Final Project Report

Group Name: Pattern Pros

Data Science: Bank Marketing Campaign

Github Repository Link:
https://github.com/slynnsin/Data-Glacier-Group-Project/tree/main/Final%20Project%20Report%20and%20Code

Team Member Details:

Name	Email-ID	Country	University	Specialization
Jay Panara	jay.panara@gmail.com	Canada	University of Waterloo	Data Science
Shreya Dwivedi	shreyad@usc.edu	USA	University of Southern California	Data Science
Sarah Sindeband	ssindeband2018@fau.edu	USA	Florida Atlantic University	Data Science
Daniel Kingswood	ddk727@gmail.com	UK	University of Bristol	Data Science

Problem Description:

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps 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).

Business Understanding:

The bank wants to use the ML model to shortlist customers whose chance of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing, etc.) can focus only on those customers who have a greater chance of buying the product.
This will save resources and their time (which is directly involved in the cost (resource billing)).
We need to develop a model with duration and without duration features and report the performance of the model.
The data is related to direct marketing campaigns 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 assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
The classification goal is to predict if the client will subscribe (yes/no) to a term deposit (variable y).
Attribute Information:

Input variables:
# bank client data:
1 - age (numeric)
2 - job: type of job (categorical:
'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed', 'services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical:
'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical:
'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course', 'university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has a housing loan? (categorical: 'no','yes','unknown')

7 - loan: has a personal loan? (categorical: 'no','yes','unknown')

# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - 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.

# other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - 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)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
# social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Project Lifecycle:

Week	Task	Deadline
Week 7	Business Understanding	19th October, 2022
Week 8	Data Understanding	26th October, 2022
Week 9	Data Cleansing and Transformation	2nd November, 2022
Week 10	EDA with Recommendation	9th November, 2022
Week 11	Presentation of EDA for business and recommended models for technical user	16th November, 2022
Week 12	Model selection and building with dashboard	23rd November, 2022
Week 13	Final Project submission and report with powerpoint	30th November, 2022

Data Intake Report

Name: Bank Marketing Campaign
Report date: 10/17/2022
Internship Batch: LISUM13: 30
Version: 1.1
Data intake by: Sarah Sindeband
Data intake reviewer:<Jay Panara>
Data storage location:
https://github.com/slynnsin/Data-Glacier-Group-Project/tree/main/Final%20Project%20Report%20and%20Code

Tabular data details: bank.csv

Total number of observations	45211
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	450 KB

Tabular data details: bank-full.csv

Total number of observations	4521
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	439 MB

Data Understanding:

Data Source:

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

bank-full.csv:

This data set is ordered by date and is broken down into four sections. The first section has columns pertaining to the specific client. The second section of columns has information related to the last contact of the current campaign. The third section of columns have information pertaining to previous marketing campaigns. The final column is whether or not the client subscribed to the term deposit (output variable).

bank.csv:

This file is 10% randomly selected from the bank_full.csv file set aside for testing a machine learning model.

Data type for analysis:

Column names	Data type
age	numeric
job	categorical
martial	categorical
education	categorical
default	binary
balance	numeric
housing	binary
loan	binary
contact	categorical
day	numeric
month	categorical
duration	numeric
campaign	numeric
pdays	numeric
previous	numeric
poutcome	categorical
y	binary

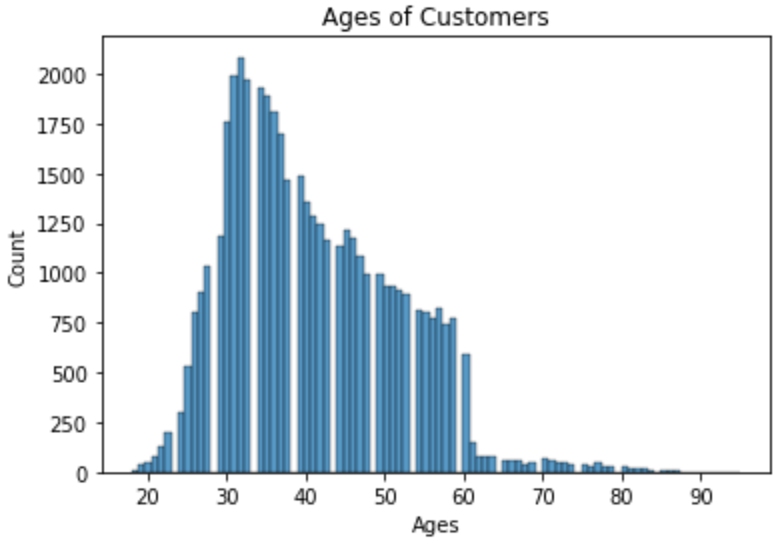
Problems in data:

- 5 columns have skewed values (balance, duration, campaign, pdays and previous)
- 4 columns have unknown values, 2 with large proportion (job, education, contact, poutcome)
- Outliers
- No client ID number

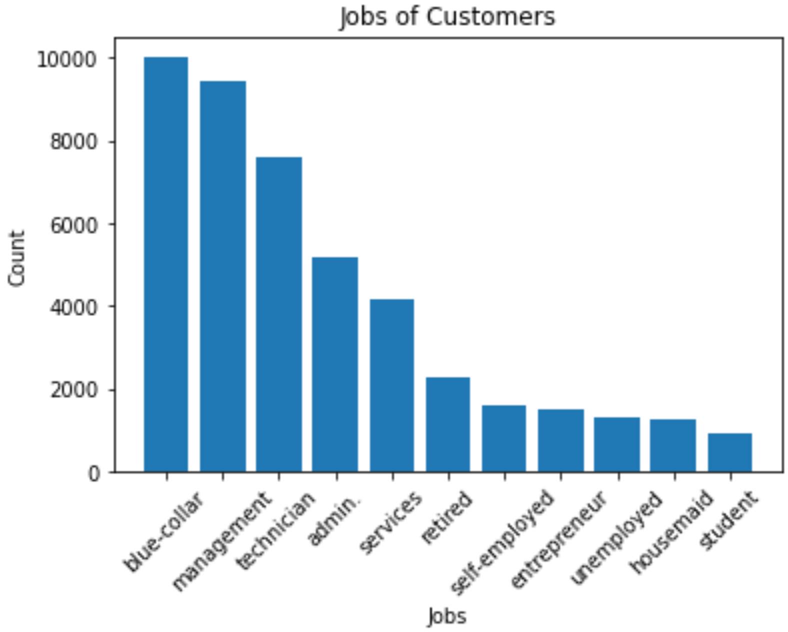
Approaches:

