Analysis of Bank Telemarketing Data

Kaushik Holla, Kaavya Gowthaman and Niyati Chopra

Understanding the Data

Client

Age

Job

Marital Status

Education

Credit in Default?

Housing Loan?

Personal Loan?

Last Contact

Mode of Communication

Call duration

Day, Month

Socio-Economic Factors

Quarterly Employment Variation Rate

Monthly Consumer Price Index

Monthly Consumer Confidence Index

Number of Employees

Euribor 3-month rate

Other

Outcome of previous campaign

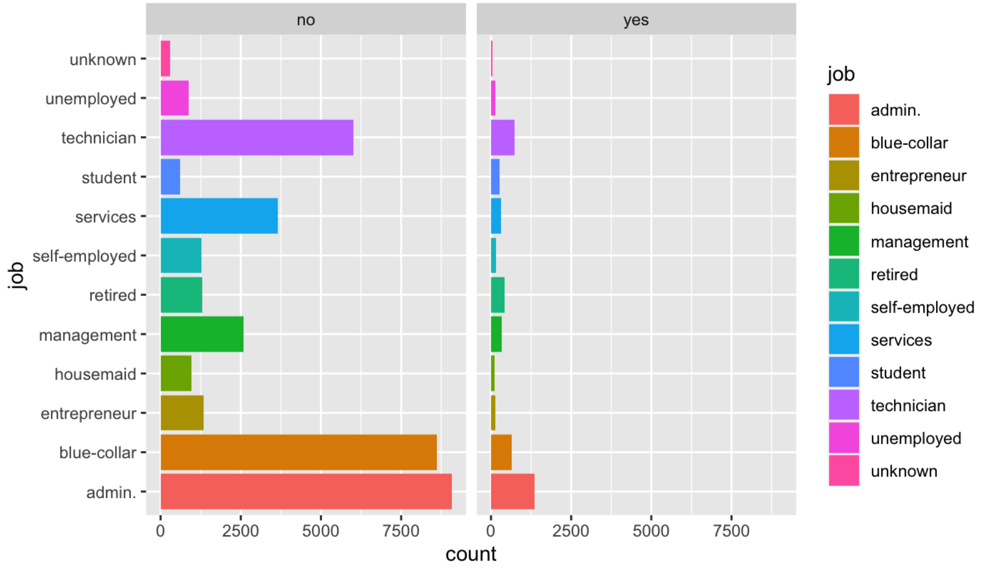
No of previous contacts in the current campaign

GOAL : The goal of the project is to build a classification model to predict if a client is going to subscribe to a term deposit

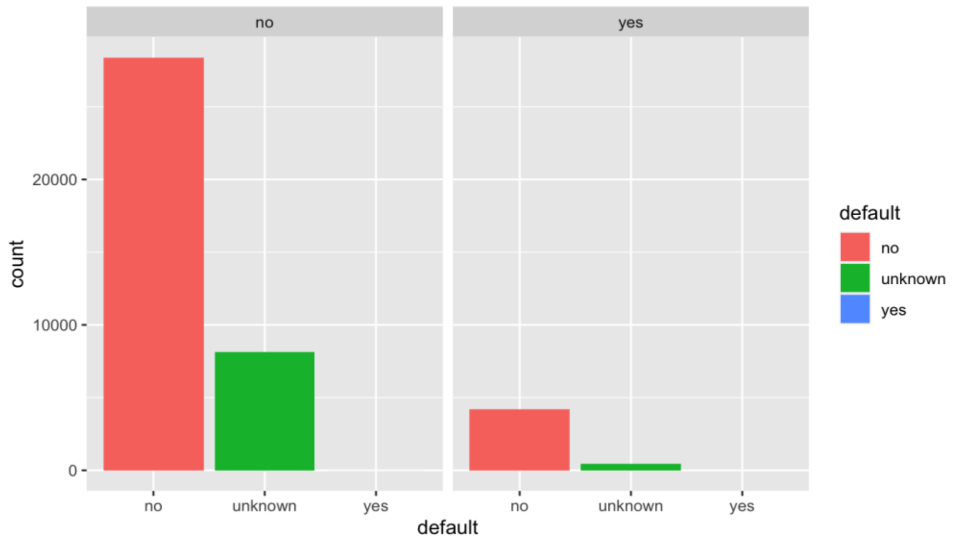
INDUSTRY RELEVANCE : The analysis is helpful in designing strategies to target prospective clients

