Project: Bank Marketing Campaign

Name: Archana Devi Ramesh
Batch Code: LISUM16
Specialization: Data Science
Github link: https://github.com/ArchanaDeviRamesh/Data-Glacier-Project/tree/main/FINAL%20PROJECT

Agenda

- Problem Statement
- Business Understanding
- Objective
- Dataset
- EDA
- Recommendation
- Model building
- Evaluation
- Results

Problem Statement

- ABC Bank wants to sell it's 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). This is an application of the organization’s marketing data.

Business Understanding

- In predicting the results of marketing campaign for each customer and interpreting which all features affect the results, will help the organization understand how to make campaign more efficient. Moreover, in categorizing which segment of customers subscribed the term deposit, helps to identify who is more likely to buy the product in future thereby developing more targeted marketing campaigns. This can be achieved using ML model that shortlists the 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. This will save resource and their time.

Objective

- Build a Classification ML model to shortlist customers who are most likely to buy the term deposit product. This would allow the marketing team to target those customers through various channels.

Data Glacier
Your Deep Learning Partner

Dataset

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

bank-additional-full.csv: 20 inputs (+1 target variable) and 41118 observations

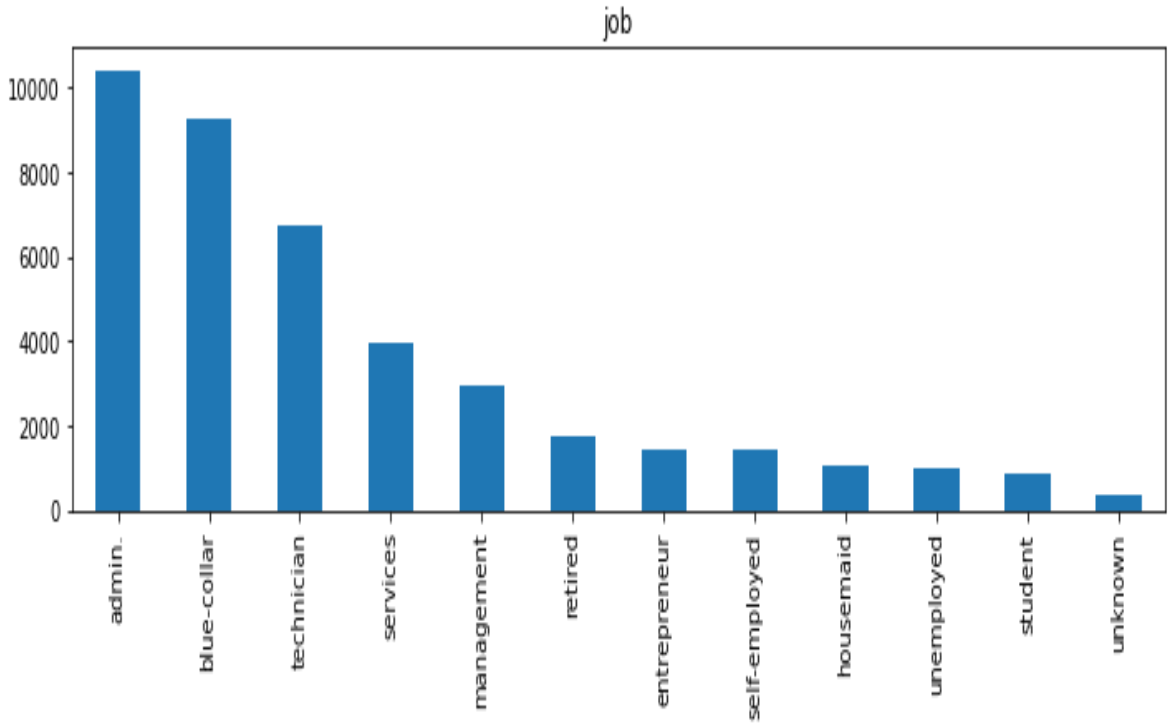
Assumptions:

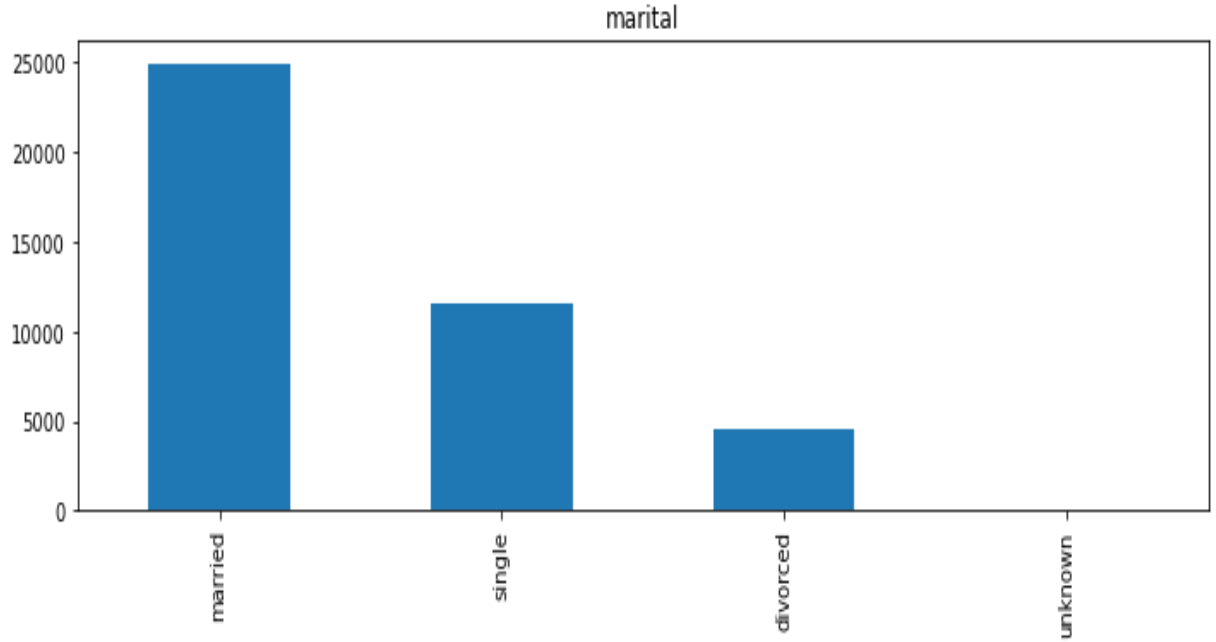
- 'Duration' feature is dropped as suggested in the dataset description
- A frequently occurring missing value 'unknown' is considered as another category for the categorical features.
- Duplicate rows were deleted from the dataset.