Problem	**Approaches**	**Why?**
Skewed values/ imbalanced dataset	Remove outliers, log transformation, normalize values to help balance the data.	The Tail region can act as an outlier for regression based models and cause a bias in the model.
Unknown values	For columns with a small amount of unknown values they could be removed. For columns with a large amount of unknown values, imputation methods could be used or predict the missing values using either a regression or classification model.	Missing data or unknown values can lead to a reduced size of the data which leads to less efficient estimates from the model. Also if the values are left as unknown the model could find patterns between unknown values which is not helpful for accurate predictions.
No client identification number	Check for duplicate entries. This can be done by comparing each line, and if the line is exactly the same as another, it could be a duplicate entry.	To avoid duplicates
Outliers	Remove outliers, check if outliers are logical, or do further statistical tests to verify the outliers	Outliers can cause a decrease in normality(skewed data), cause a bias in models, have a significant impact on mean and standard deviation of data and can also cause problems during statistical analysis

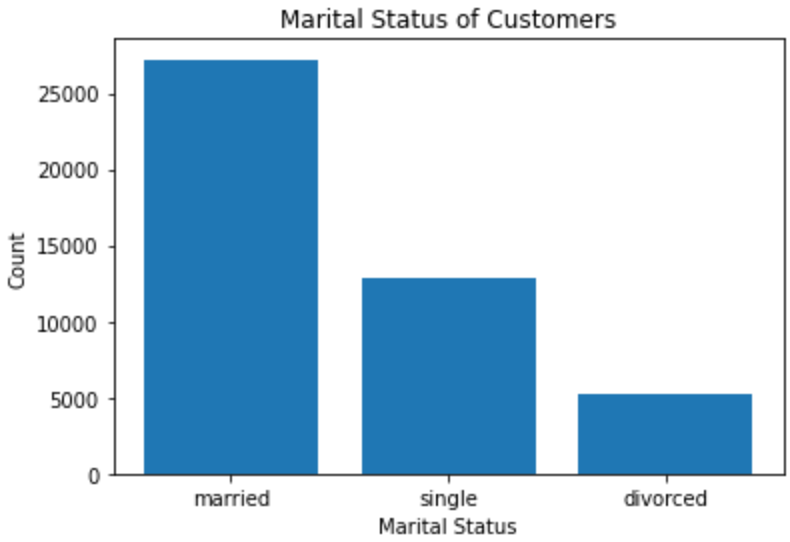
EDA Analysis:



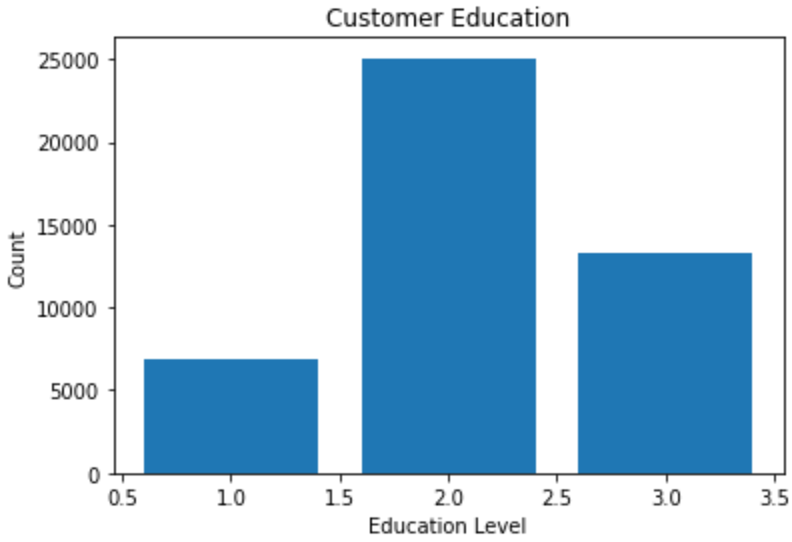
The distribution of the age of the customers show that there are more individuals who are present in their early 30s.



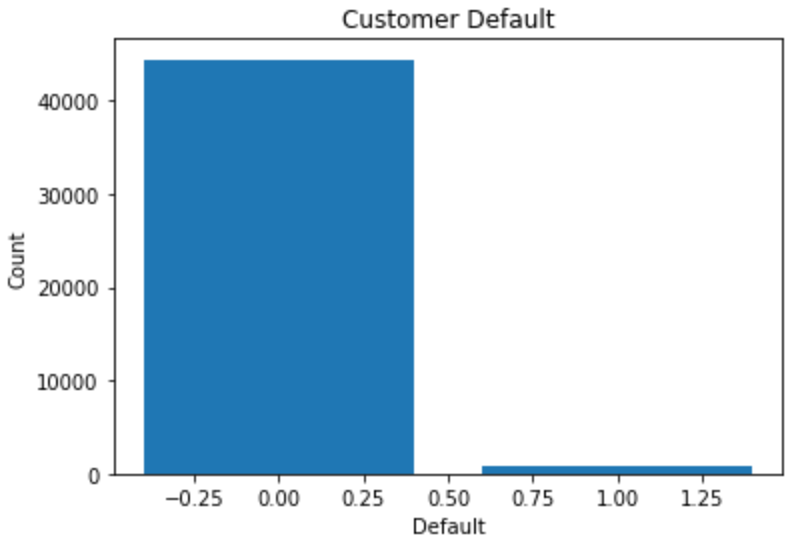
The customers with the blue-collar jobs were the highest in the data.