Exploratory Data Analysis: Client Data

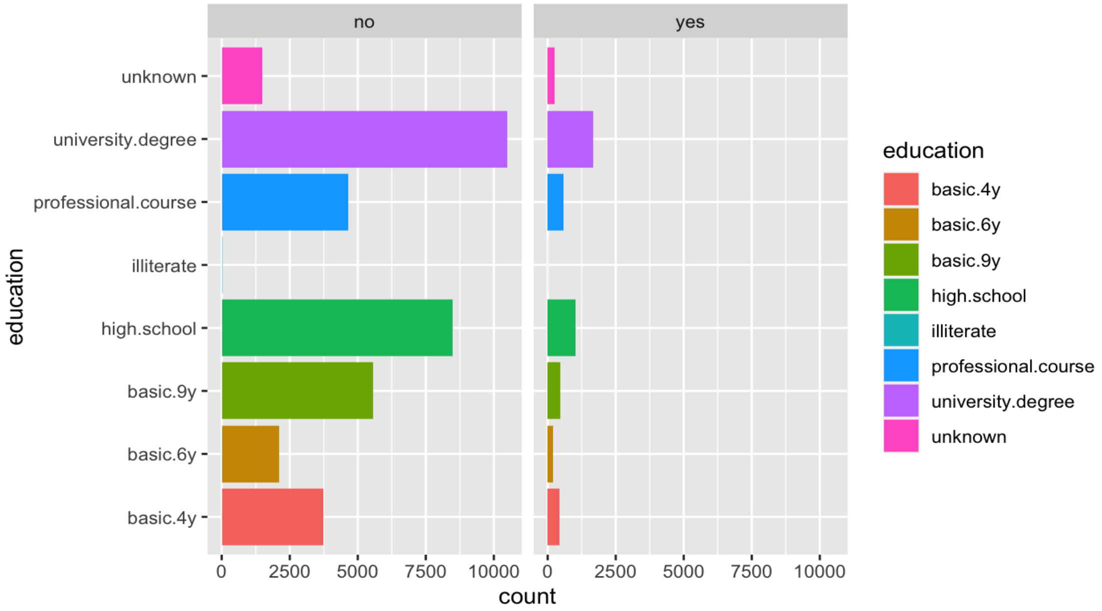
Distribution of jobs



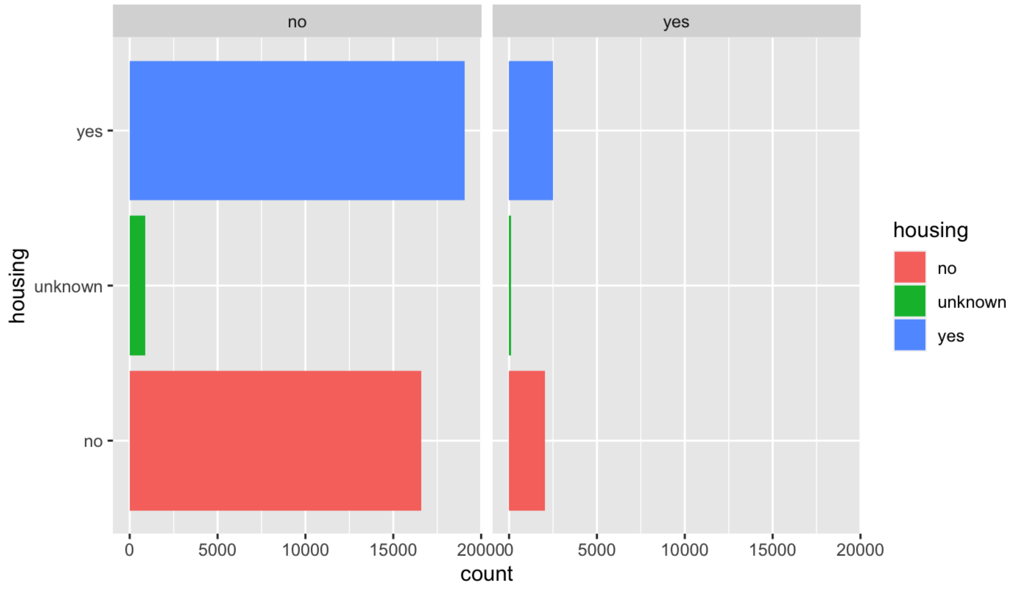
Distribution of defaulters



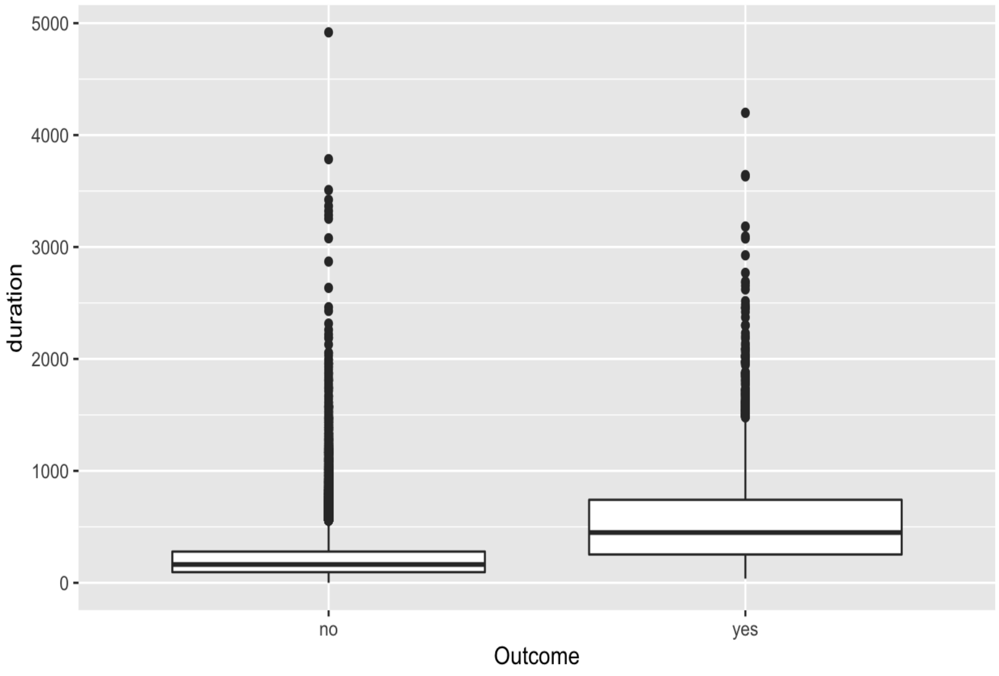