EDA

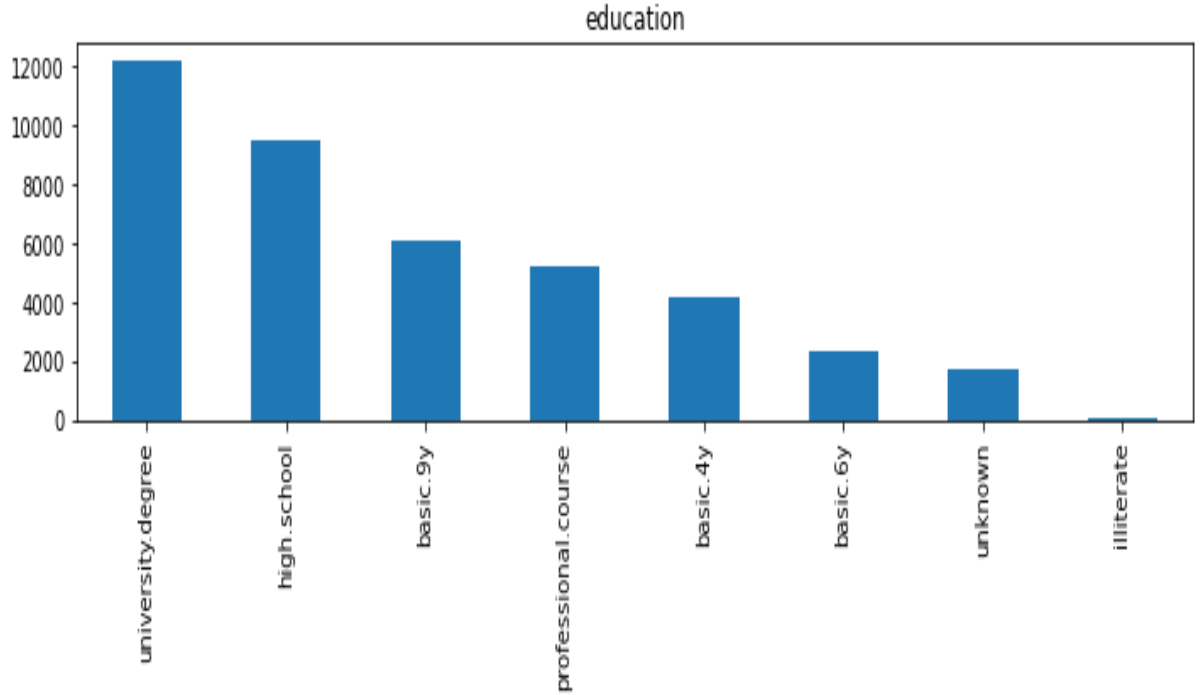
EDA
1. Univariate Analysis

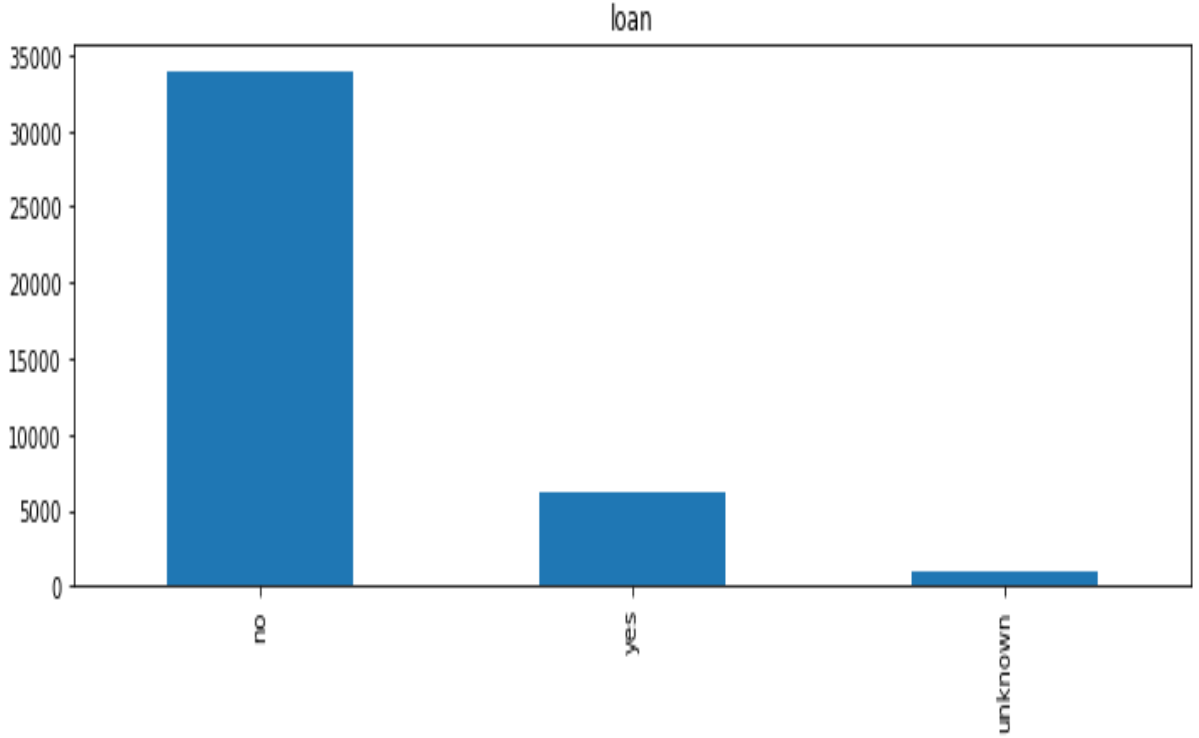
- In Jobs, most of the customers are administrative staffs and technicians.
- Married customers have been communicated more for subscription to the deposit more compared to single



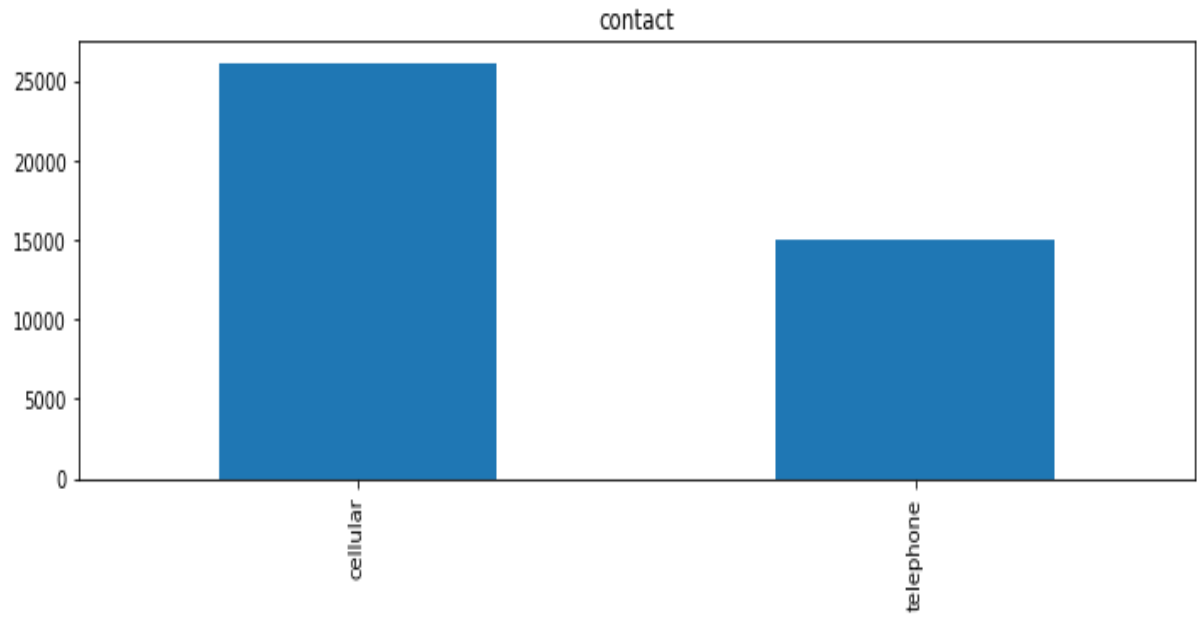


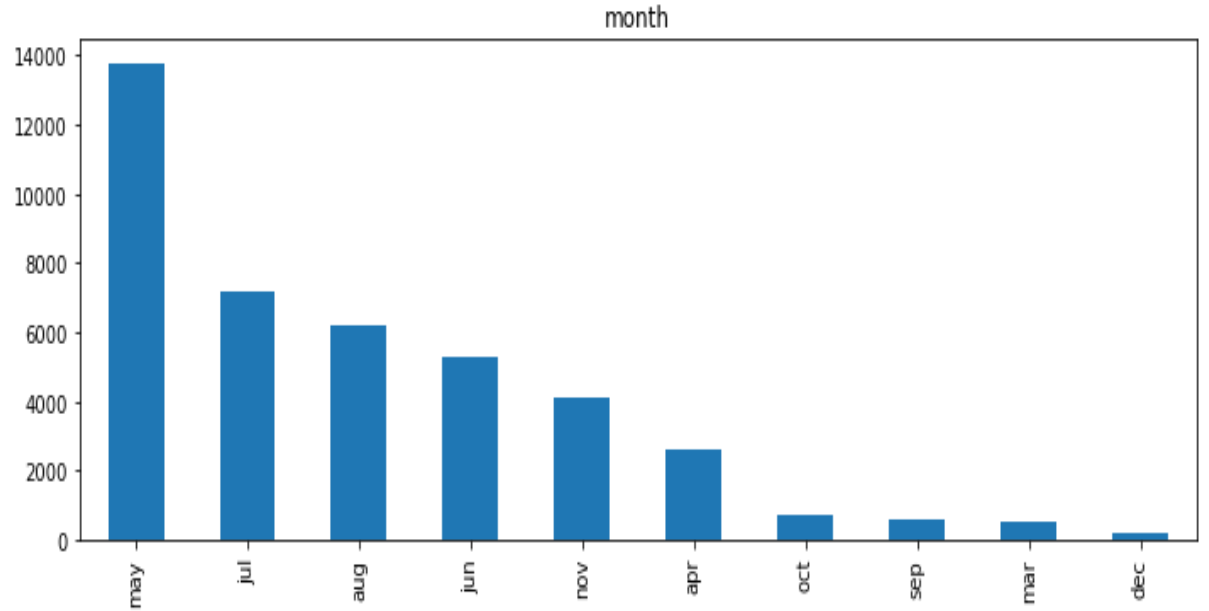
- Most of the customers are university graduates followed by high school degree
- Majority of the customers do not have a personal loan