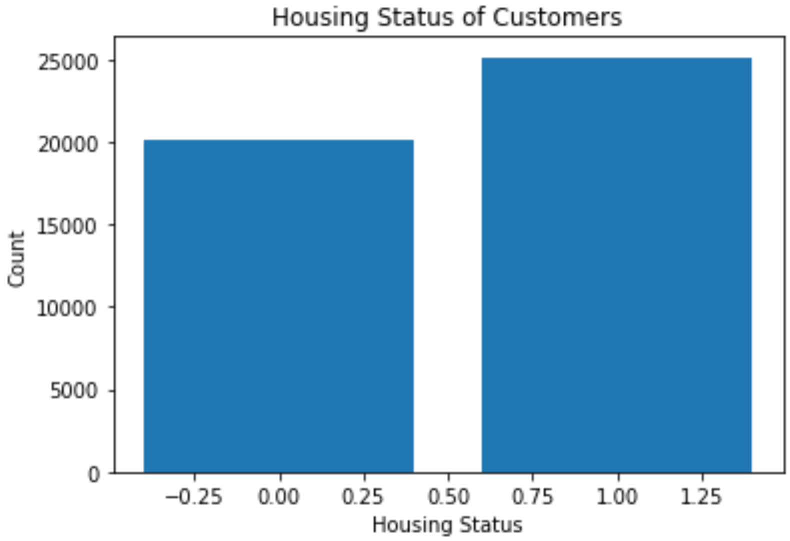
There are more customers who are married compared to single and divorced.



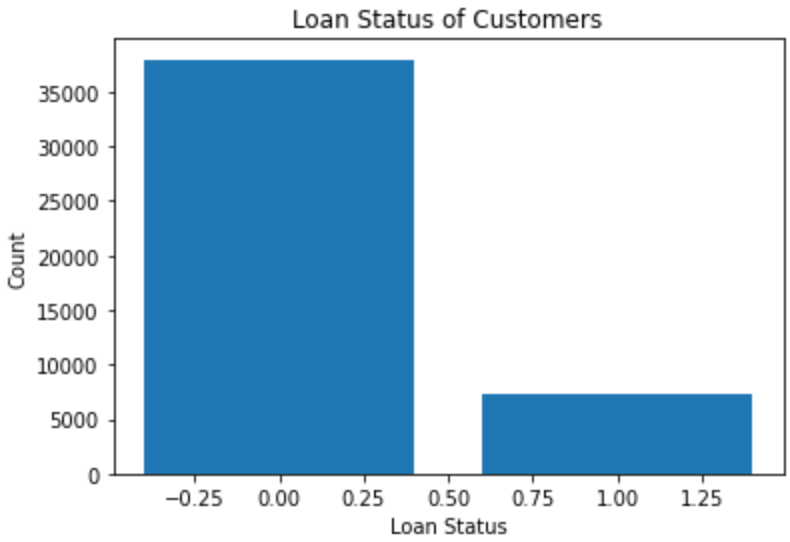
Most of the individuals have a secondary level of education.



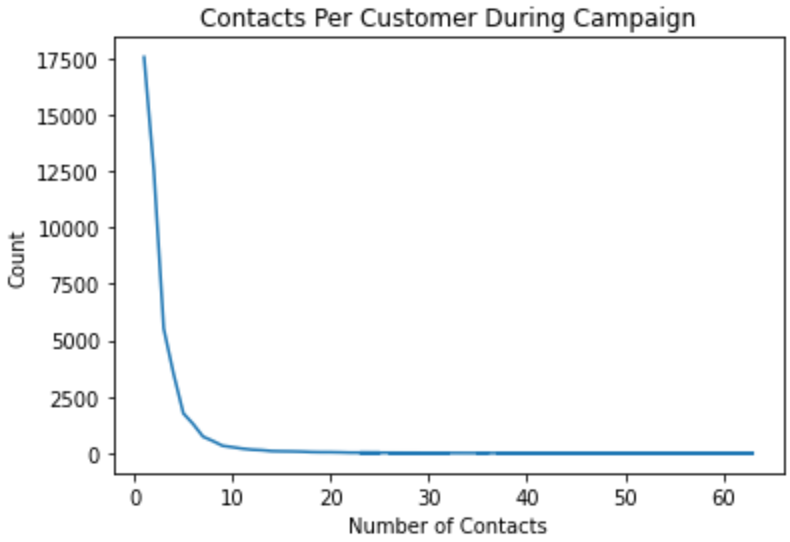
Almost all of the customers do not have their credit default.