Distribution of education

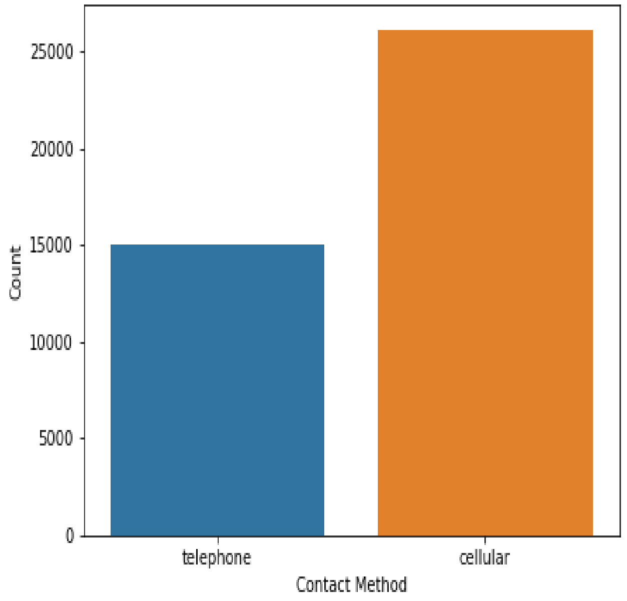


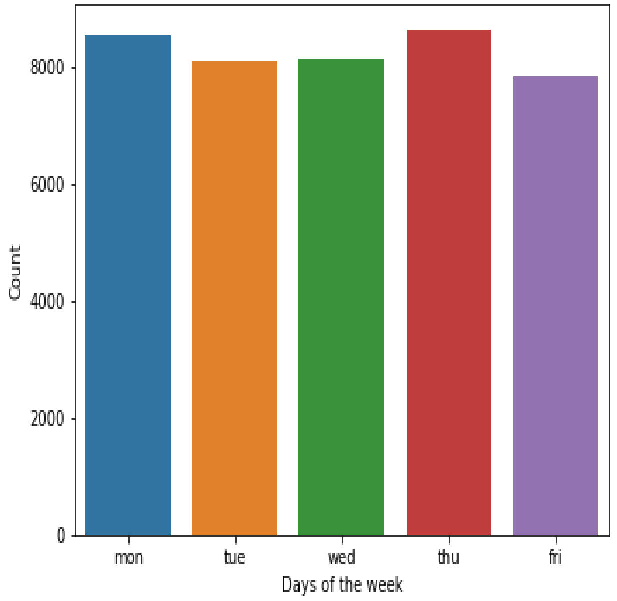
Distribution of clients with housing loans



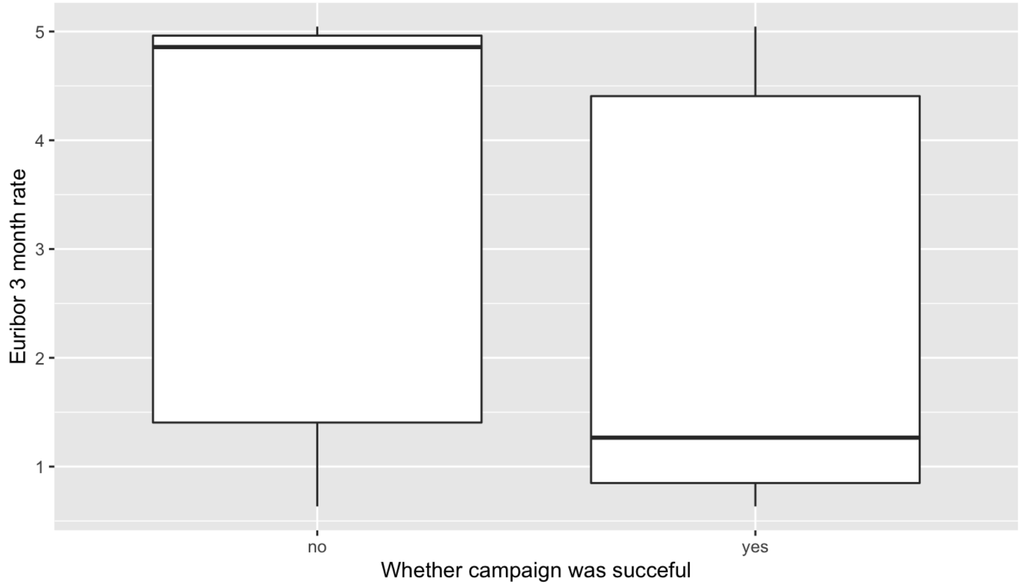
Exploratory Data Analysis: Last Contact Data