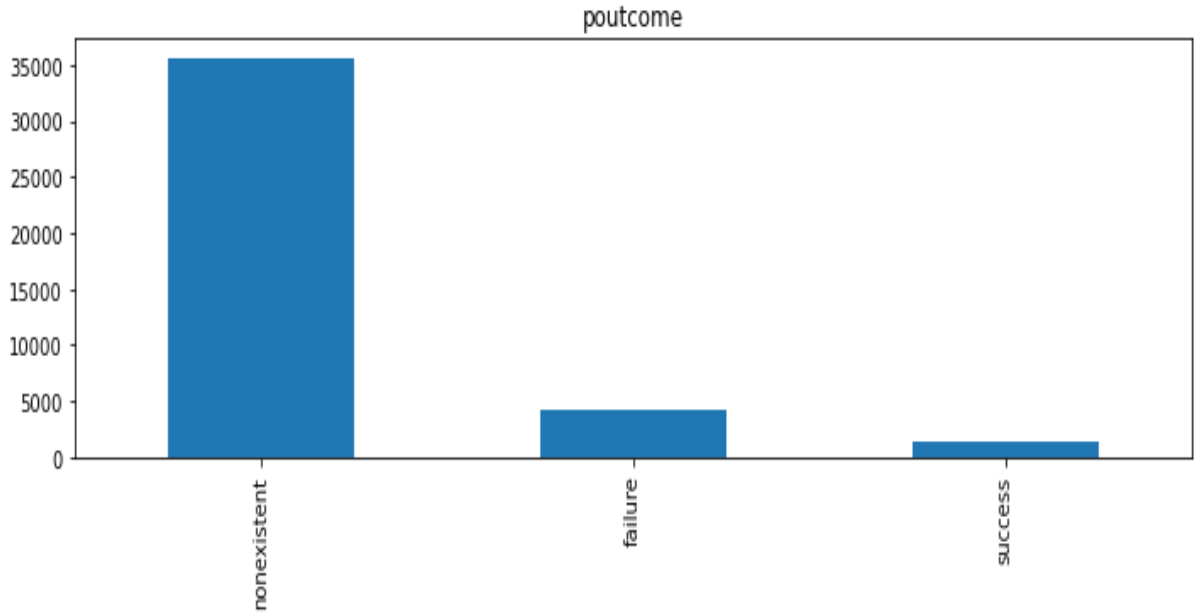


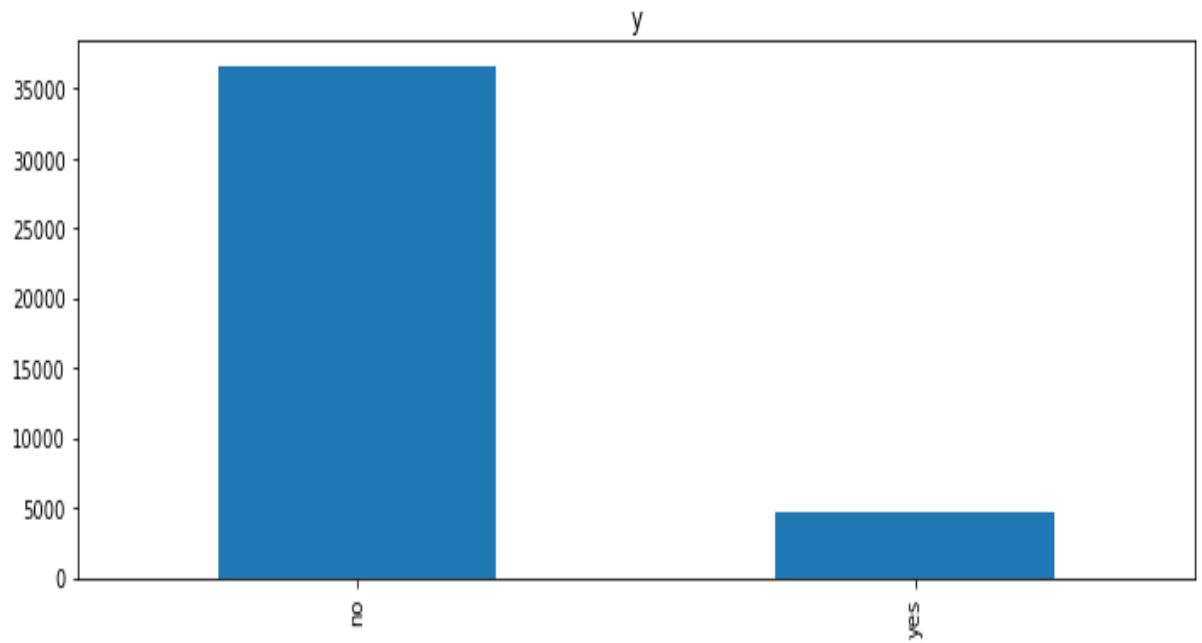
- The type of communication used to contact customers is mostly cellular compared to telephone
- May seems to be the month with most contacts made