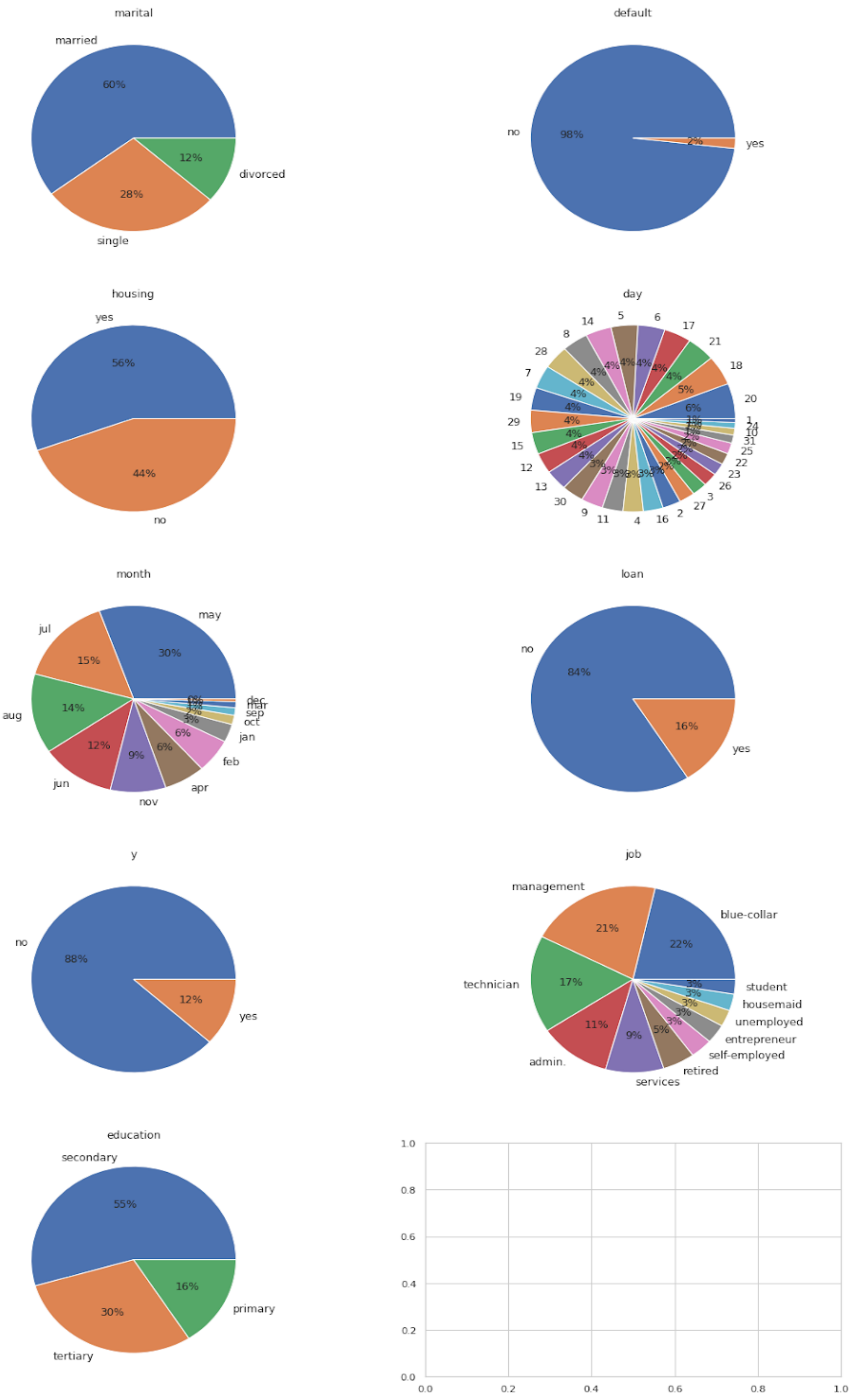
A lot of people have housing loans.



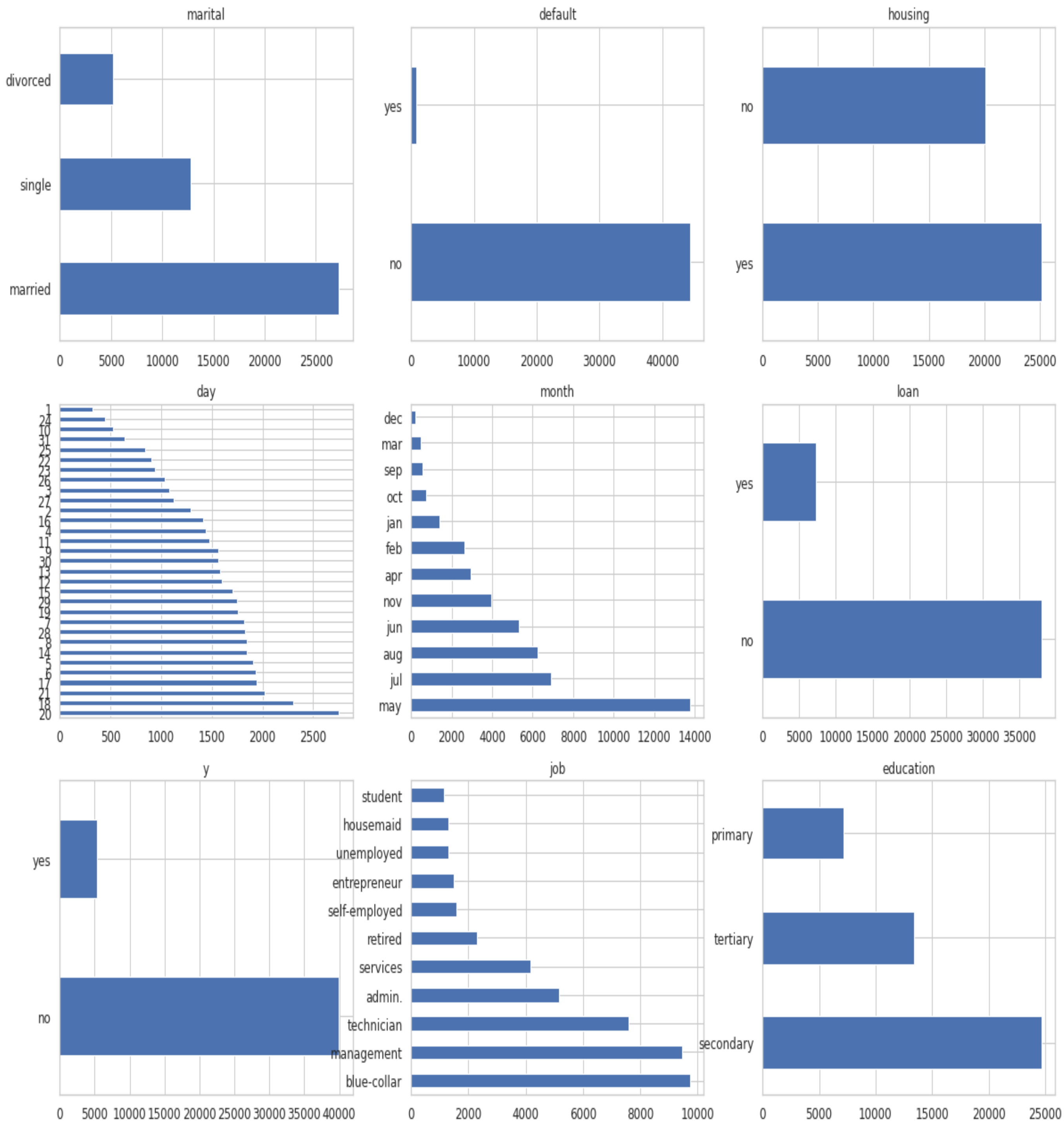
Many customers do not have any loans.