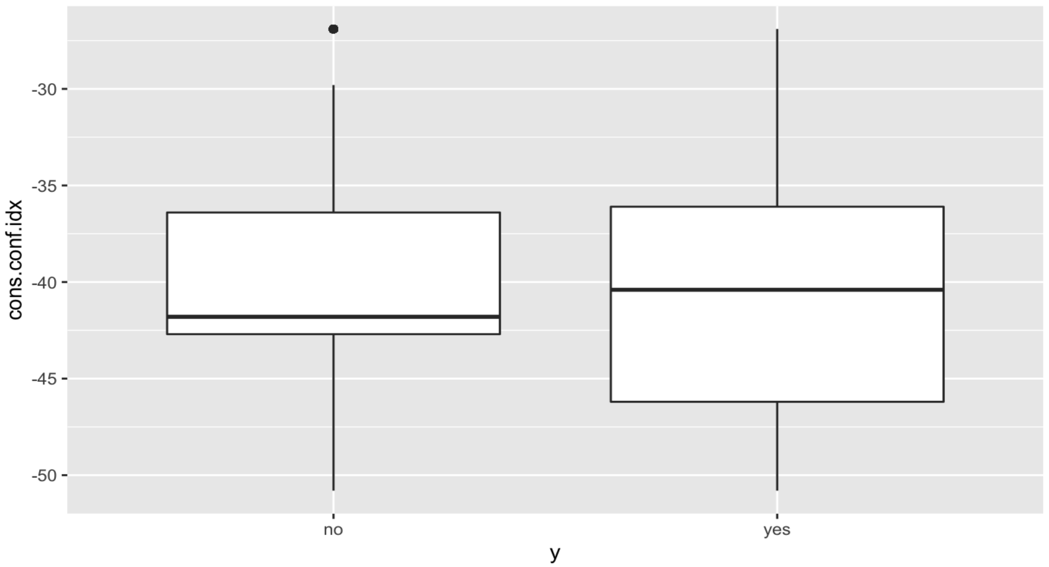

Call duration is higher for positive outcomes

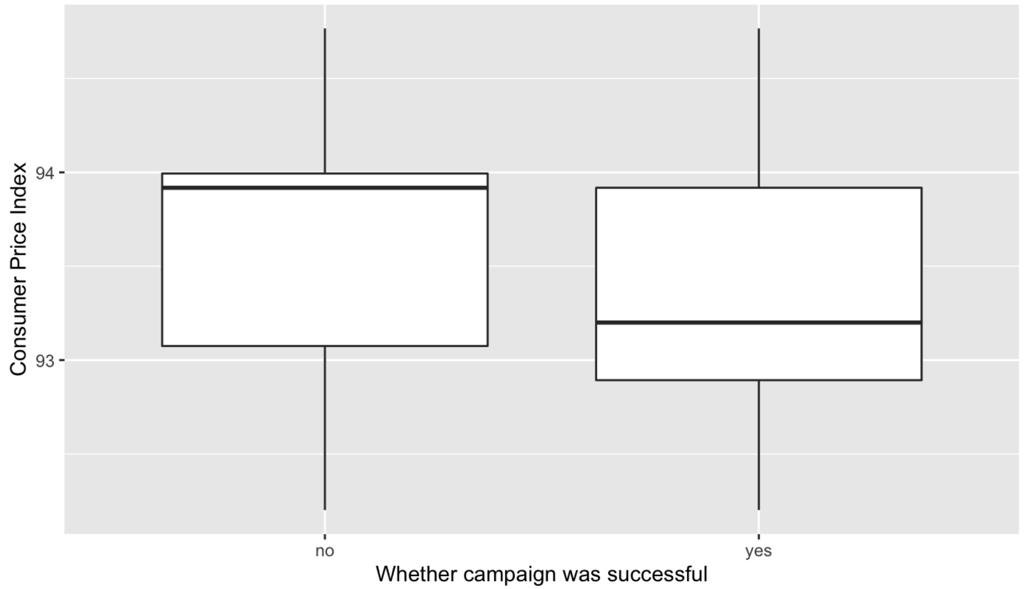

Count of Contact Methods

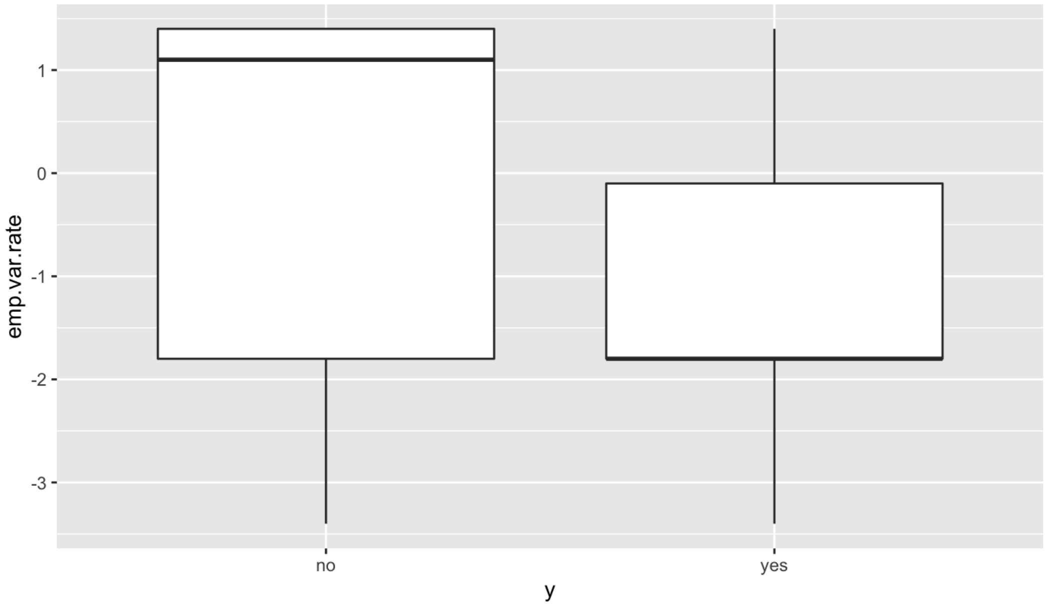

Count of Calls made on Days of the week

Exploratory Data Analysis: Socio-Economic Factors


Outcome is highly dependent on Euribor 3 month rate


Outcome might be dependent on Consumer confidence index


Outcome is dependent on Consumer price index