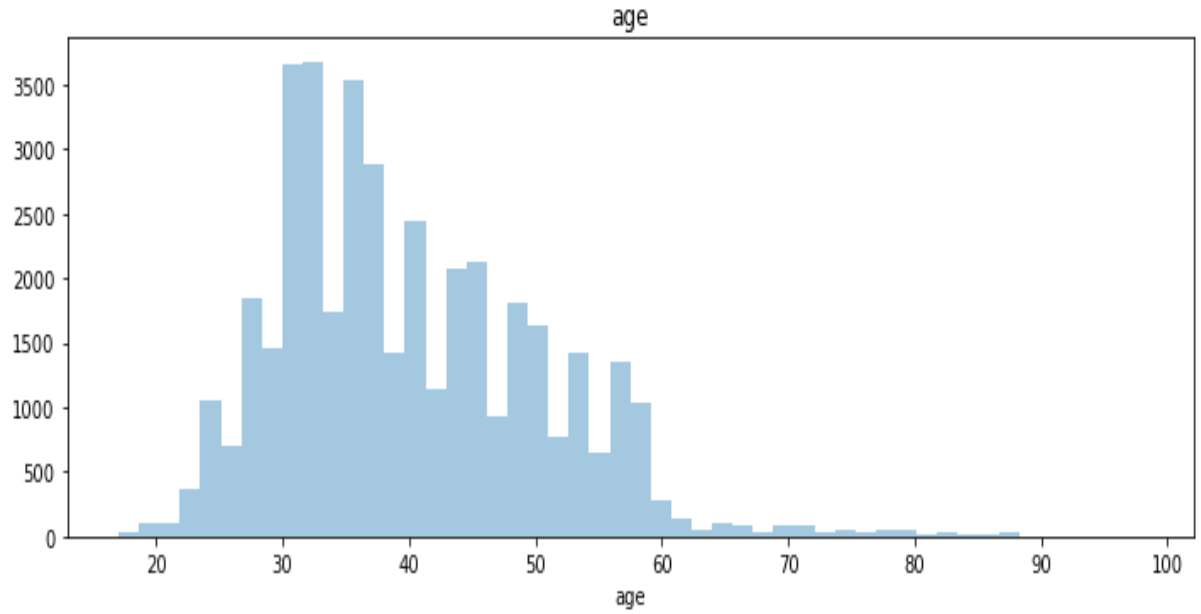


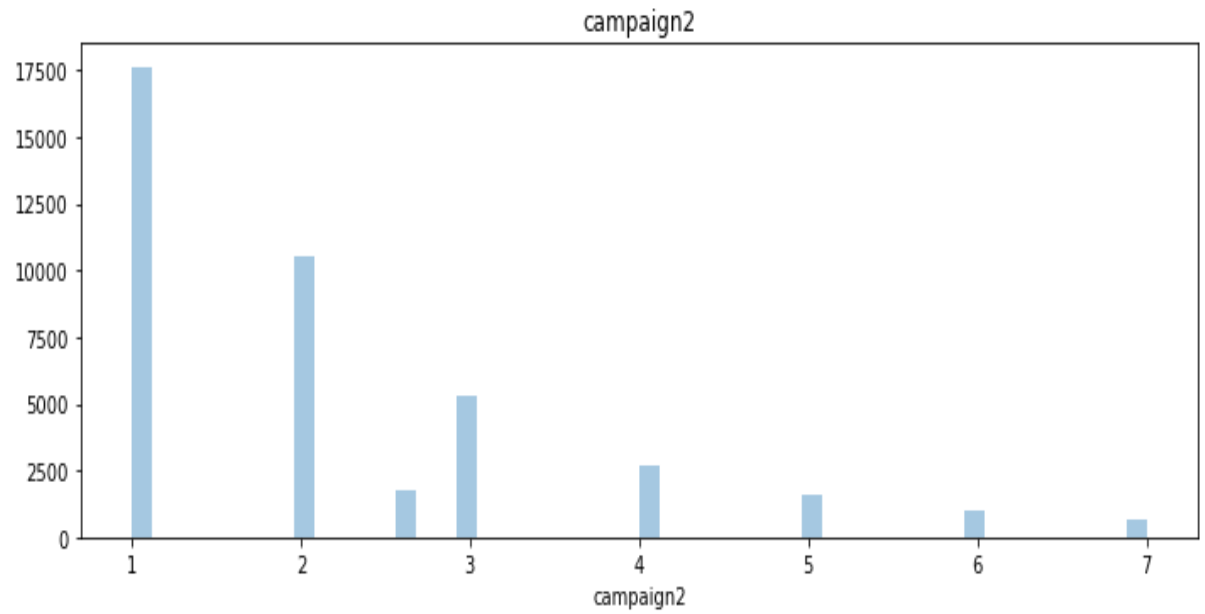
- Most of the previous campaign results whether succeeded or failed is non existent
- In the previous campaign, the percentage of people who subscribed to the deposit is less to those who did not.



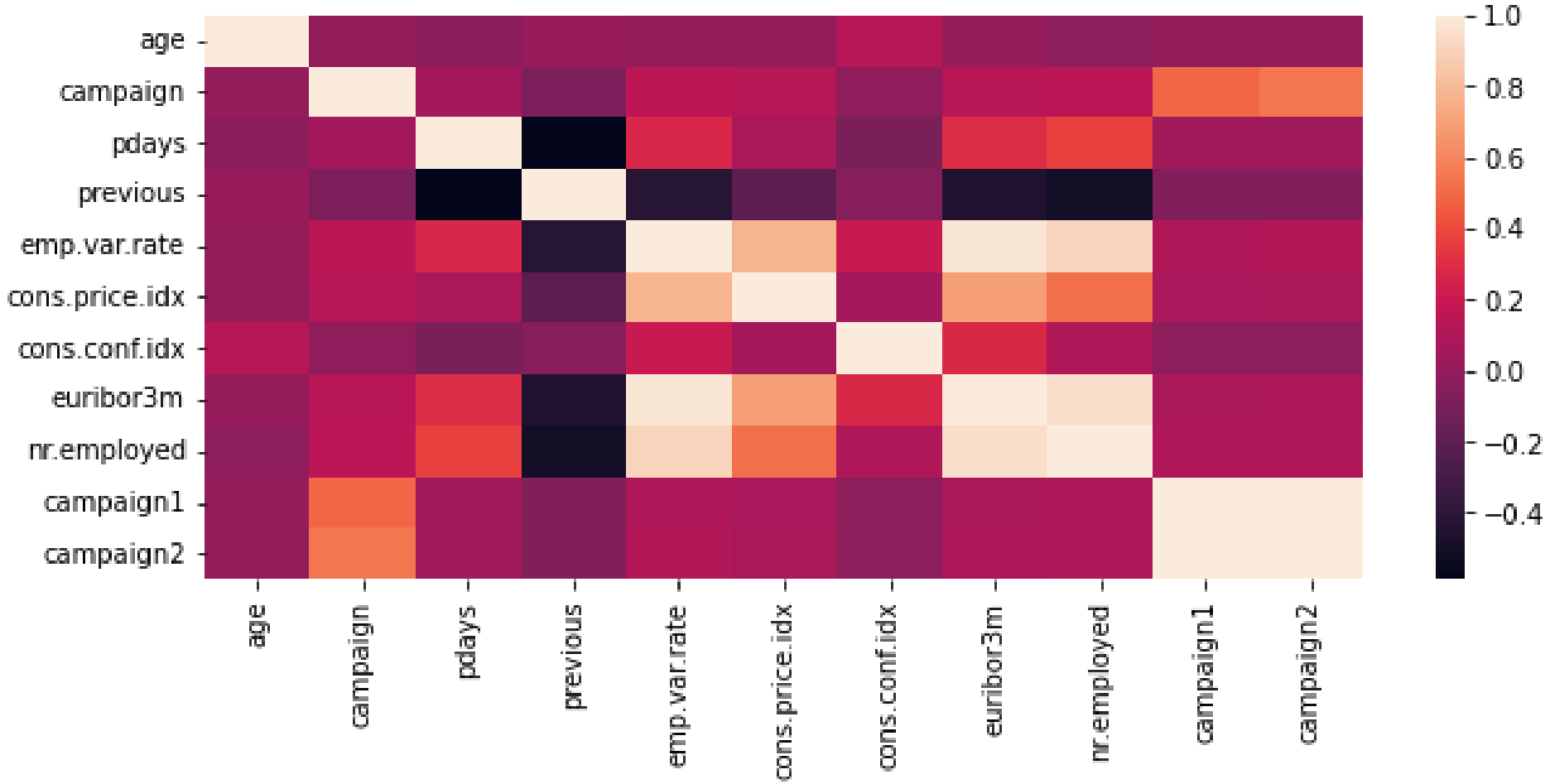


- The age group of customers contacted mostly fall between 20 to 60
- Number of follow ups made to a customer is less in the previous campaign