This graph shows the number of contacts that have been made during the campaign.

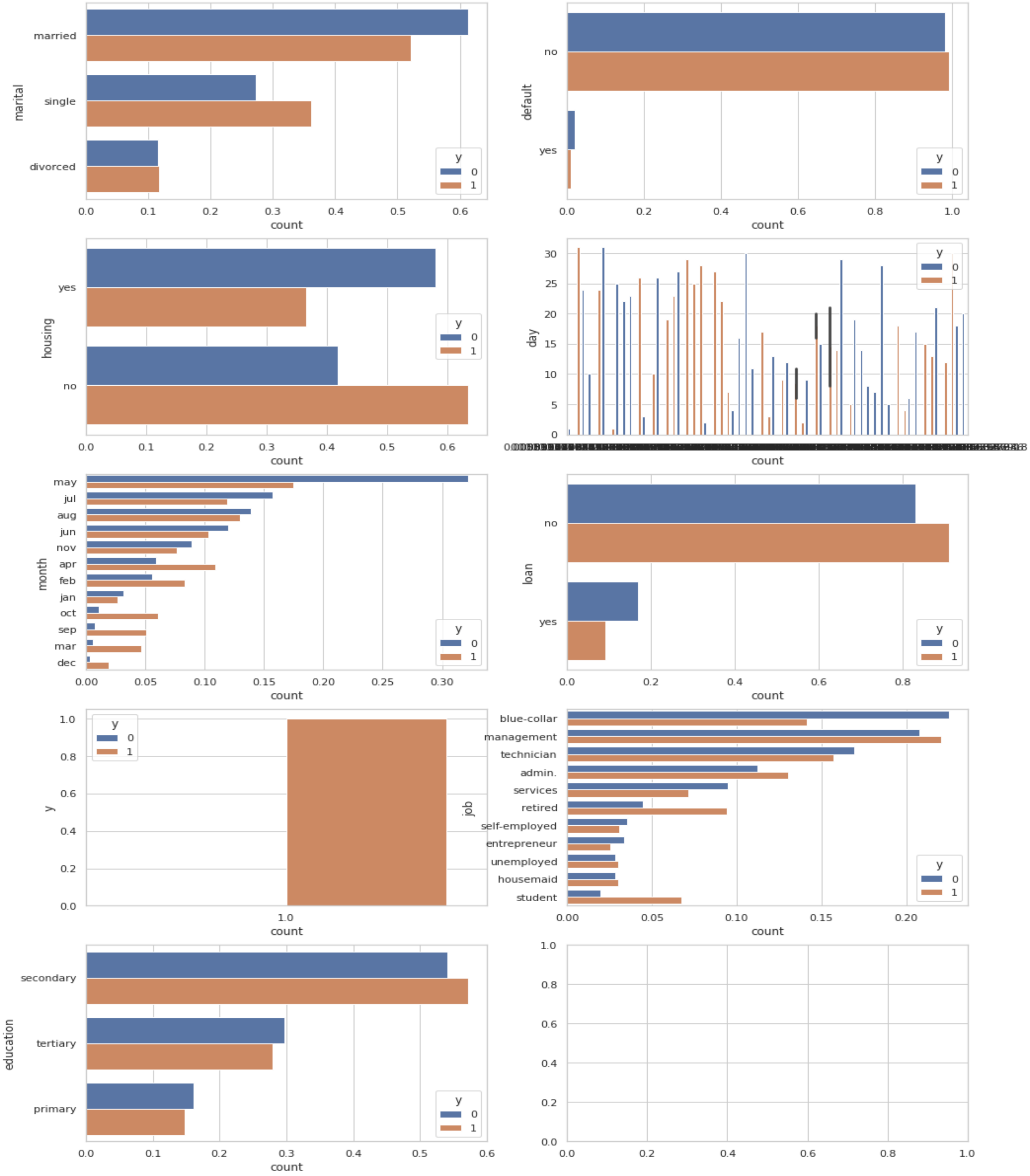

marital
married
60%
28%
12%
divorced
single
default
no
98%
2%
yes
housing
yes
56%
44%
no
day
5
6
17
21
18
20
14
8
28
7
19
29
15
12
13
30
9
11
4
16
2
27
3
26
23
22
25
31
10
24
1
6%
5%
4%
3%
2%
1%
month
may
30%
jul
15%
aug
14%
12%
jun
9%
nov
6%
apr
6%
feb
3%
jan
2%
oct
sep
mar
dec
0%
1%
loan
no
84%
16%
yes
y
no
88%
12%
yes
job
management
21%
blue-collar
22%
technician
17%
11%
9%
5%
3%
student
housemaid
unemployed
entrepreneur
self-employed
retired
services
admin.
education
secondary
55%
30%
16%
primary
tertiary
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

- 60% of the individuals are married and there are 12% divorced individuals.
- The default has a mere 2% ‘Yes’ and the remaining 98% as ‘No’.
- 56% of the individuals have a housing loan..
- The last contact day is evenly distributed as seen in the pie chart.
- The number of people that have been contacted in the month of May is highest with 30% followed by the month of july.
- 84% of people do not have a personal loan while 16% of individuals do not have a personal loan.
- Most of the individuals in the dataset have a blue collar percentage of 22% which is followed by management with 21%. The technicians take the third spot with 17% and the remaining jobs are distributed with every percentage.
- Most of the individuals have secondary education and constitute 55% while 30 % have advanced education and the remaining 16% have primary education.