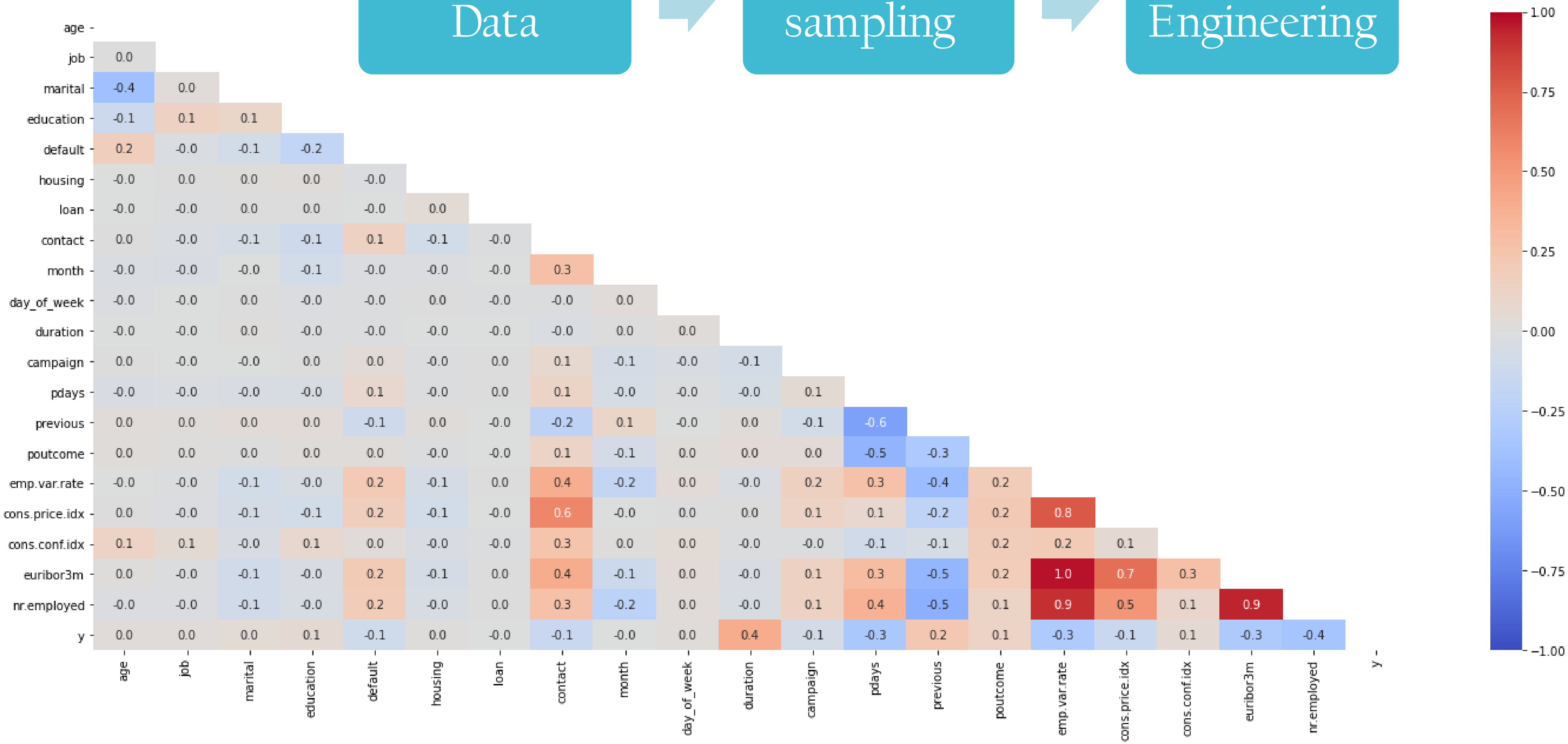

Outcome is highly dependent on Employment Variation Rate

Conclusions drawn from EDA

- Socio-economic factors highly influence the outcome. There might be possible correlation between these variables that needs to be checked.
- Client Data like education, job, marital status etc. does not influence the outcome.
- Call duration influences the outcome. The duration is higher for people who subscribed to the term deposit.

