Bank Marketing Campaign

Quicken Loans

Ly Nguyen
lilydq@yahoo.com

Content

- Problem Understanding
- Data Exploratory Analysis
- Feature Importance
- Model Building
- Hyper-parameter Tuning
- Model Evaluation
- Results

Problem Understanding

- European banking institution perform marketing campaigns based on phone calls to reach out to target customers in order to recruit them to Term Deposit subscriptions.

- The bank develop a prediction model to predict customers who are receptive to subscriptions.

- The campaigns were based on phone calls. Customers are contacted more than one time in order to access Term Deposit.

Dataset

- 41188 records
- 20 features
- From May 2008 – November 2010

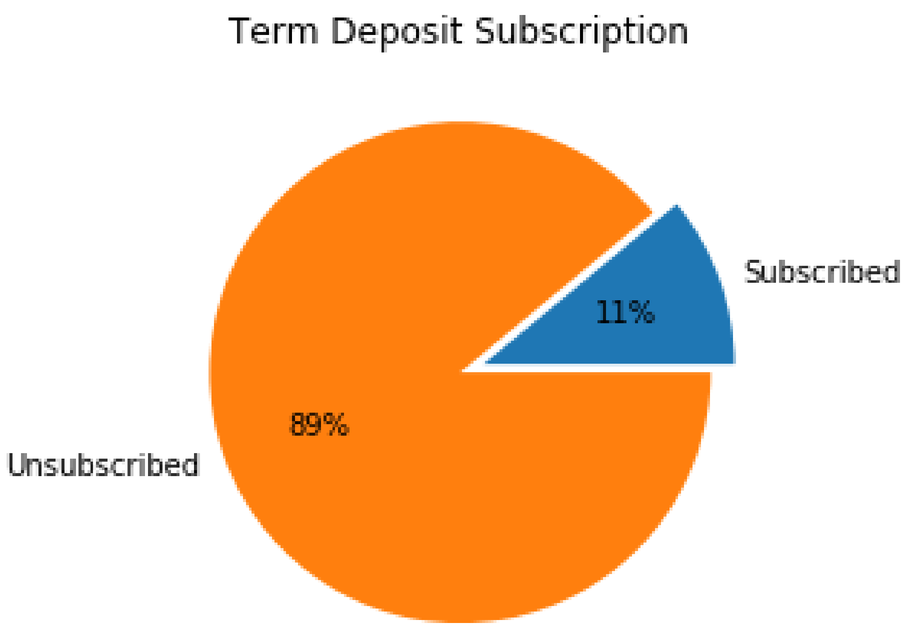
Data Exploratory Analysis

- Features are both categorical and numerical
  - Categorical features: 10
  - Numerical features: 9

- Target variable: binary (Yes/No)

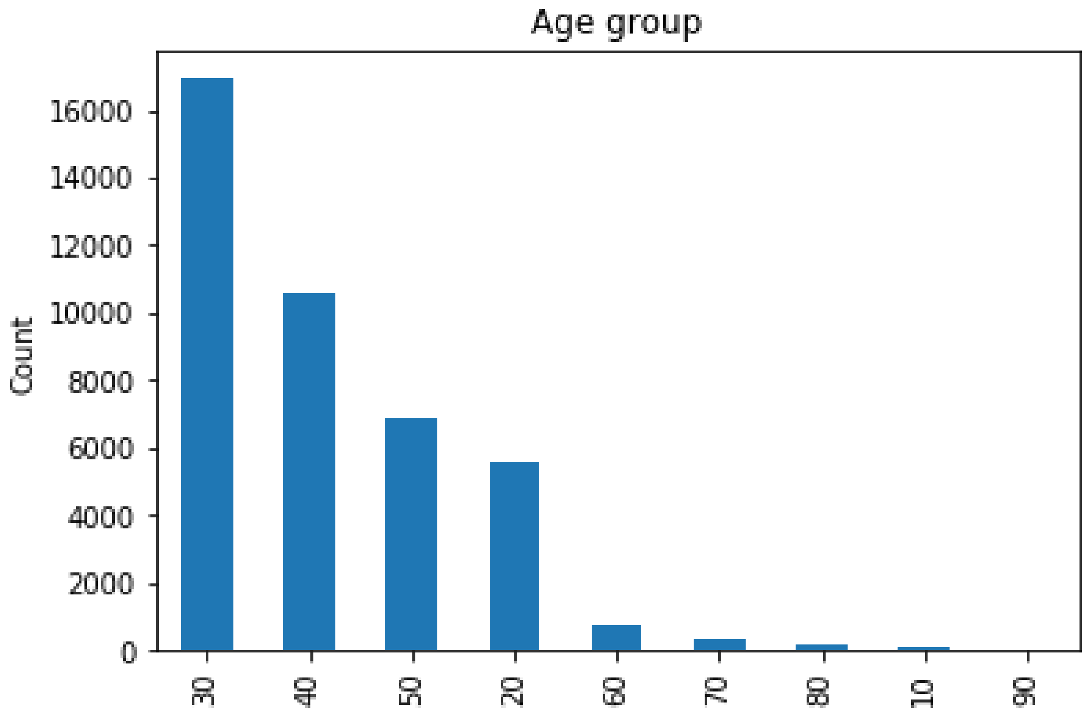
Data Exploratory Analysis

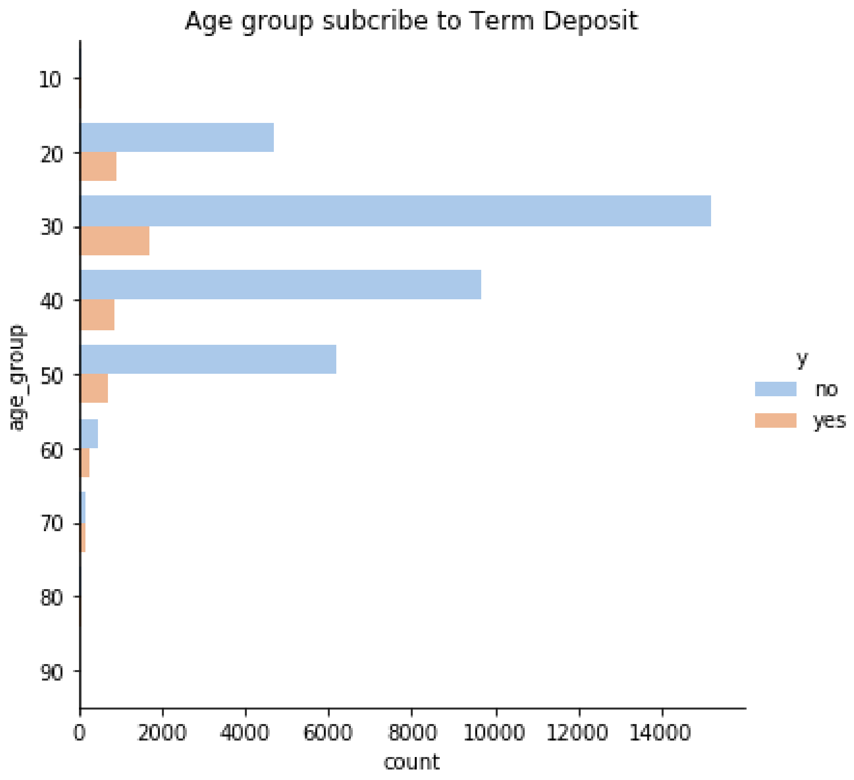
- **Has the client subscribed a term deposit? ('yes', 'no')**