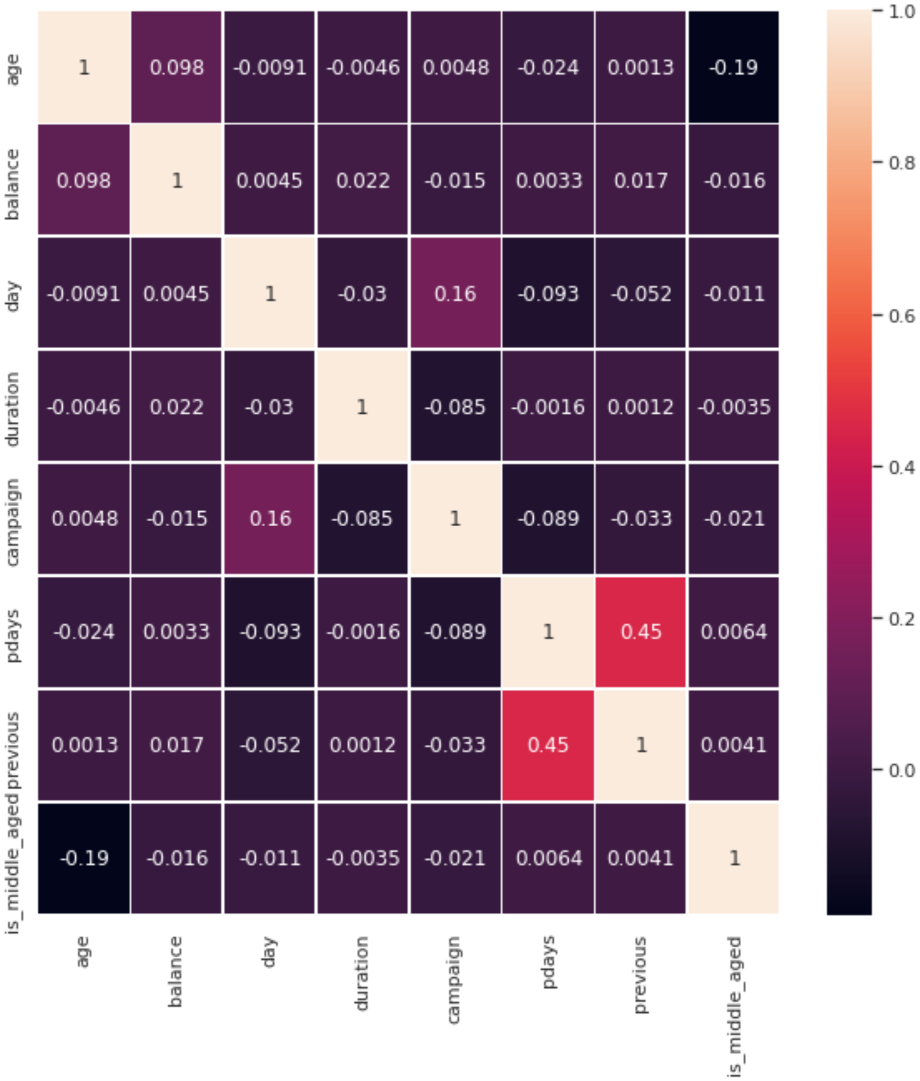


- Less number of students and more number of management and technician customers
- Most of married customers
- Most customers education levels is secondary
- Most customers are not defaulted in past

- More than 50% have taken housing loan
- Nearly 85% have taken personal loan
- Major communication type is cellular
- Most of the customers were last contacted in the month of May
- Most customers where not contacted in previous month


marital
married
single
divorced
count
0.0
0.1
0.2
0.3
0.4
0.5
0.6
y
0
1
default
no
yes
0.0
0.2
0.4
0.6
0.8
1.0
housing
yes
no
day
30
25
20
15
10
5
0
month
may
jul
aug
jun
nov
apr
feb
jan
oct
sep
mar
dec
0.00
0.05
0.10
0.15
0.20
0.25
0.30
loan
no
yes
y
1.0
0.8
0.6
0.4
0.2
0.0
job
blue-collar
management
technician
admin.
services
retired
self-employed
entrepreneur
unemployed
housemaid
student
0.00
0.05
0.10
0.15
0.20
education
secondary
tertiary
primary

- Management, retire, self-employed, unemployed and students tend to subscribe more
- Singles subscribe more than married and divorced
- Customers with tertiary level of education will subscribe
- Customers with without housing and personal loan tend to subscribe to team deposit
- Customers approached by cellular communication have subscribed
- Subscription rate is more during start(jan,feb,march,apr) and end of the year(oct,sept,dec)
- Customers who subscribed during the previous campaign tend to subscribe more.