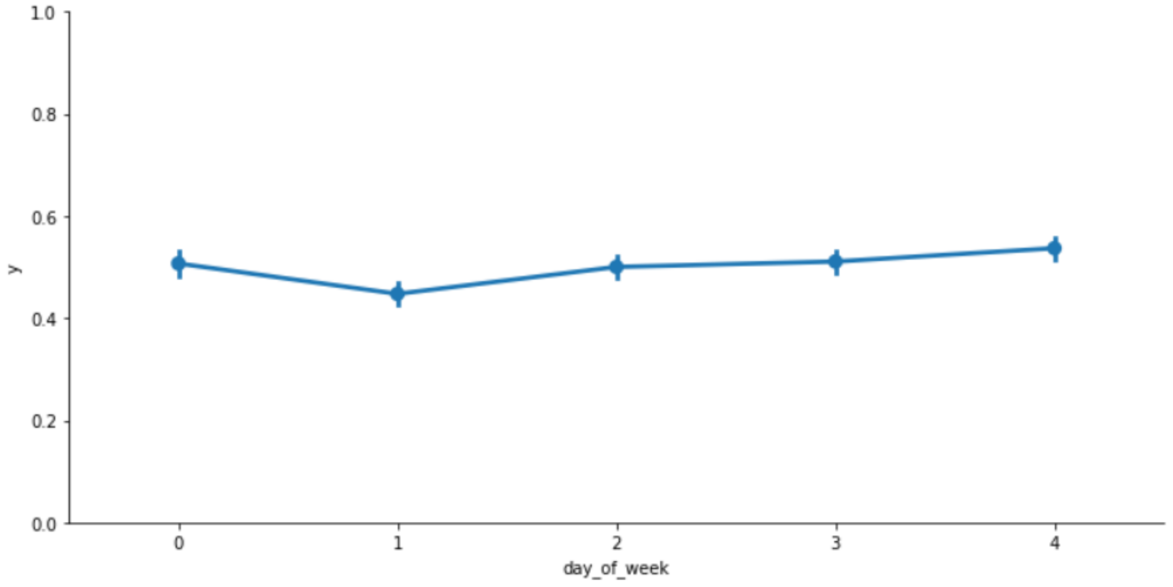
Feature Engineering

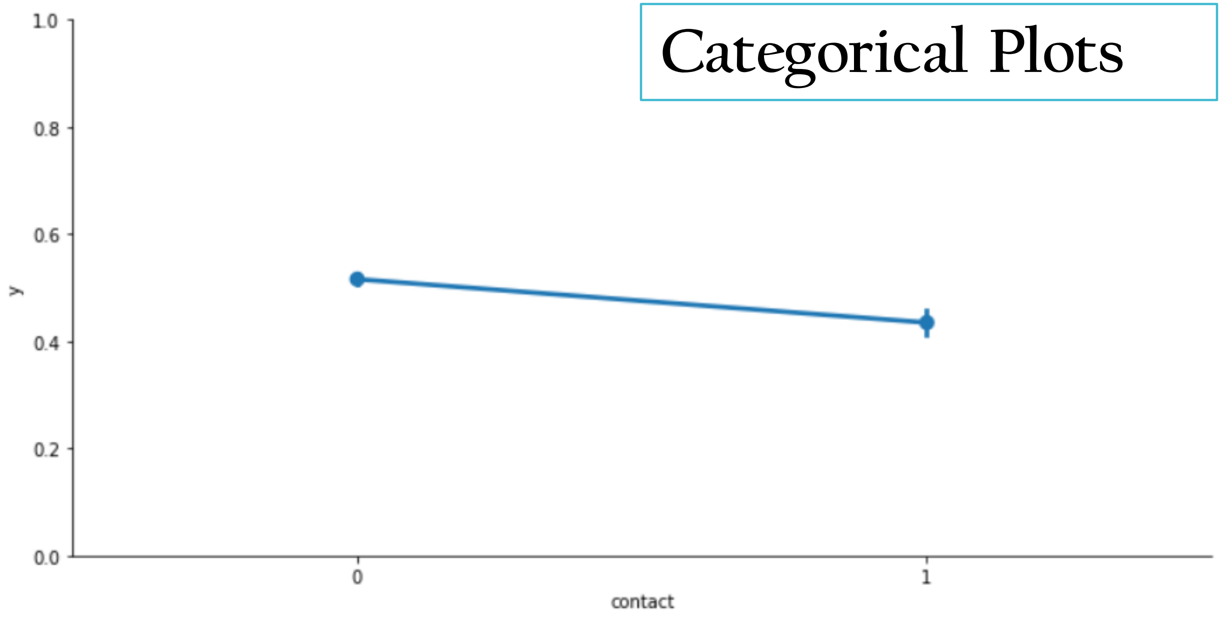
Imbalanced Data → Under-sampling → Feature Engineering