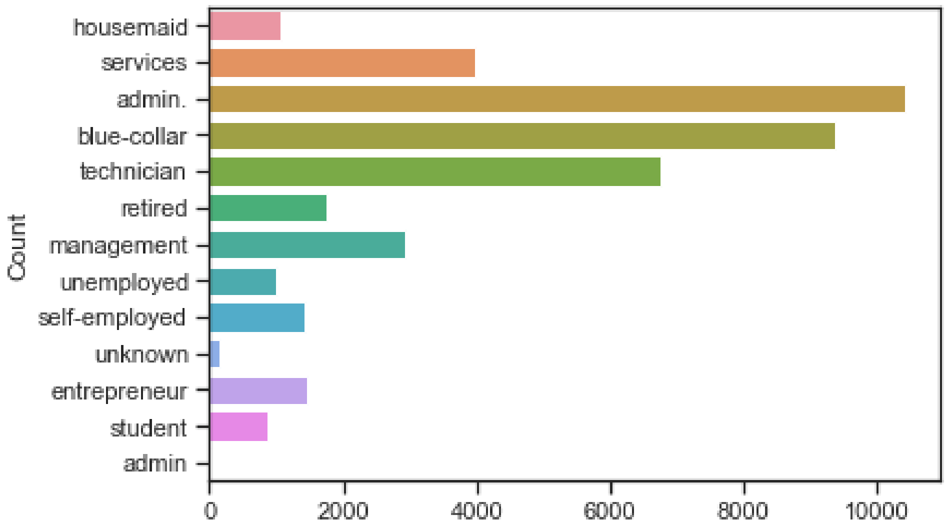
Data Exploratory Analysis

- **Customer's Age**

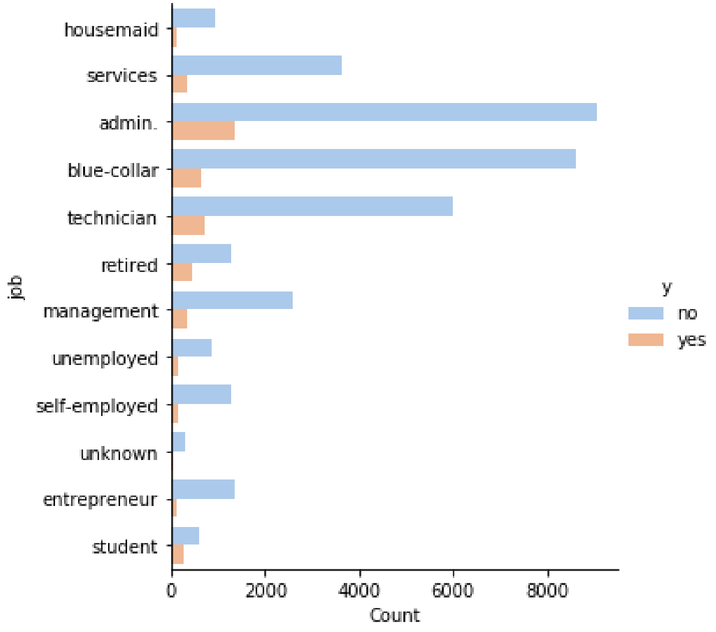


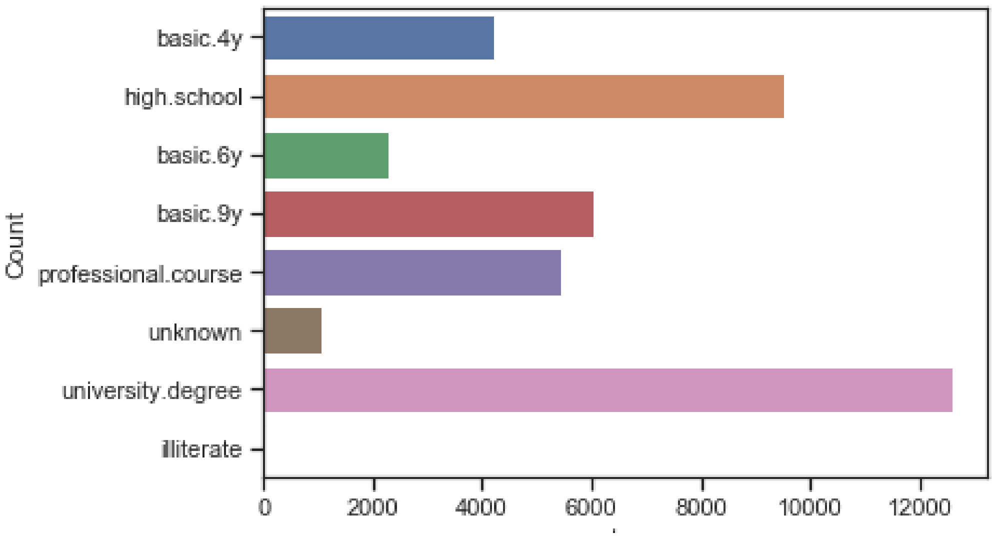


Data Exploratory Analysis

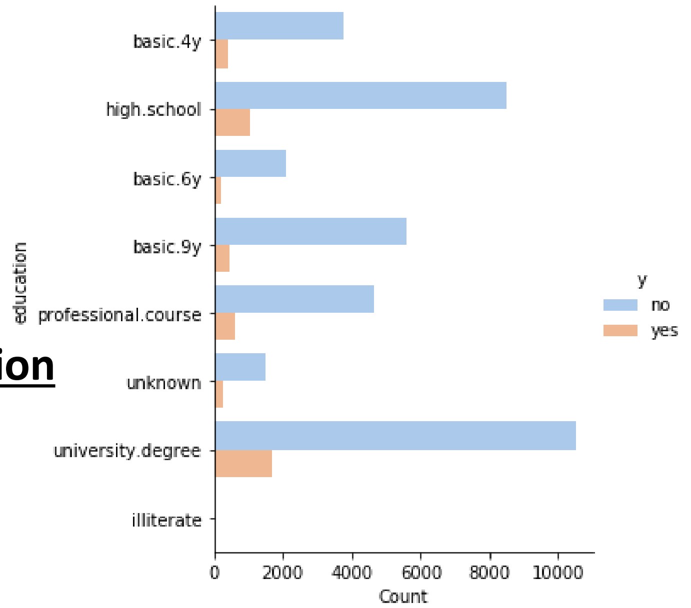


Job

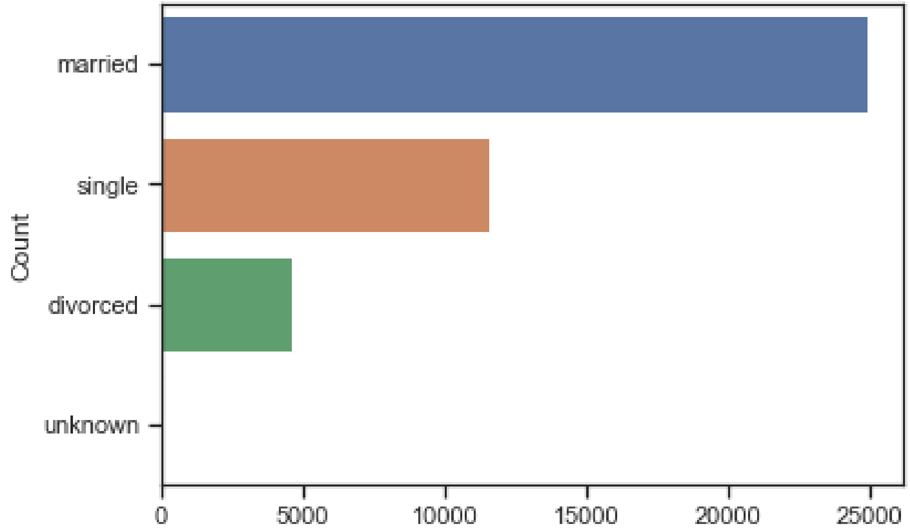




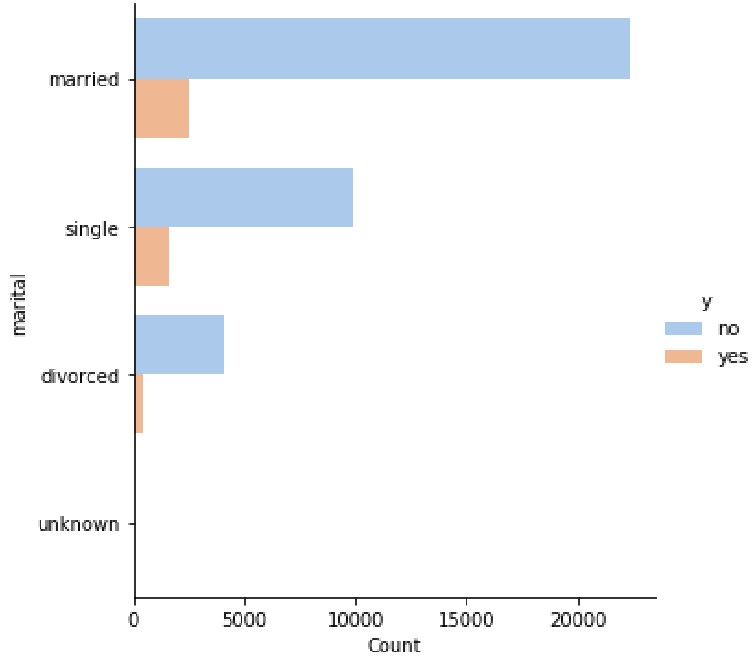
Education

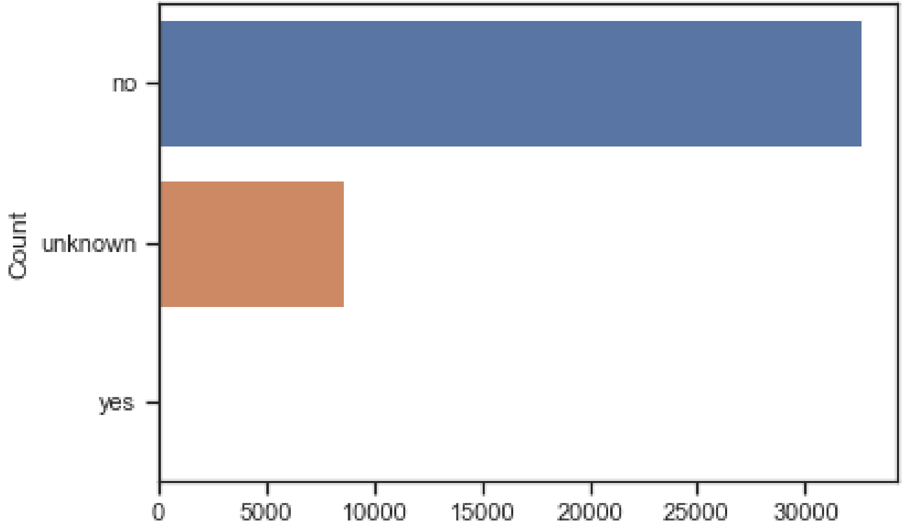


Data Exploratory Analysis