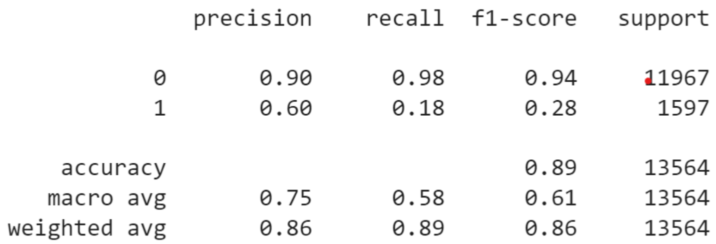
	age	balance	day	duration	campaign	pdays	previous	is_middle_aged
age	1	0.098	-0.0091	-0.0046	0.0048	-0.024	0.0013	-0.19
balance	0.098	1	0.0045	0.022	-0.015	0.0033	0.017	-0.016
day	-0.0091	0.0045	1	-0.03	0.16	-0.093	-0.052	-0.011
duration	-0.0046	0.022	-0.03	1	-0.085	-0.0016	0.0012	-0.0035
campaign	0.0048	-0.015	0.16	-0.085	1	-0.089	-0.033	-0.021
pdays	-0.024	0.0033	-0.093	-0.0016	-0.089	1	0.45	0.0064
previous	0.0013	0.017	-0.052	0.0012	-0.033	0.45	1	0.0041
is_middle_aged	-0.19	-0.016	-0.011	-0.0035	-0.021	0.0064	0.0041	1

From the correlation matrix, there seems to be strong correlation between the features p_days and previous. We can verify that using hypothesis tests.

Machine Learning Model Results

Columns with categorical values were converted to arbitrary numerical variables with One-Hot Encoding.

Logistic Regression

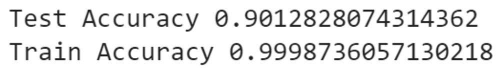


```
              precision    recall  f1-score   support

           0       0.90      0.98      0.94     11967
           1       0.60      0.18      0.28      1597

    accuracy                           0.89     13564
   macro avg       0.75      0.58      0.61     13564
weighted avg       0.86      0.89      0.86     13564
```

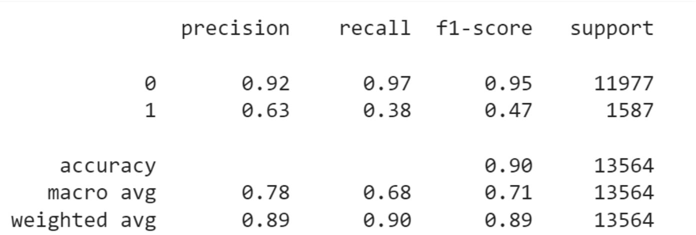
Accuracy: 89%

Random Forest

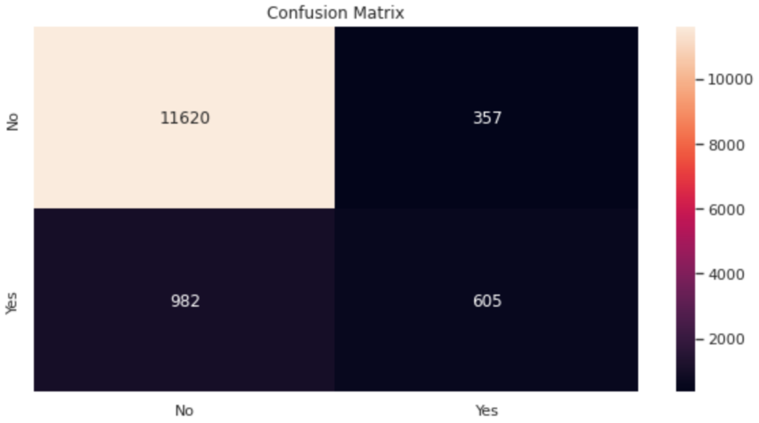
- Data was stratified when split for test and train data
- Test size = 30%
- n_estimators = 50 (number of trees in Random Forest model)



```
Test Accuracy 0.9012828074314362
Train Accuracy 0.9998736057130218
```



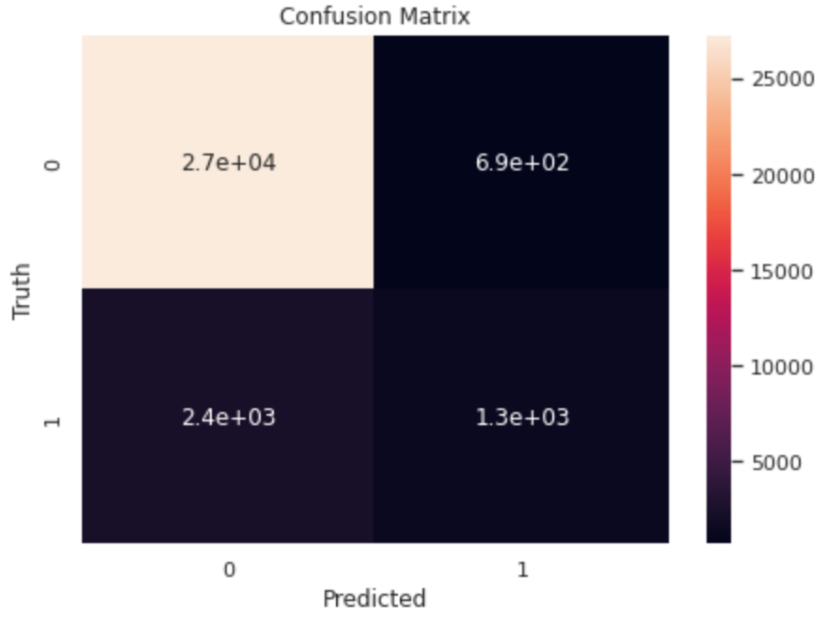
```
              precision    recall  f1-score   support

           0       0.92      0.97      0.95     11977
           1       0.63      0.38      0.47      1587

    accuracy                           0.90     13564
   macro avg       0.78      0.68      0.71     13564
weighted avg       0.89      0.90      0.89     13564
```