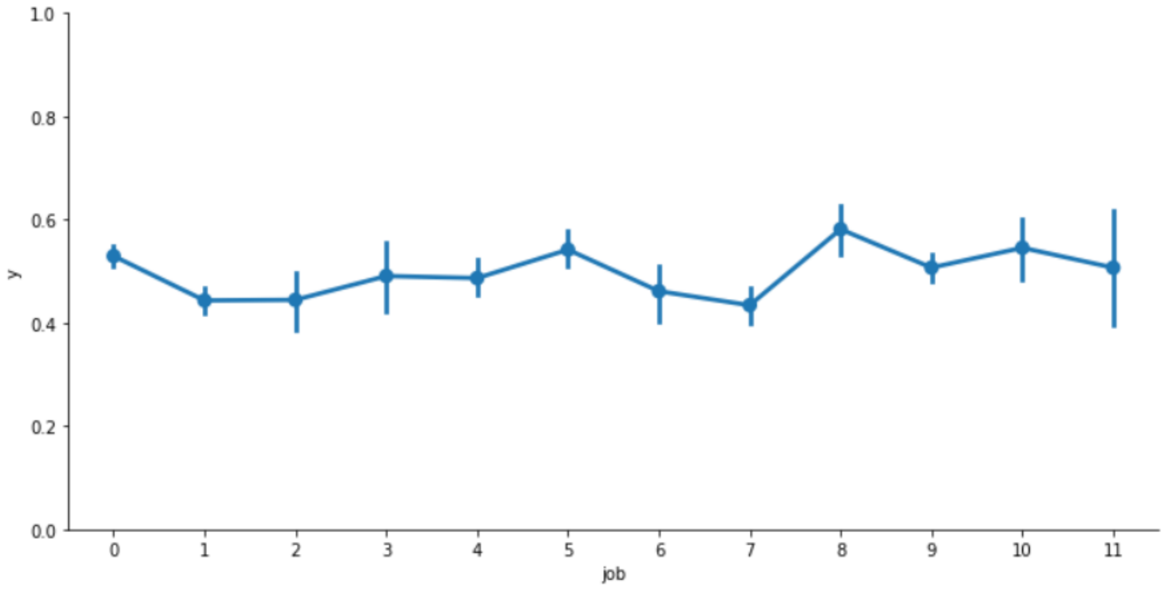


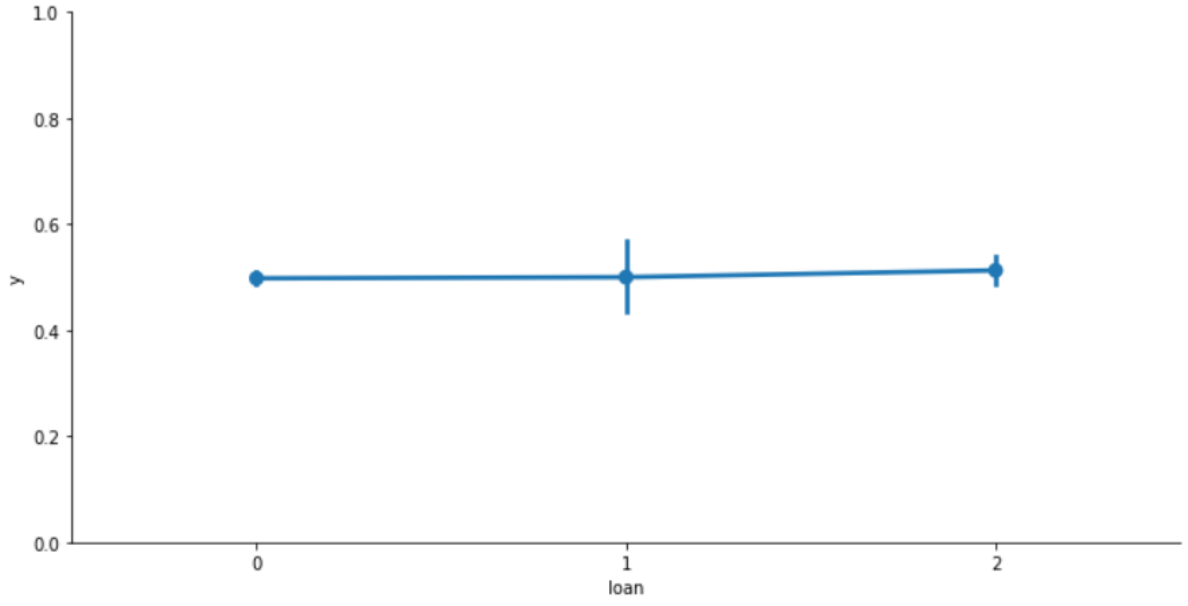
	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
age																					
job	0.0																				
marital	-0.4	0.0																			
education	-0.1	0.1	0.1																		
default	0.2	-0.0	-0.1	-0.2																	
housing	-0.0	0.0	0.0	0.0	-0.0																
loan	-0.0	-0.0	0.0	0.0	-0.0	0.0															
contact	0.0	-0.0	-0.1	-0.1	0.1	-0.1	-0.0														
month	-0.0	-0.0	-0.0	-0.1	-0.0	-0.0	-0.0	0.3													
day_of_week	-0.0	-0.0	0.0	-0.0	-0.0	0.0	-0.0	-0.0	0.0												
duration	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0											
campaign	0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	0.1	-0.1	-0.0	-0.1										
pdays	-0.0	-0.0	-0.0	-0.0	0.1	-0.0	0.0	0.1	-0.0	-0.0	-0.0	0.1									
previous	0.0	0.0	0.0	0.0	-0.1	0.0	-0.0	-0.2	0.1	-0.0	0.0	-0.1	-0.6								
poutcome	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.1	-0.1	0.0	0.0	0.0	-0.5	-0.3							
emp.var.rate	-0.0	-0.0	-0.1	-0.0	0.2	-0.1	0.0	0.4	-0.2	0.0	-0.0	0.2	0.3	-0.4	0.2						
cons.price.idx	0.0	-0.0	-0.1	-0.1	0.2	-0.1	-0.0	0.6	-0.0	0.0	0.0	0.1	0.1	-0.2	0.2	0.8					
cons.conf.idx	0.1	0.1	-0.0	0.1	0.0	-0.0	-0.0	0.3	0.0	0.0	-0.0	-0.0	-0.1	-0.1	0.2	0.2	0.1				
euribor3m	0.0	-0.0	-0.1	-0.0	0.2	-0.1	0.0	0.4	-0.1	0.0	-0.0	0.1	0.3	-0.5	0.2	1.0	0.7	0.3			
nr.employed	-0.0	-0.0	-0.1	-0.0	0.2	-0.0	0.0	0.3	-0.2	0.0	-0.0	0.1	0.4	-0.5	0.1	0.9	0.5	0.1	0.9		
y	0.0	0.0	0.0	0.1	-0.1	0.0	-0.0	-0.1	-0.0	0.0	0.4	-0.1	-0.3	0.2	0.1	-0.3	-0.1	0.1	-0.3	-0.4	