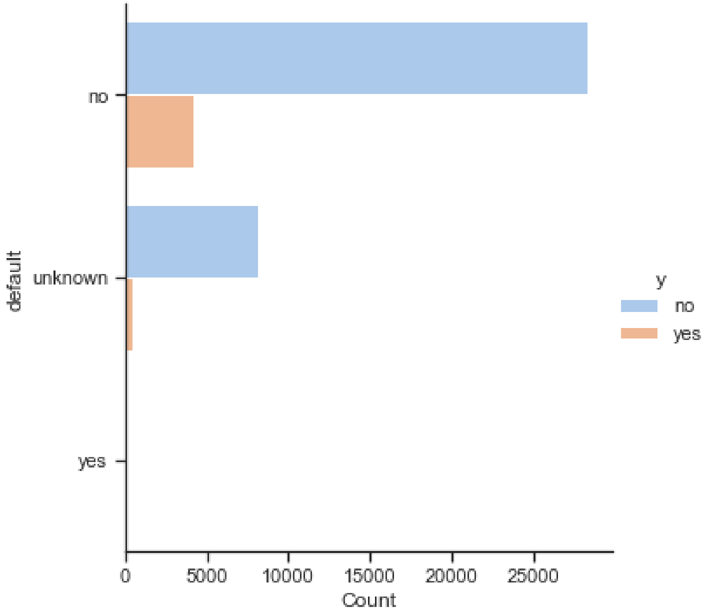


Marital

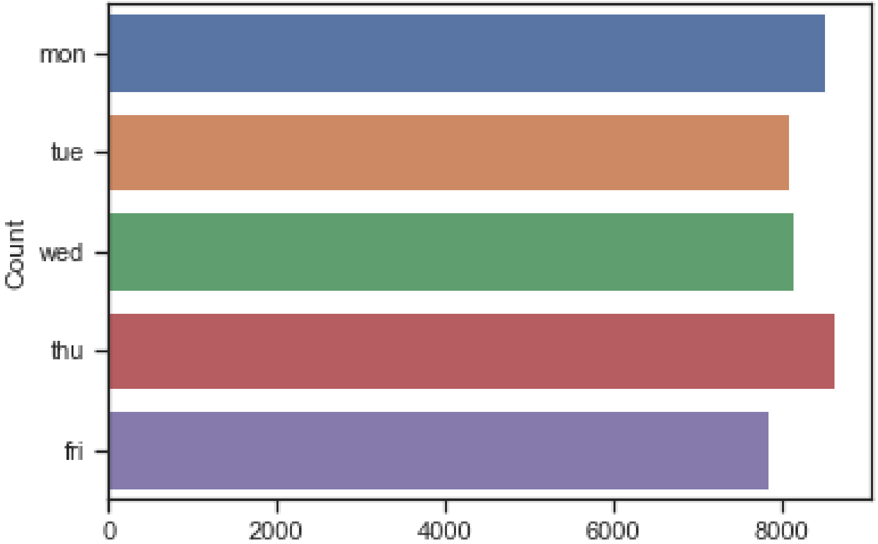




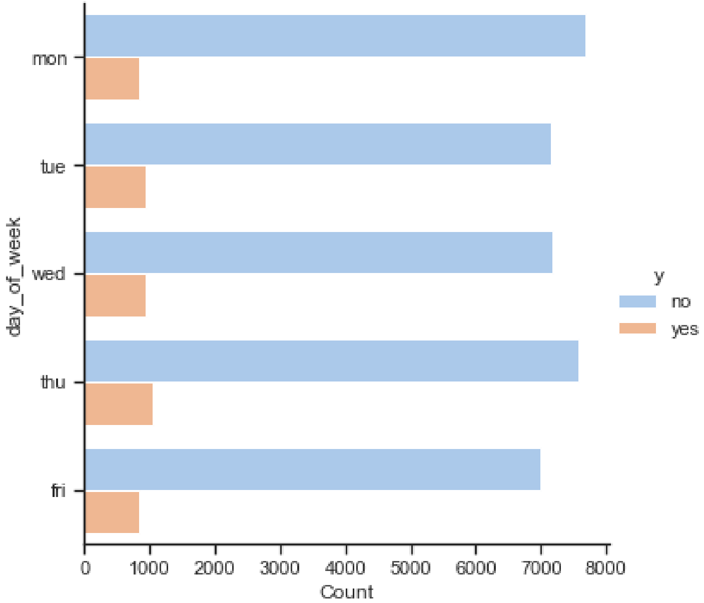
Default

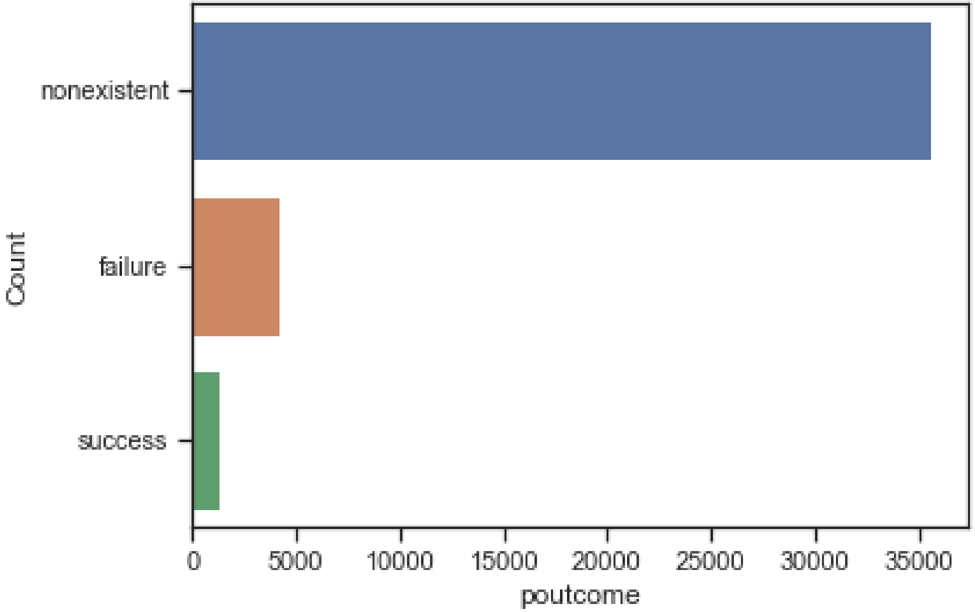


Data Exploratory Analysis

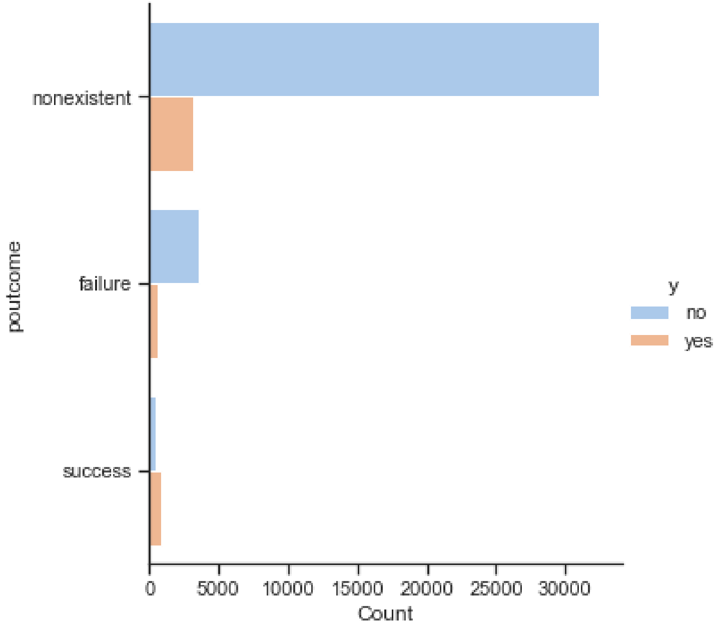


Day_of_week

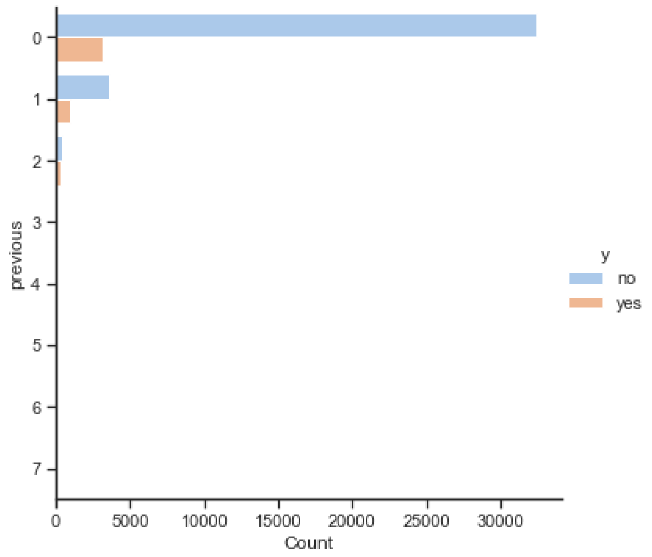


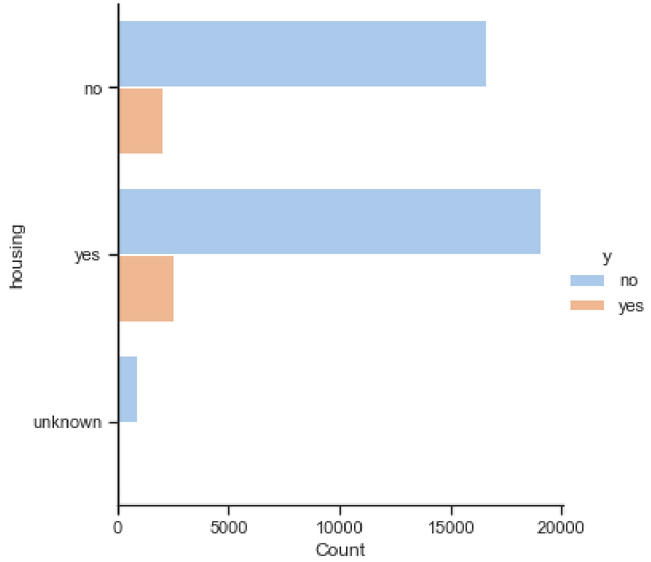


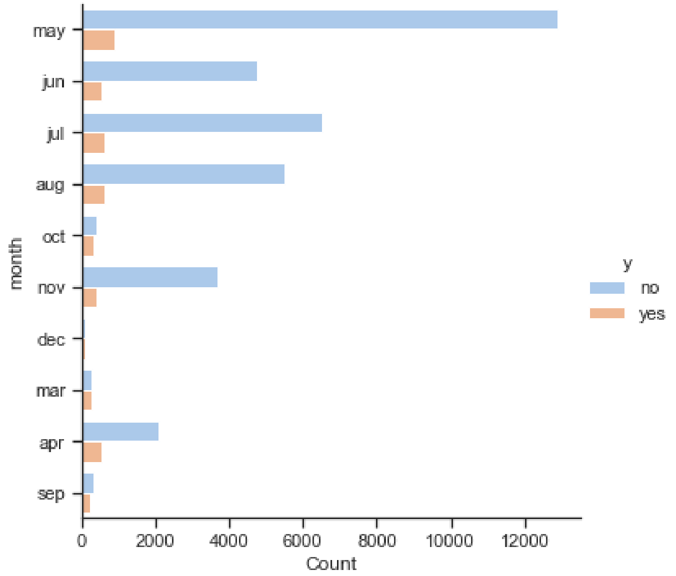