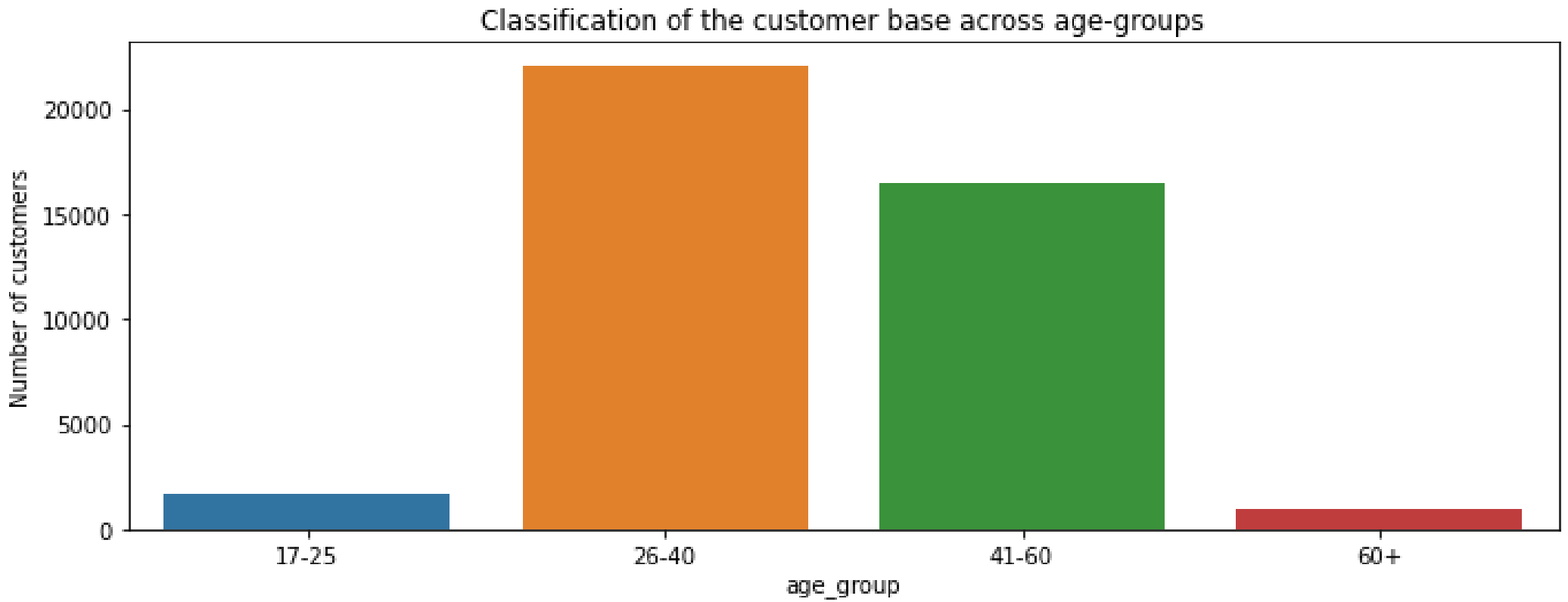


Correlation map

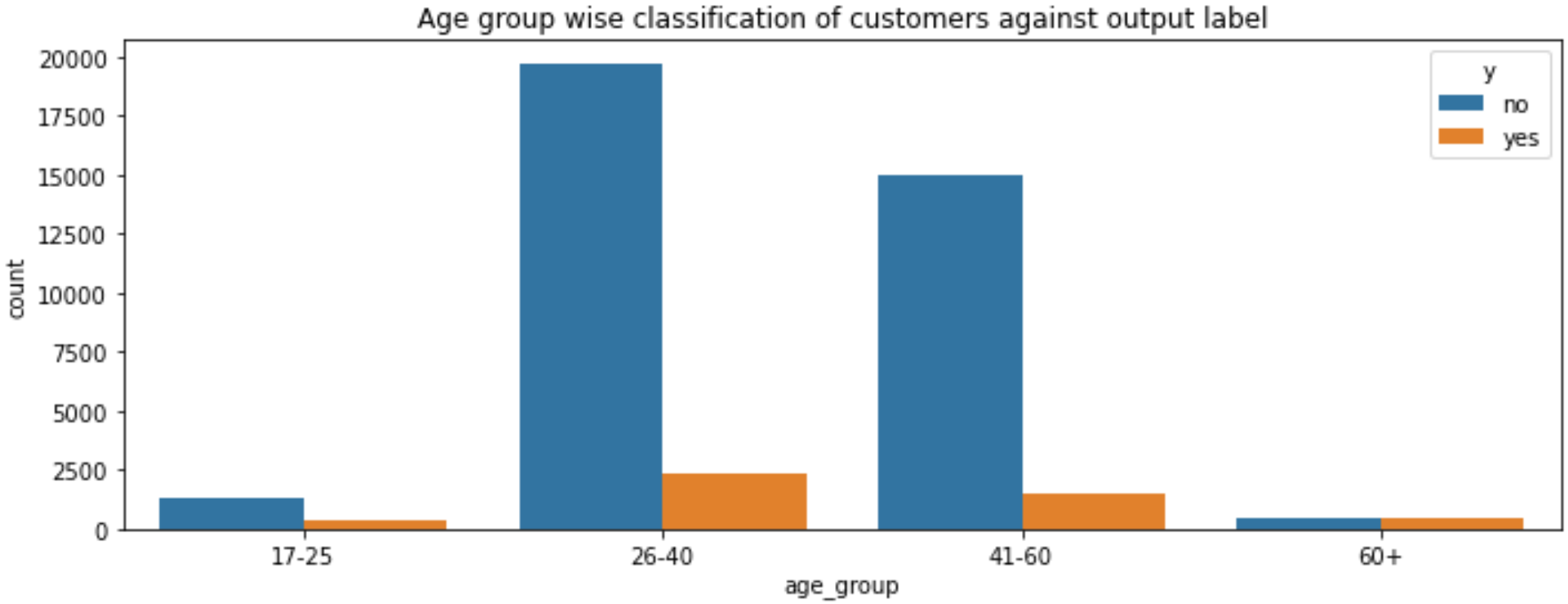


EDA
2. Bivariate Analysis

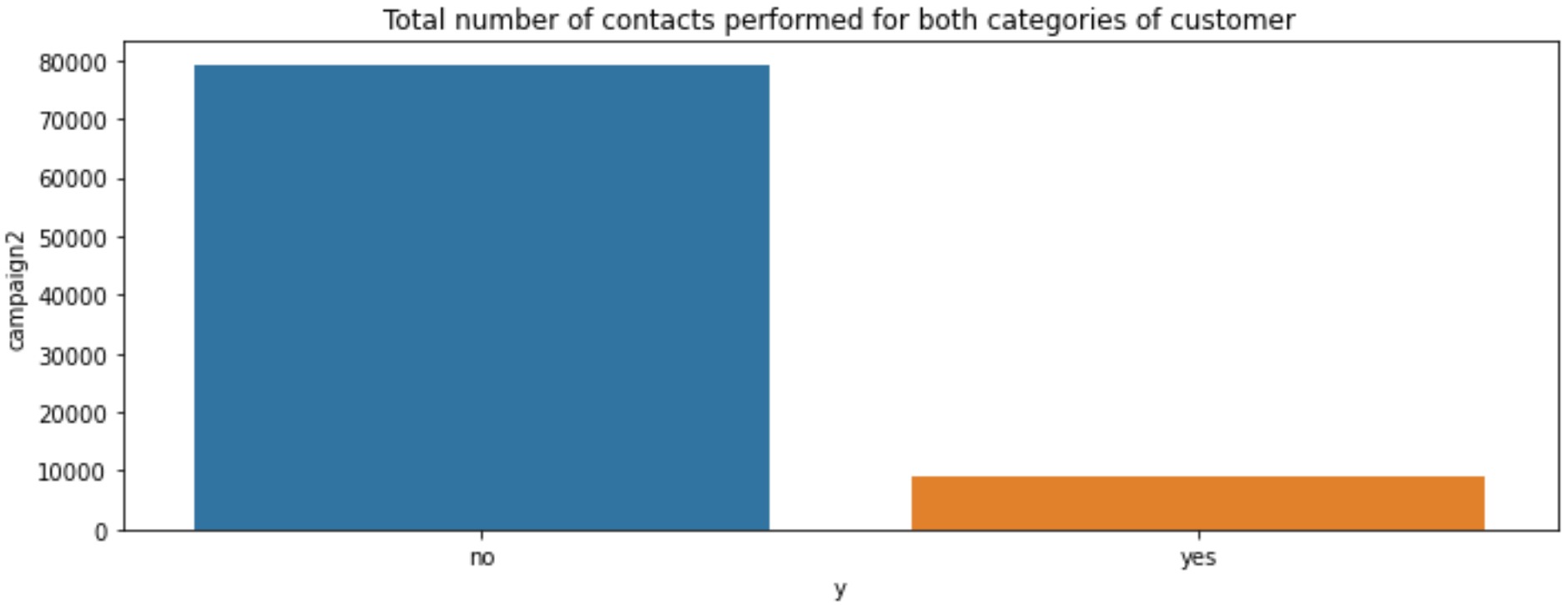
- Classification of the customer base across age-groups



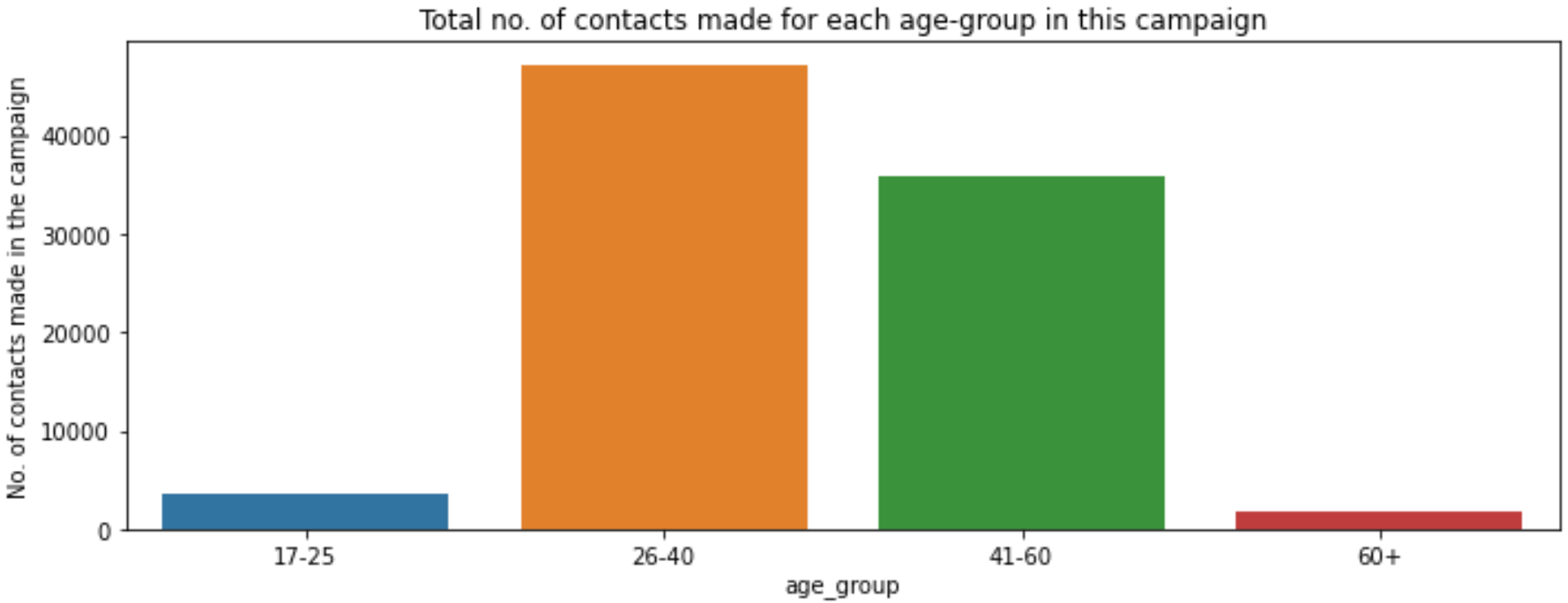
- Looking at relation between different age groups and the output label y
- In the age-groups of 26-40 and 41-60 yrs, majority of the people are not subscribed to the term deposit plan