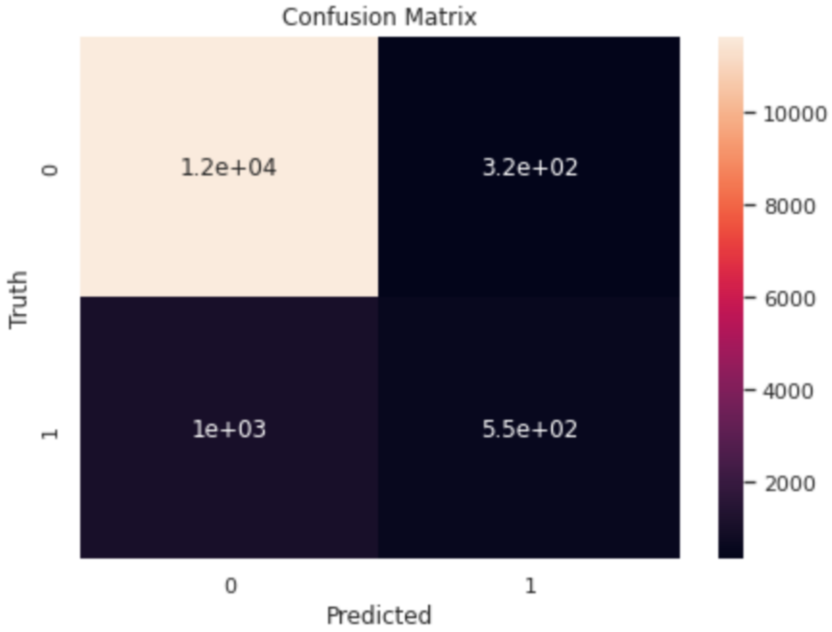
Boosting

- Test size = 30%
- alpha = 1; accuracy score = 0.8993

```
Train Confusion Matrix
```

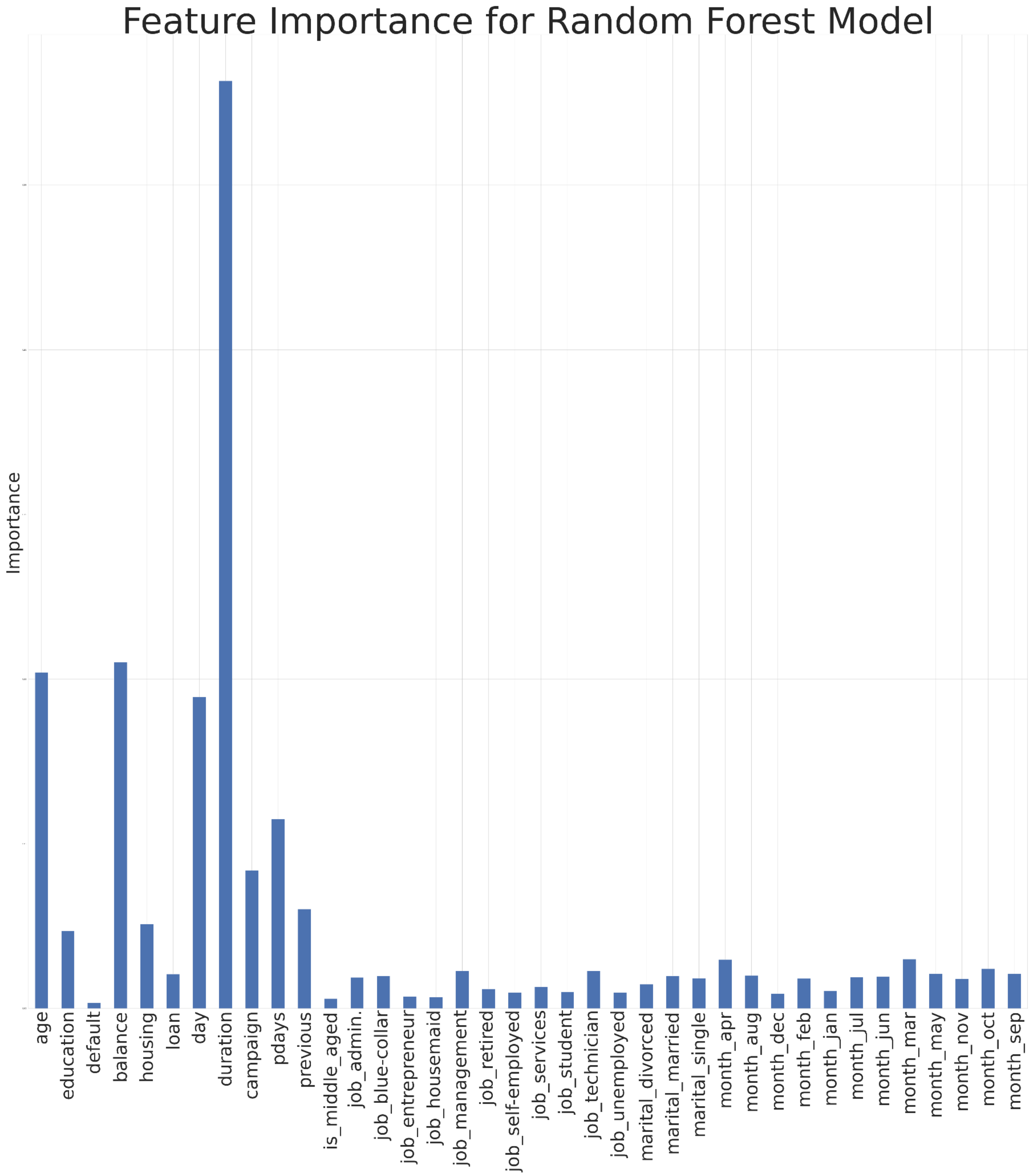


```
Test Confusion Matrix
```



Conclusion

Out of the 3 models tested, the random forest had the highest accuracy of 90% when used on the testing data. This model is recommended to help forecast outcomes in future marketing campaigns.



Based on the Feature Importance graph above, the duration of the contact seems to be the most important attribute in the model. This will be an important factor to consider when creating a strategy for the next marketing campaign.