poutcome

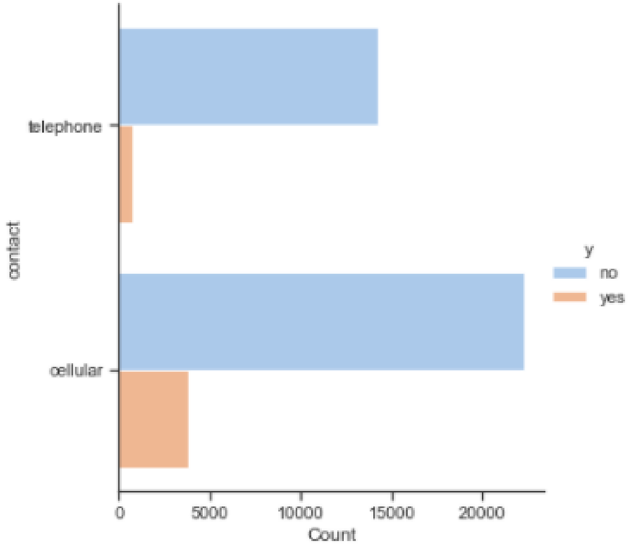


Data Exploratory Analysis

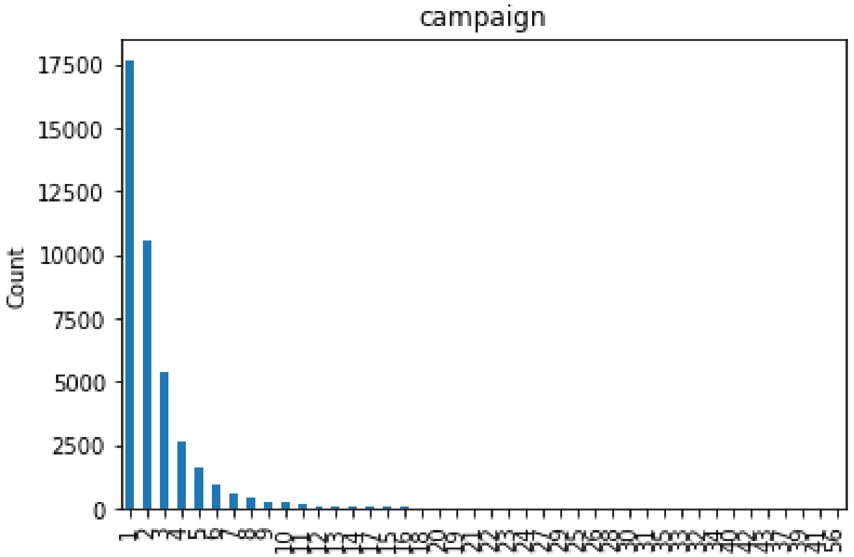


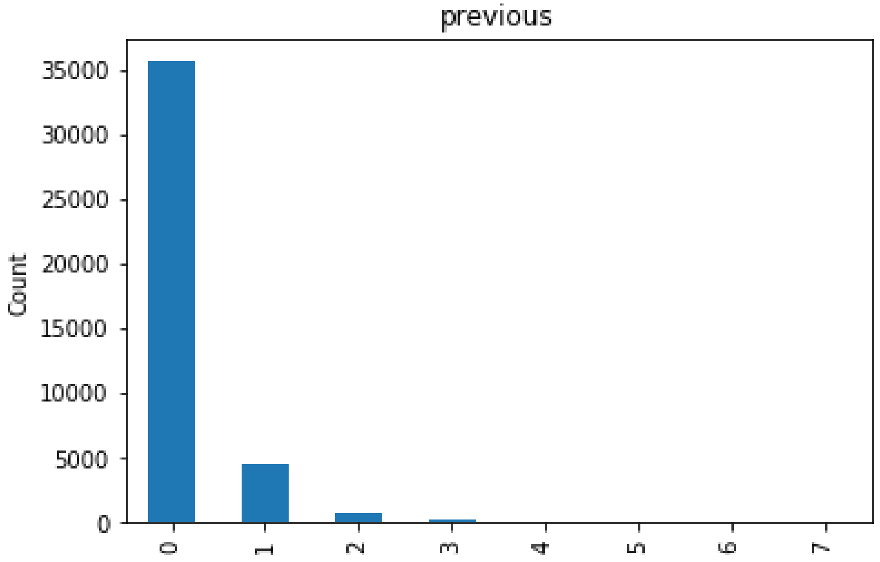


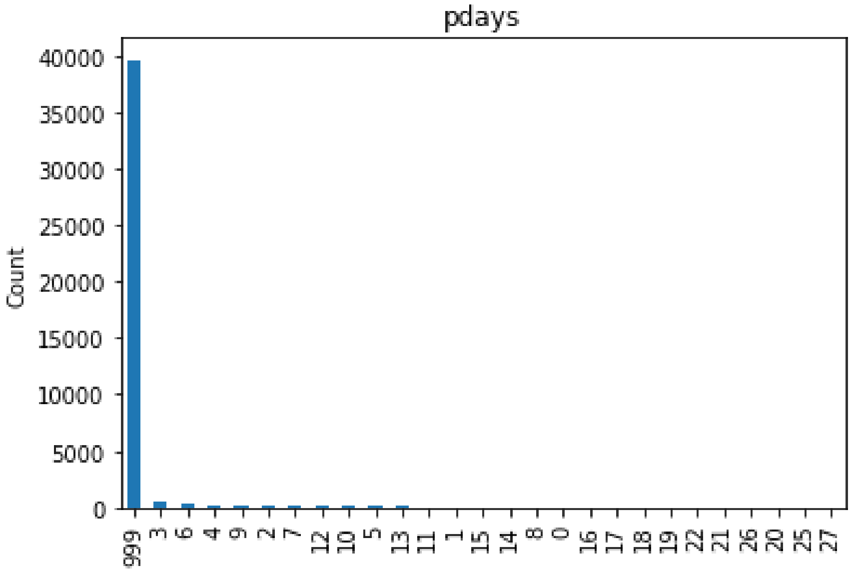


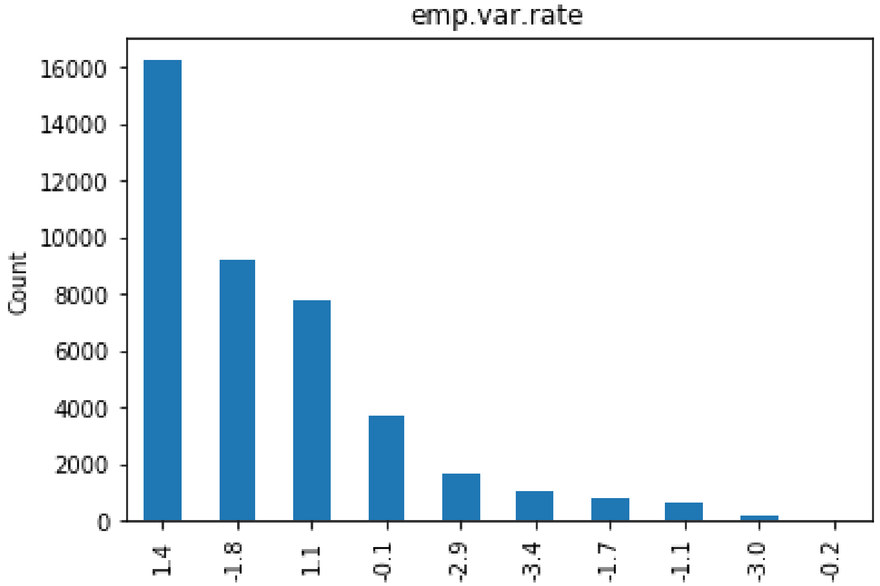


Data Exploratory Analysis

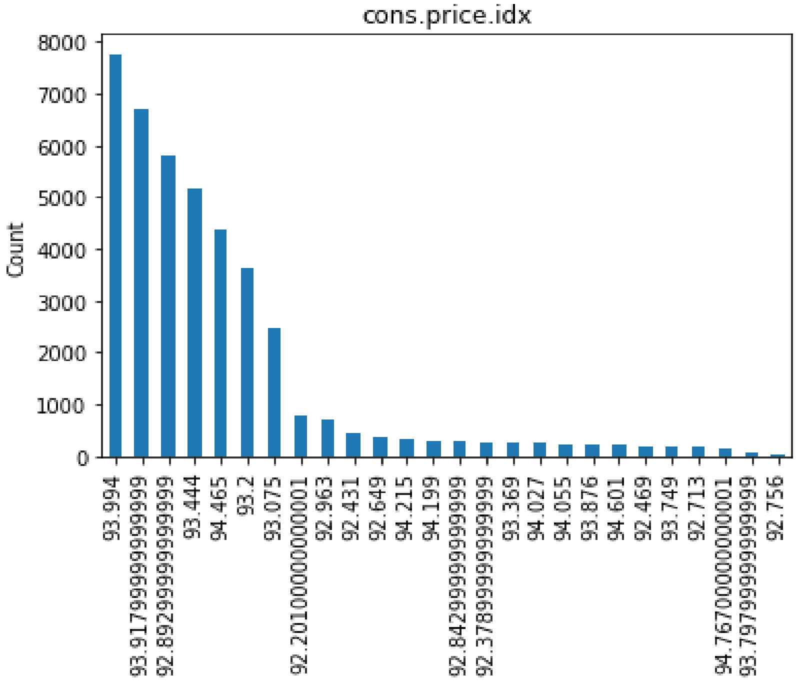


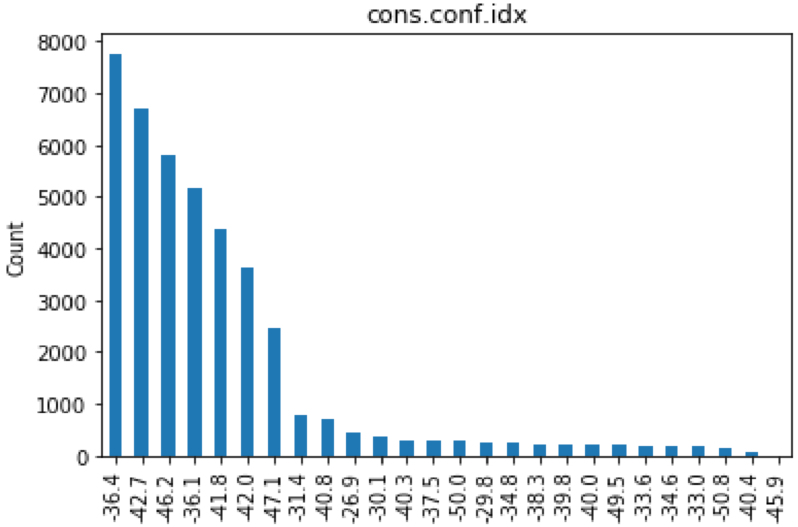


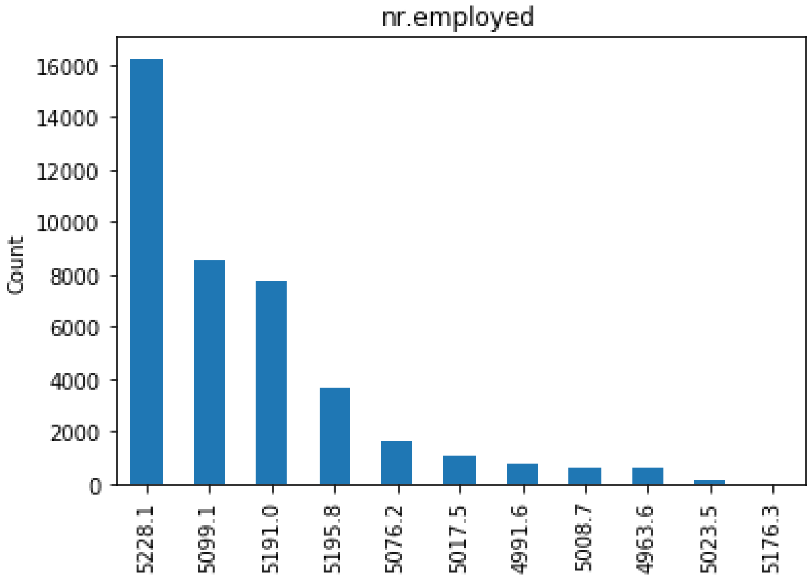