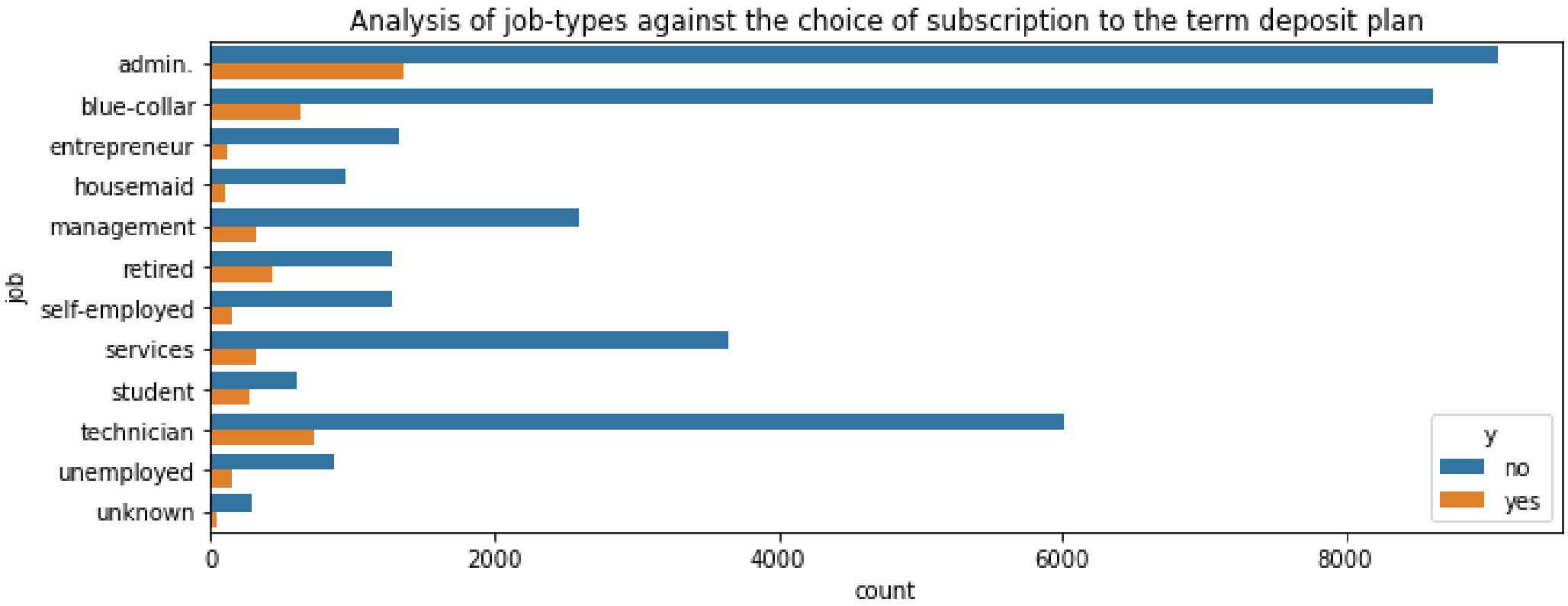
- Looking at relation between Number of contacts made to the customer (campaign) and the output label y
- When a greater number of contacts is made to the customer, they haven't subscribed to the term deposit plan



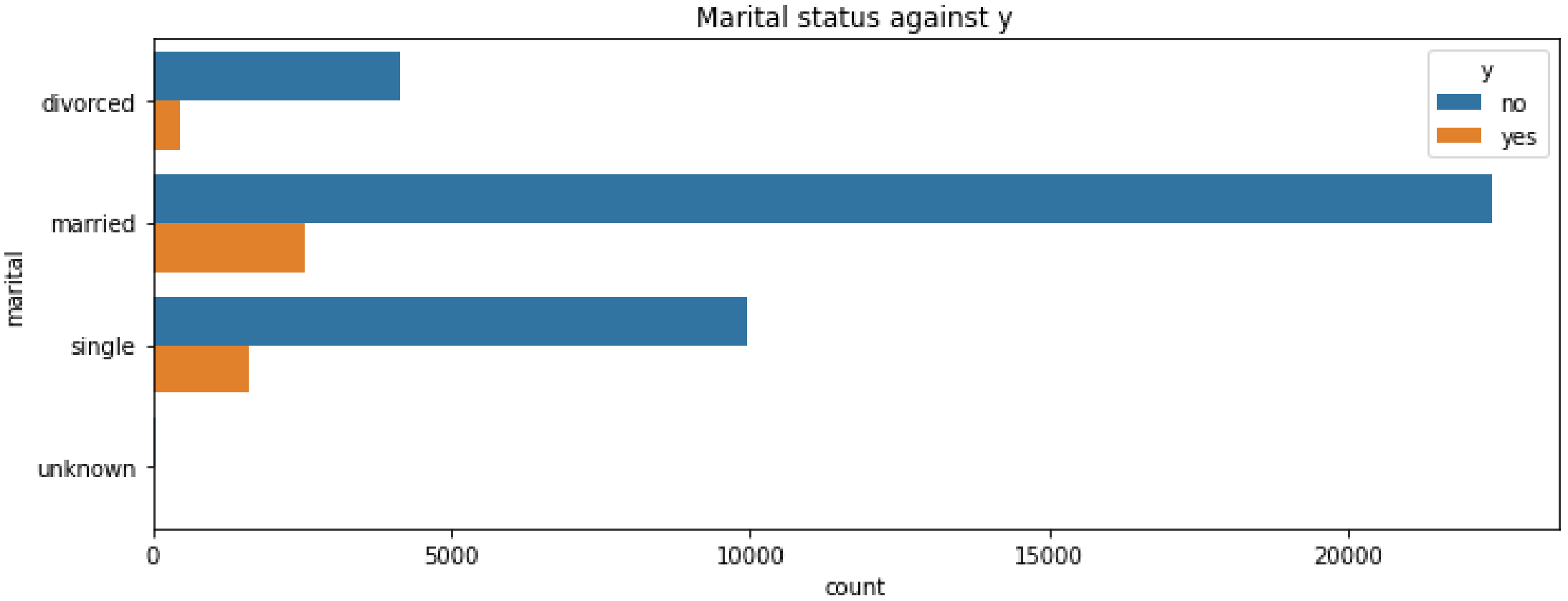
- Looking at relation between 'age_group' and 'campaign' that is number of contacts performed for each age group
- The 26-40 and 41-60 age-groups witness majority of the contacts made in this campaign. These two age-groups seem to the target groups for the bank.



