,' and 'y.'

A screenshot of a computer program

Description automatically generated

To perform one hot encoding, dummy variables are created for the features job, marital, education, contact and poutcome.

A close up of a white background

Description automatically generated

Encoding month and day of the week using the equivalent numbers.

Encoding default, housing, loan and y with 1, 0 and -1 for yes, no and unknown.

A white background with black and red text

Description automatically generated

A screenshot of a computer

Description automatically generatedNow the features are all encoded into numbers.

### Training using Machine Learning classifiers

A screenshot of a computer

Description automatically generatedThe dataset is split into Train and Test

A white screen with black text

Description automatically generatedTo perform feature scaling, StandardScaler() is used.

I have chosen the following models for testing.

1. Logistic Regression (Linear)
2. Decision Tree (Linear)
3. Naïve Bayes (Linear)
4. Random Forest (Ensemble)
5. Gradient Boosting (Boosting)

For evaluating the model, cross validation testing is used with the number of folds as 10 and accuracy as the metric.

A computer screen shot of a computer

Description automatically generated

From all the above Models, **Logistic Regression** performed the best with an accuracy of **84%.**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **Logistic Regression** | **84%** |
| Decision Tree | 28% |
| Naïve Bayes | 72% |
| Random Forest | 45% |
| Gradient Boosting | 53% |

### Hyperparameter Tuning

Since Logistic Regression gave better performance, hyper parameter tuning was performed on it to identify best features.

A computer screen shot of a computer code

Description automatically generated

The best parameters values were identified as

|  |  |
| --- | --- |
| **Parameter** | **Values** |
| C | 0.1623776739188721 |
| penalty | 'l1' |
| Solver | liblinear |
| max\_iter | 80 |

With Hyperparameter tuning the accuracy increased from **84% to 89.58%.**

**Evaluation metrices**

The classification report and the confusion matrix are reported as

**Observations:**

1. Confusion matrix results tell us that we have 10785 + 144 Correct predictions and 1135+289 incorrect
2. Classification report shows precision as 90% which is the ability of a classification model to identify only the relevant data points, that is in this case people who would be subscribing to the term deposit is correctly classified.

A screenshot of a computer screen

Description automatically generated

### Evaluating using AUC-ROC curve

The model is also evaluated using AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve which is an evaluation metric for binary classification problems. An ROC curve is a graph showing the performance of a classifier. ROC is a probability curve plotted with True Positive Rate (also called Recall or Sensitivity) on the y-axis against False Positive Rate (also called as Precision) on the x-axis. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

A graph of a logistic regression

Description automatically generated

\*\*Github Repo Link\*\*

**https://github.com/supinhooda/assignment/tree/main/Final%20Project**