Data Exploratory Analysis







Data Exploratory Analysis

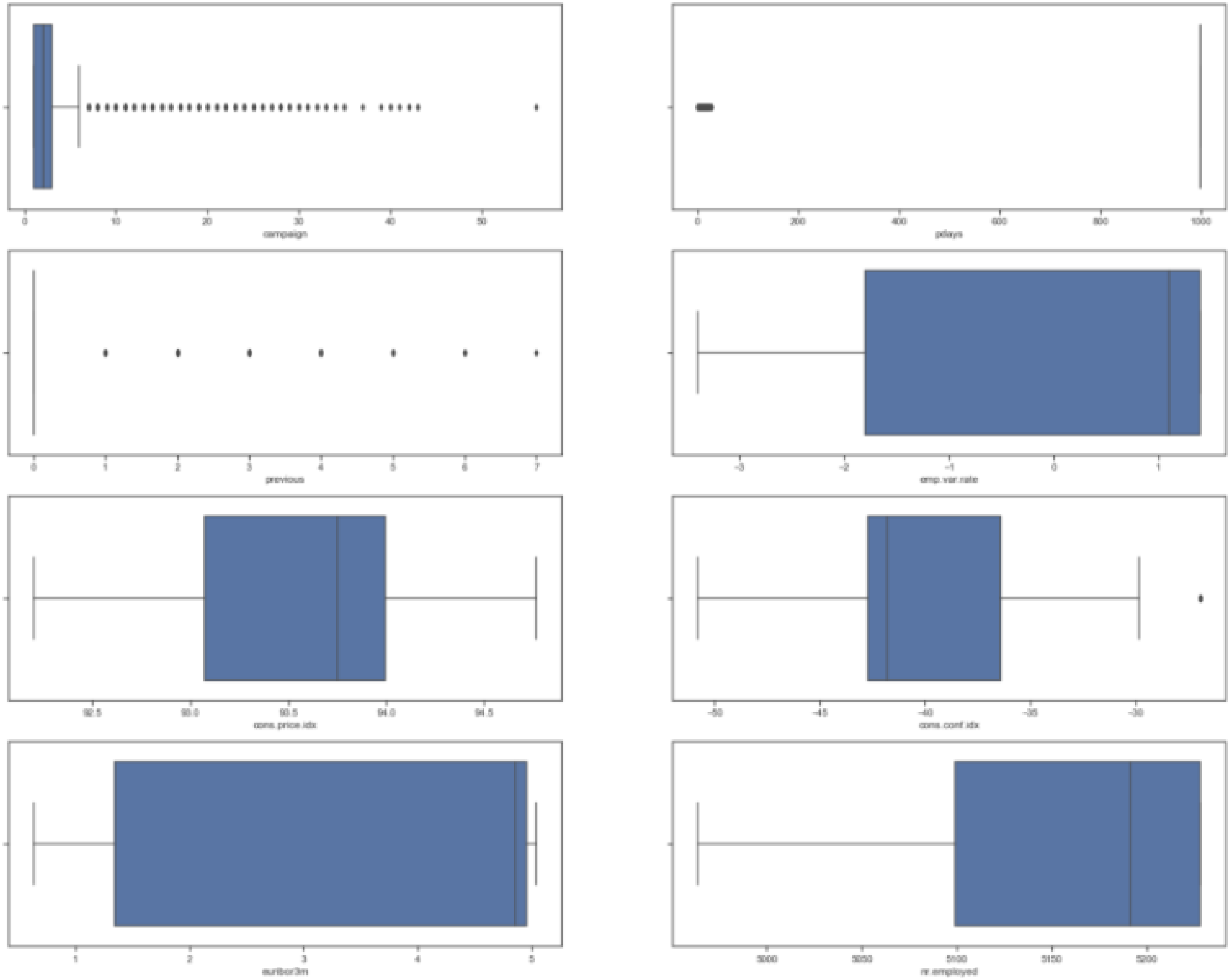
- Missing values:
  - There are no missing values (e.g. NA, null), but there are unknown values in 'job', 'marital', 'education', 'default', 'housing', 'loan'
  - Use inferences to create rules to impute the unknown data.
    - E.g. Age > 60 can be retired. The statistics of people who are over 60:
    - E.g. There will be a relationship between job and education. For example, 'admin', 'management', and 'technician' normally have university degree. Most of technicians have professional development.

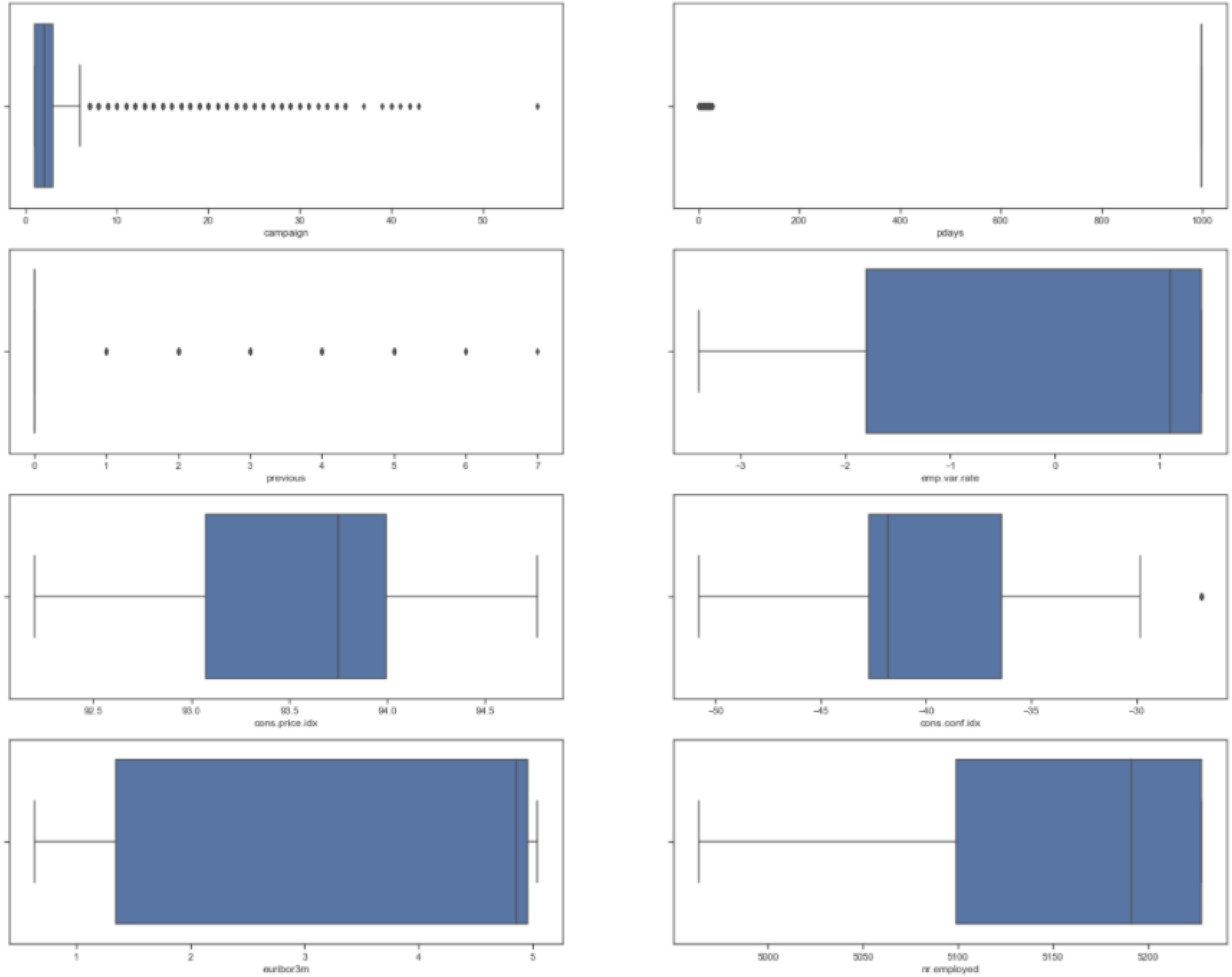
retired	785
admin.	106
housemaid	67
management	58
technician	49
blue-collar	43
unknown	29
self-employed	21
entrepreneur	17
unemployed	10
services	8