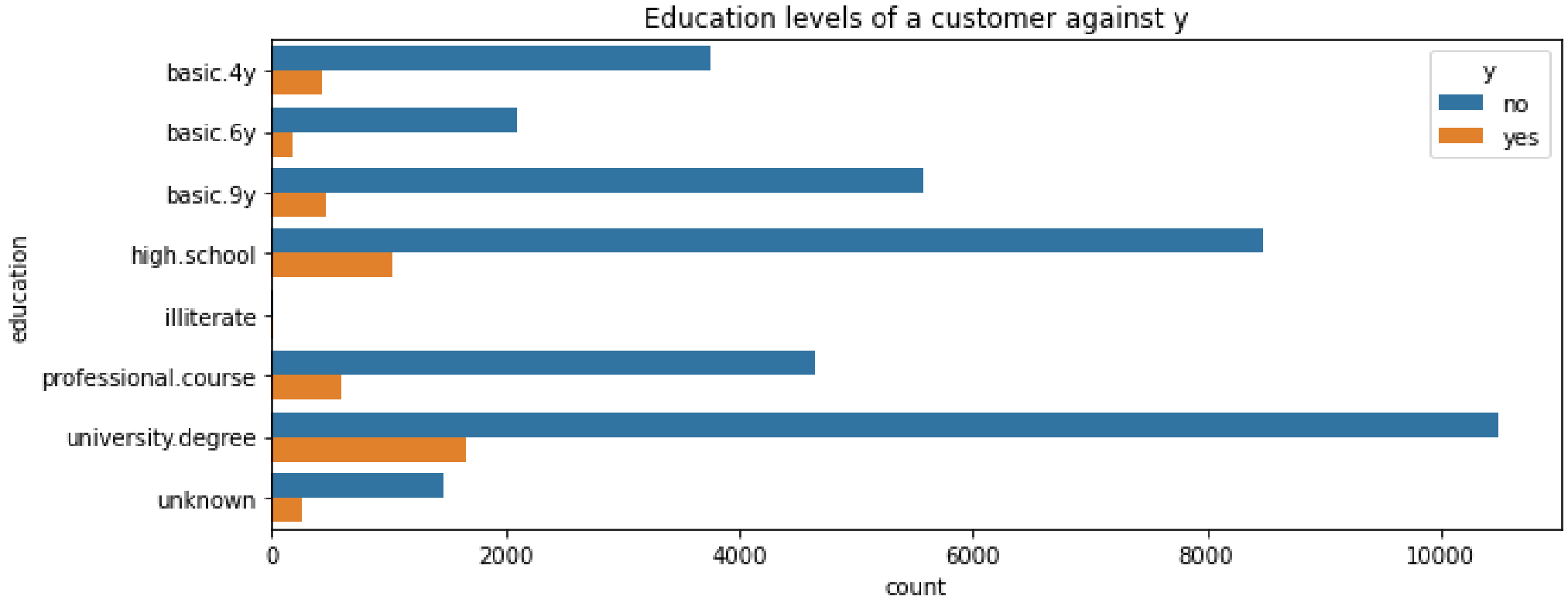
- Looking at relation between job and the output label y
- Looking at the jobs, 'admin', 'blue-collar' and 'technician' are the prominent jobs and most of the customers in these jobs have rejected the term deposit plan.



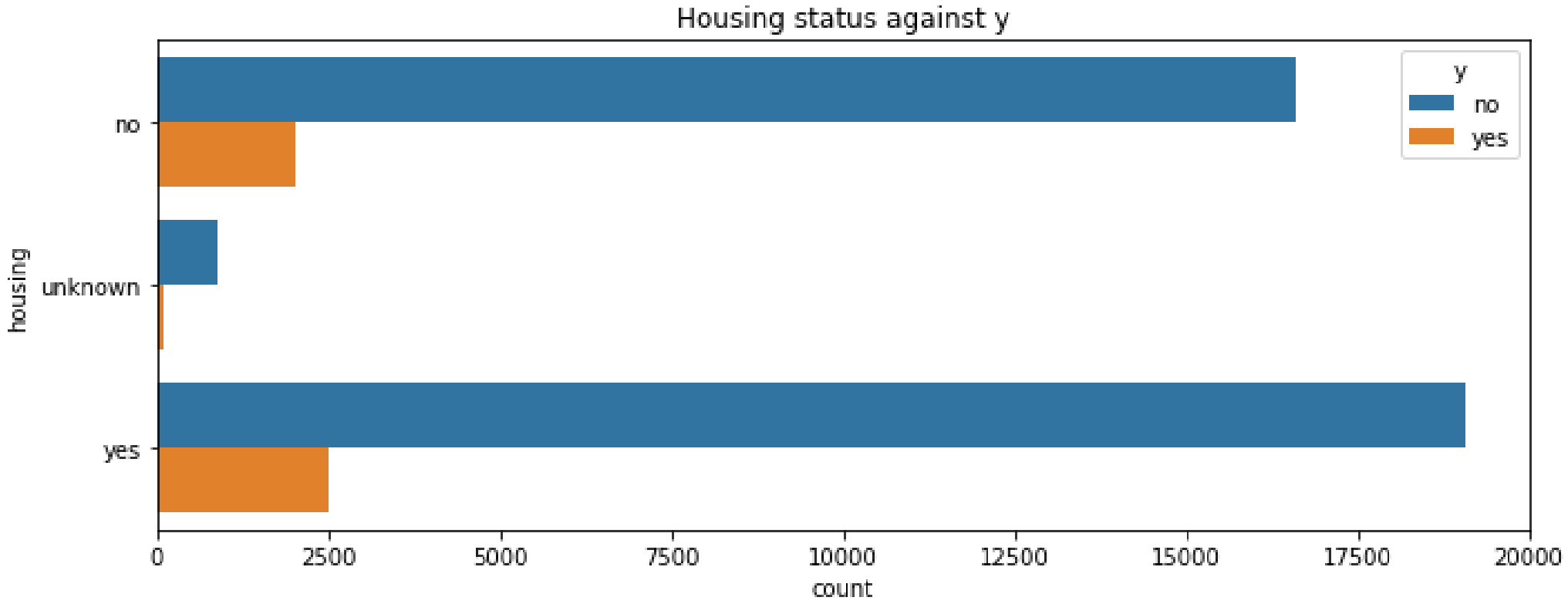
- Analysing marital status and the output label
- married and single customers are the majority of the customer base and comparatively married customers have taken the term deposit



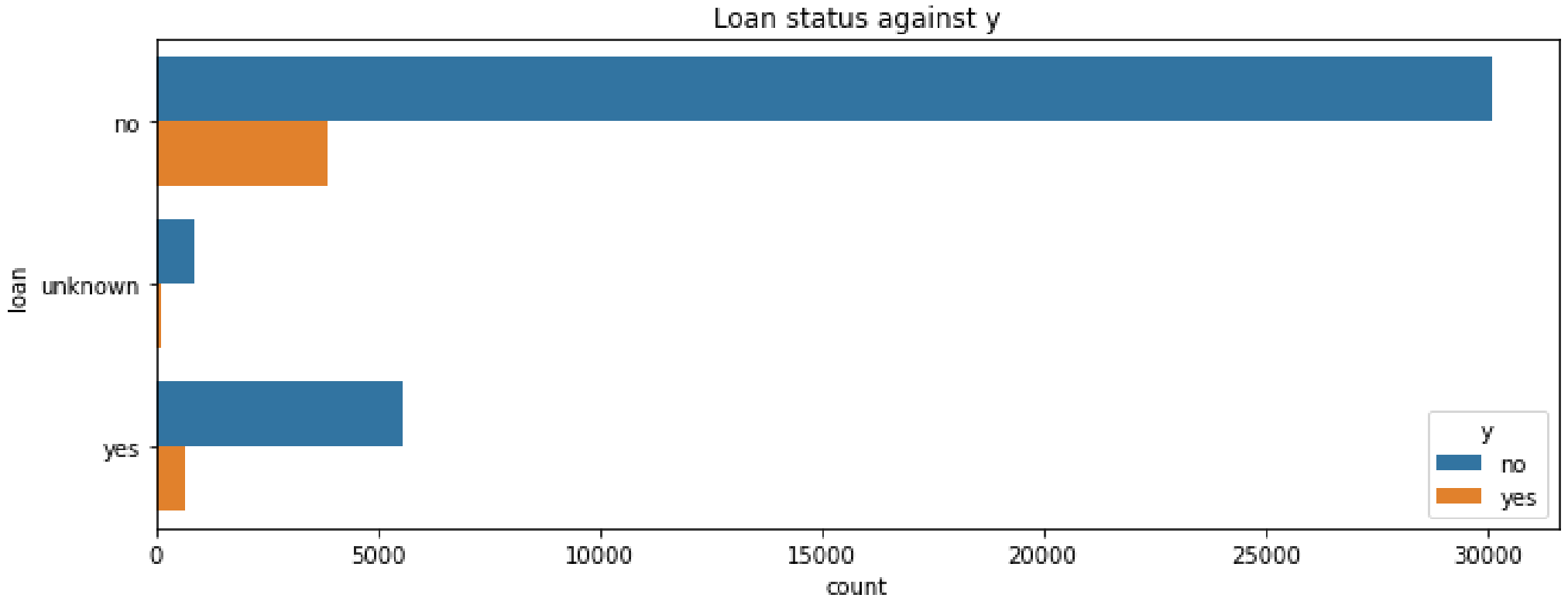
- Analysing the different education levels of a customer against the choice of subscription
- Customers with university degree have subscribed to the term deposit more