Data Exploratory Analysis

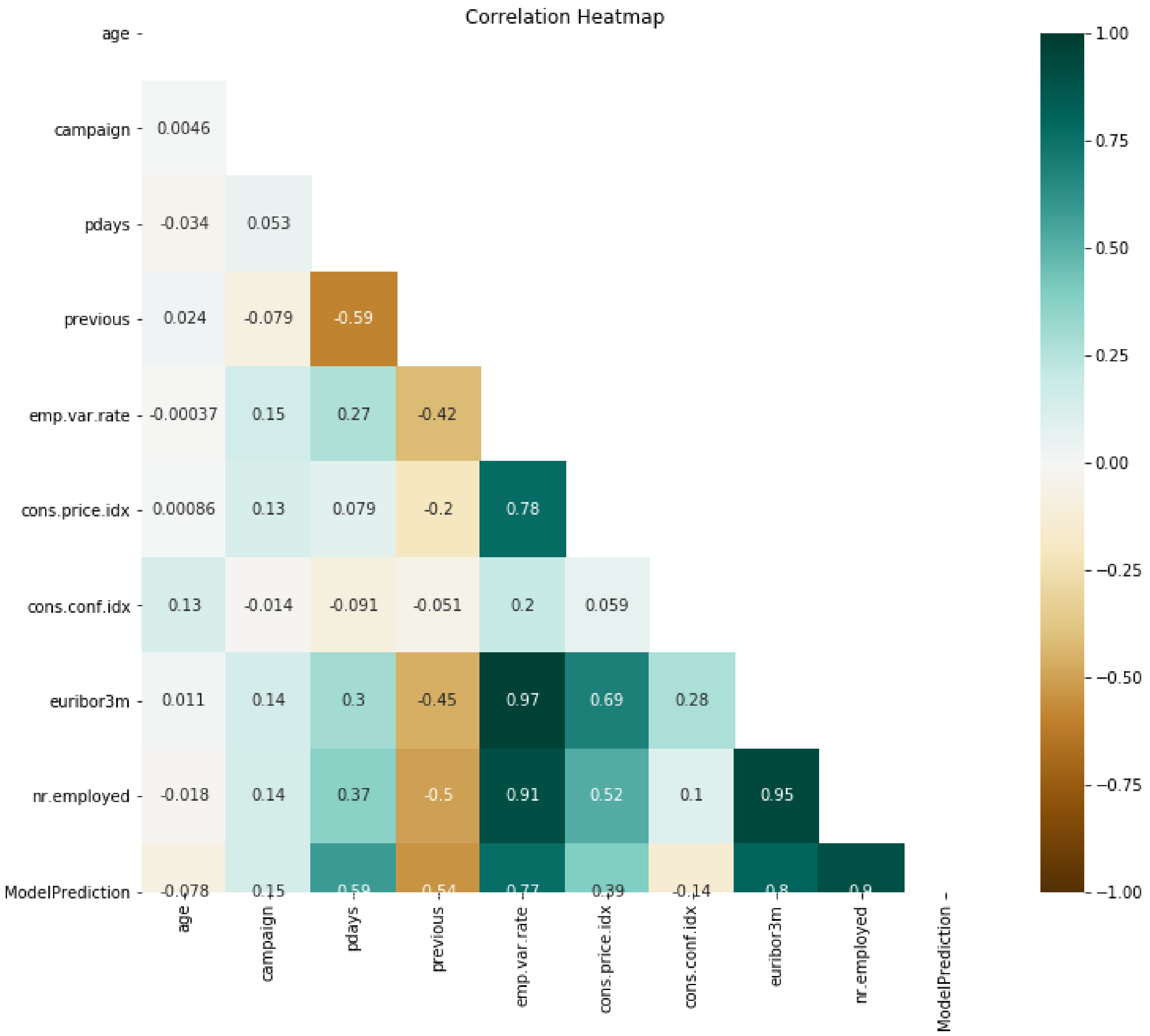
- Outliers:
  - we can see 'age', 'campaign' and 'previous' are the features that have outliers. However, the value are acceptable in the real world so we do not need to remove them.

	age	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000


campaign
pdays
previous
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed

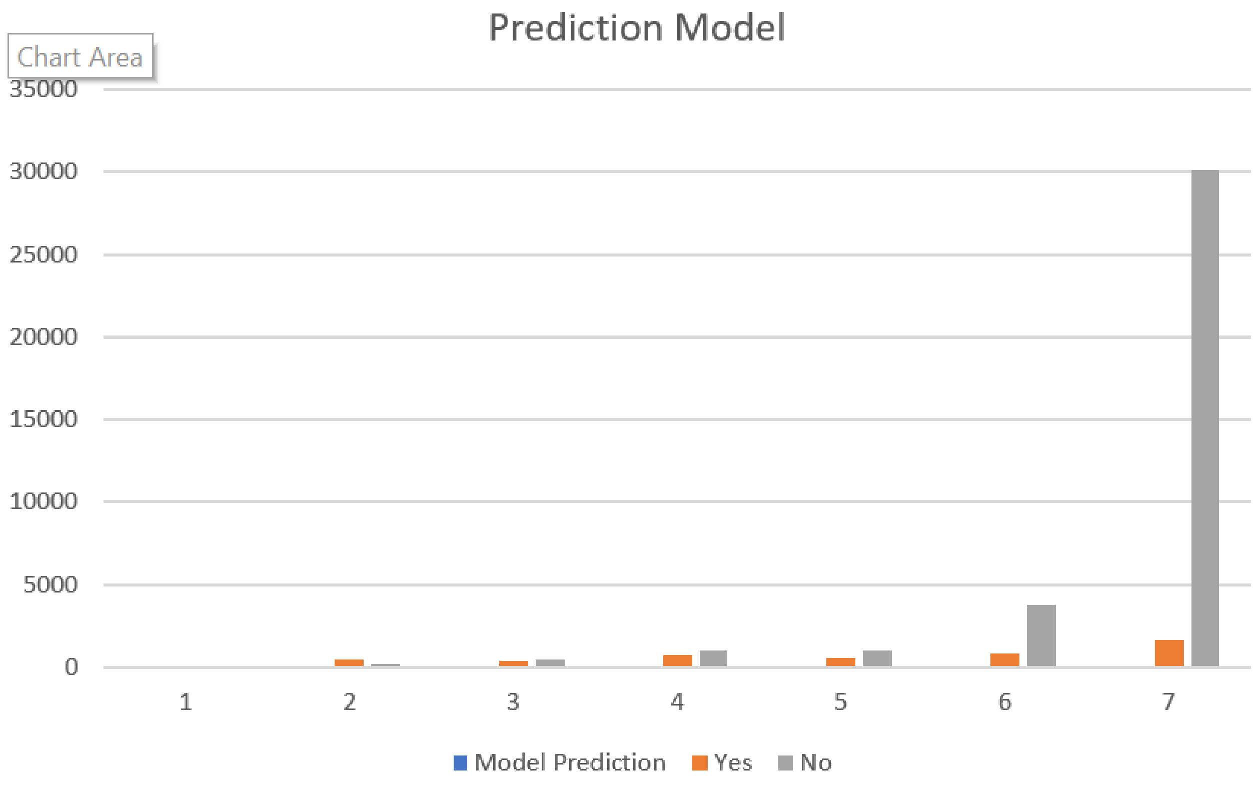

0 10 20 30 40 50
campaign
0 200 400 600 800 1000
pdays
0 1 2 3 4 5 6 7
previous
−3 −2 −1 0 1
emp.var.rate
92.5 93.0 93.5 94.0 94.5
cons.price.idx
−50 −45 −40 −35 −30
cons.conf.idx
1 2 3 4 5
euribor3m
5000 5050 5100 5150 5200
nr.employed

Correlation Analysis



Model Prediction

Model Prediction	Yes	No	Total
0.3	74	22	95
0.4	468	156	623
0.5	376	427	802
0.6	748	1042	1790
0.7	528	1010	1538
0.8	794	3762	4556
0.9	1659	30136	31795



Findings

1. This is imbalanced dataset (89% of the data are not subscribed to Term Deposit and only 11% of the data subscribed)

2. There is a feature that is unnecessary but highly affects the output target, which need to be deleted:
   -*duration*

3. There are unknown values that we can impute to improve the dataset quality:
   -*job*
   -*education*

4. There are highly correlated features that need to be taken care of:
   -*euribor3m, nr.employed, emp.var.rate*

5. Data cleaning:
   -the data values are consistent and no need further processing

6. Outliers:

   - Not need to removed

Findings

- Given the prediction model values, assume the threshold to predict Term Deposit =‘Y’ will be equal or larger than 0.6, the model has poor performance in prediction that need to be taken care of.
  - E.g. resampling

Feature Enginering

- Label Encoding:
  - Use one hot encoding and label encoding for categorical variables and target variable

Training/Test sets

- We split the dataset into 70%-30% for the training and test sets
- After encoding, the data we have:

```
X_train: (28831, 63)
X_test: (12357, 63)
y_train: (28831,)
y_test: (12357,)
```

Target variable

```
0    10969
1     1388
```

Logistic Regression

	precision	recall	f1-score	support
0	0.91	0.99	0.95	10969
1	0.67	0.19	0.30	1388
accuracy			0.90	12357
macro avg	0.79	0.59	0.62	12357
weighted avg	0.88	0.90	0.87	12357

Under-sampling

- The imbalanced datasets causes a skewness in the data distribution, create the minority class and the majority class.
- The bias in the data cause the machine learning model ignore the minority class.
- To address the problem of class imbalance, we will randomly resample the dataset using under-sampling. Under-sampling means to delete examples from the majority class.

```
X1.shape
```

```
(9280, 63)
```

```
y1.value_counts()
```

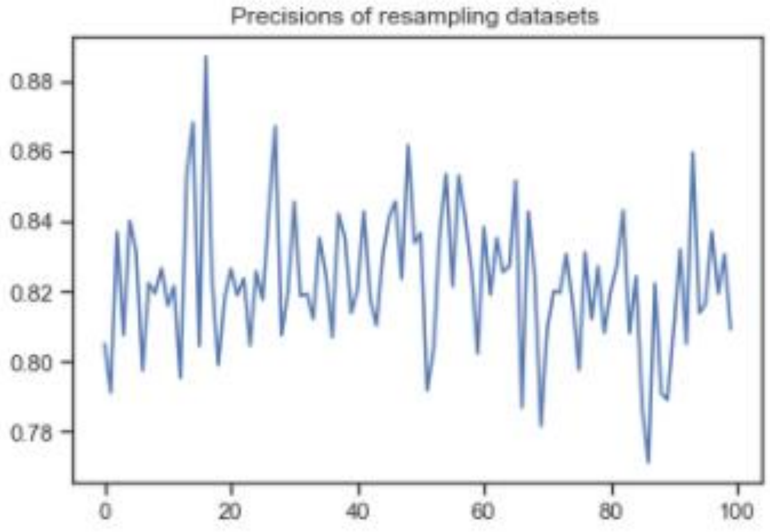
```
1    4640
0    4640
dtype: int64
```

Logistic Regression with under-sampling data

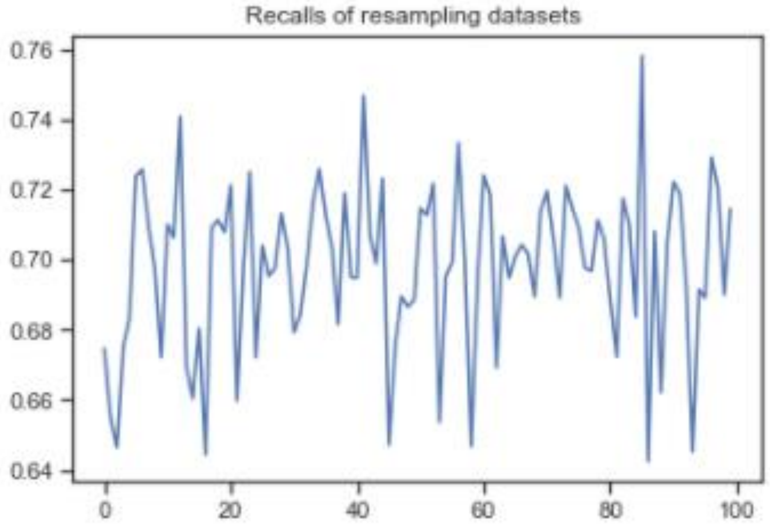
```
[Test] Accuracy score (y_predict_test, ytestlr): 0.7618534482758621

[Training] Accuracy score: (ylr, y_predict_training) 0.7800377155172413
              precision    recall  f1-score   support

           0       0.72      0.85      0.78       912
           1       0.82      0.68      0.74       944

    accuracy                           0.76      1856
   macro avg       0.77      0.76      0.76      1856
weighted avg       0.77      0.76      0.76      1856
```

Logistic Regression with under-sampling data

Average Precisions 0.8229140290755669

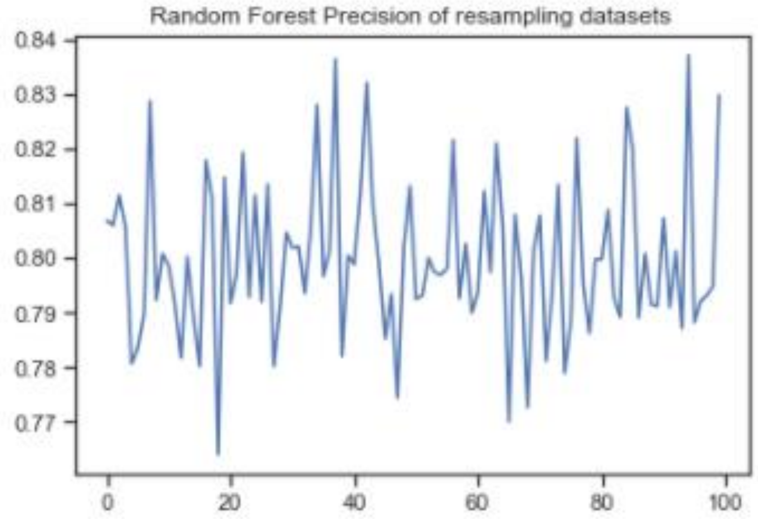


Average Recalls 0.697866322715883

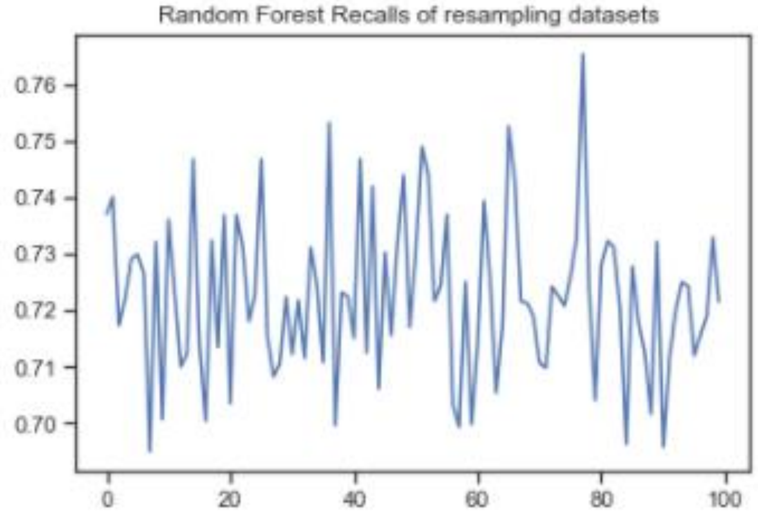


Random Forest with under-sampling data

Average Precisions 0.7997878194743473



Average Recalls 0.7225062687271955



Conclusions

- Explored and analyzed the dataset
- Handle unknown values
- Further studies can be performed to improve the models e.g. cross validation, feature importance, hyper parameter tuning, etc.

Recommendations

Target Customers

Occupation	admins, technicians, blue-collar, management, retired clients
Age	From 20s-50s, especially the group of 30s
Education	Target to customers with high education like university degree, professional education, or high school
Marital	Married
New customers	Target to new customers who have been contacted before
Contact type	Preferred cellular phone