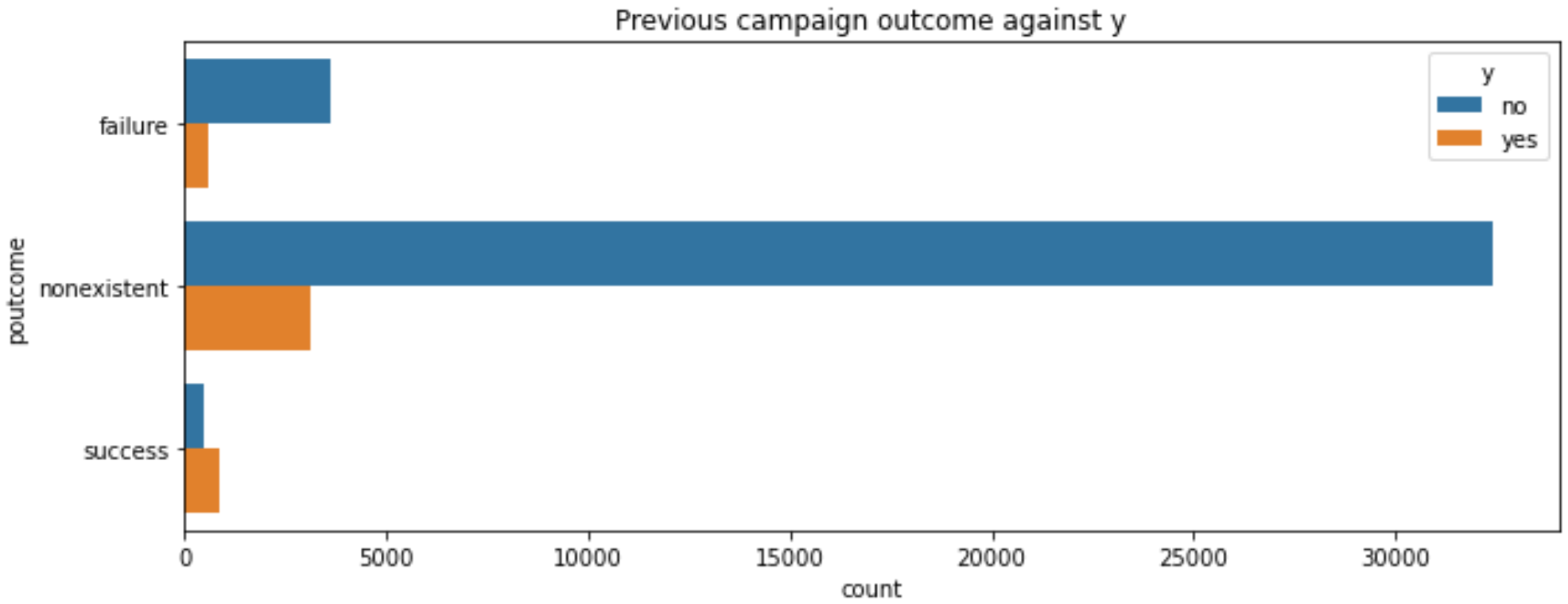
- Analysing housing status and y
- Number of customers who have subscribed to the term deposit is comparatively more for those with housing loan



- Analysing loan status and y
- Number of customers who have subscribed to the term deposit is comparatively less for those with personal loan



- Analysing poutcome and y
- The success rate of previous marketing campaign has resulted in more number of people subscribing to the term deposit



Recommendation to improve campaign

1. May is the most effective month to contact customers
2. Increase the time of contacts made per customer
3. Give more focus on university graduate students and high school degree students
4. age-groups of 26-40 and 41-60 have a higher proportion among customers, therefore these groups present a profitable target for the marketing team.
5. Target the admins and technicians for more subscriptions

Recommended models for this dataset

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

Hyper parameter tuning and model evaluation will be performed in order to determine the best model and the important features

Model Building

- Categorical features like 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome', 'campaign2' and ‘y’ are converted to numerical features
- Feature scaling is performed on all features to achieve global minimum fast in gradient descent using StandardScaler()
- Dataset is split into train and test using train_test_split()
- The dataset is trained using the five recommended models, Logistic Regression (Linear), Decision Tree (Linear), Naïve Bayes (Linear), Random Forest (Ensemble) and Gradient Boosting (Boosting)
- For evaluating the model, cross validation testing is used with the number of folds as 10 and accuracy as the metric.

Accuracy results

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

From all the above Models, Logistic Regression performed the best with an accuracy of 84% therefore I recommend this model for production purpose

Hyper parameter tuning

- **Since Logistic Regression gave better performance, hyper parameter tuning was performed on it to identify best features.**
- **The model was run using the following parameters**

Parameter	Values
C	100, 10, 1.0, 0.1, 0.01
penalty	'l1', 'l2'
Solver	'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'
max_iter	Between 80 and 120

- **The best parameters values were identified as**

Parameter	Values
C	1.0
penalty	'l1'
Solver	'saga'
max_iter	87

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

Confusion matrix and classification report

- **Confusion matrix results tell us that we have 10780 + 149 Correct predictions and 1138+286 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.**

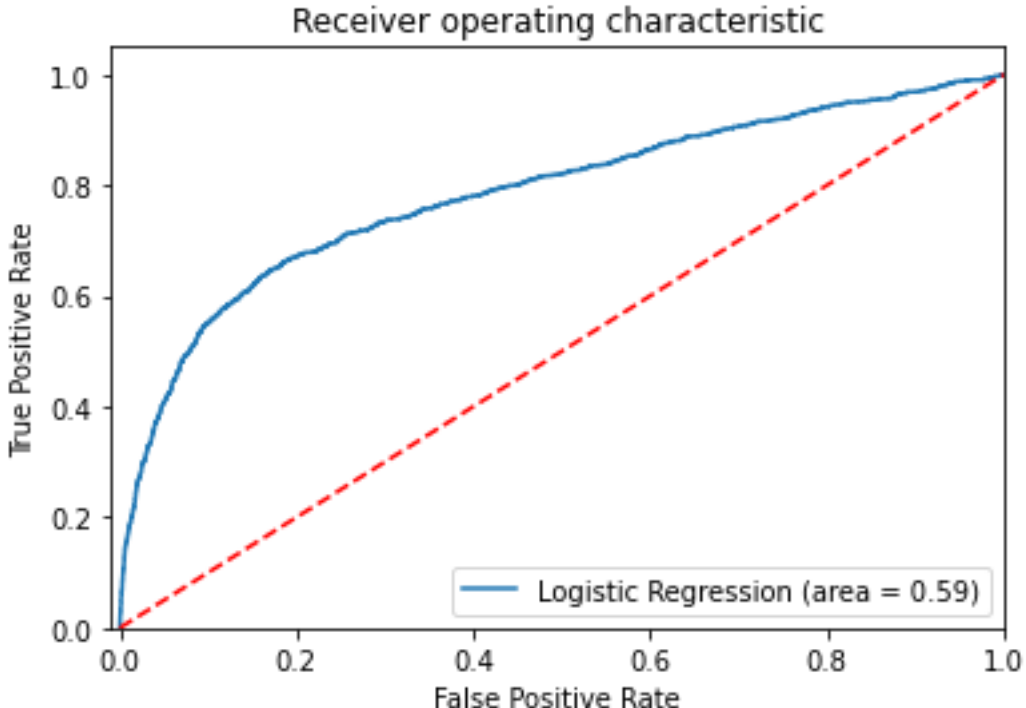
```
Confusion Matrix:
 [[10780   149]
 [ 1138   286]]
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.99      0.94     10929
           1       0.66      0.20      0.31      1424

    accuracy                           0.90     12353
   macro avg       0.78      0.59      0.63     12353
weighted avg       0.88      0.90      0.87     12353
```

AUC-ROC curve

- 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.



Thank You