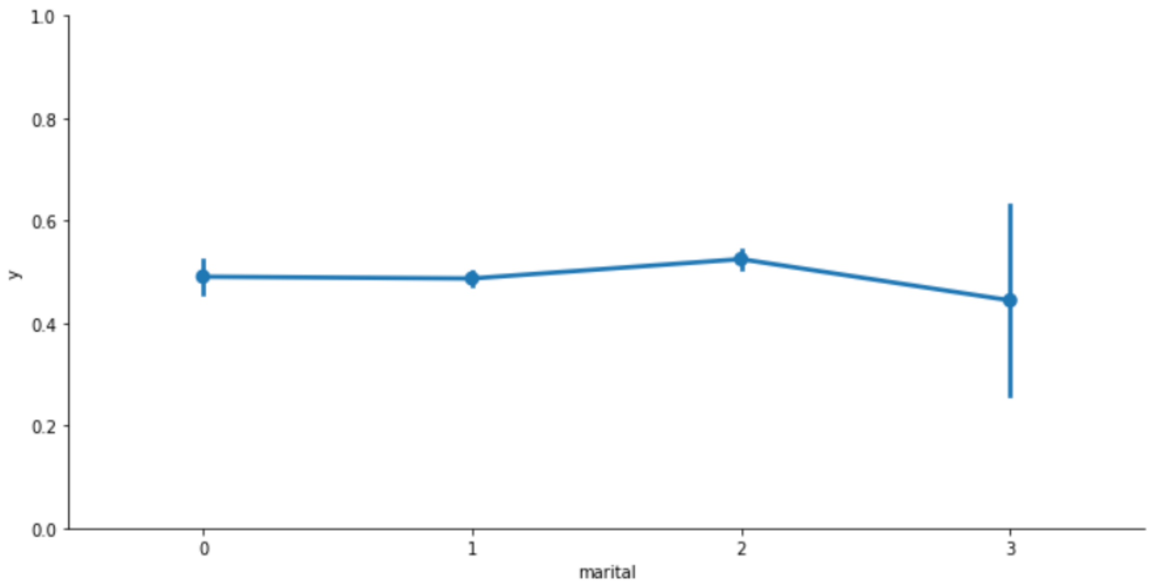
Categorical Plots

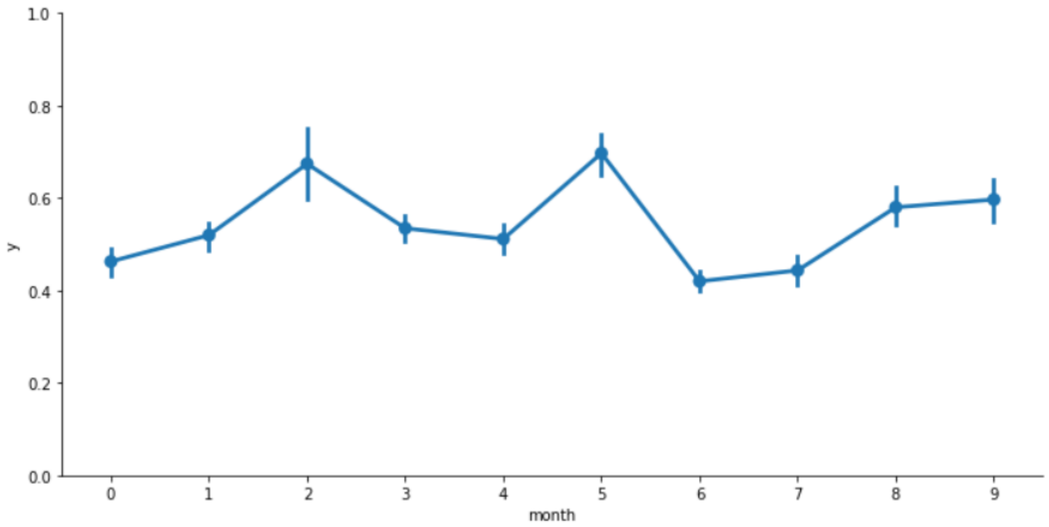


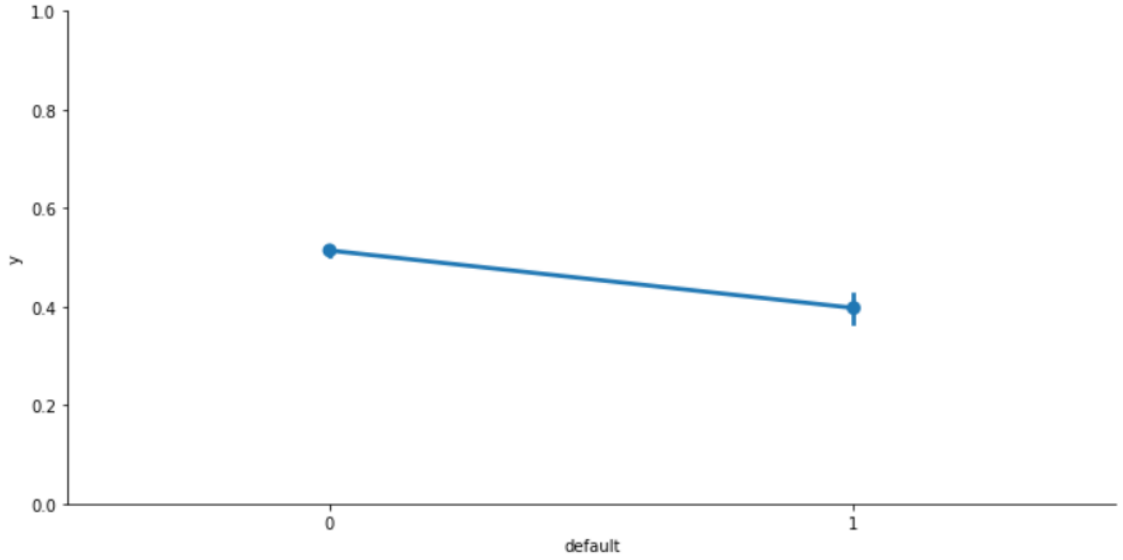


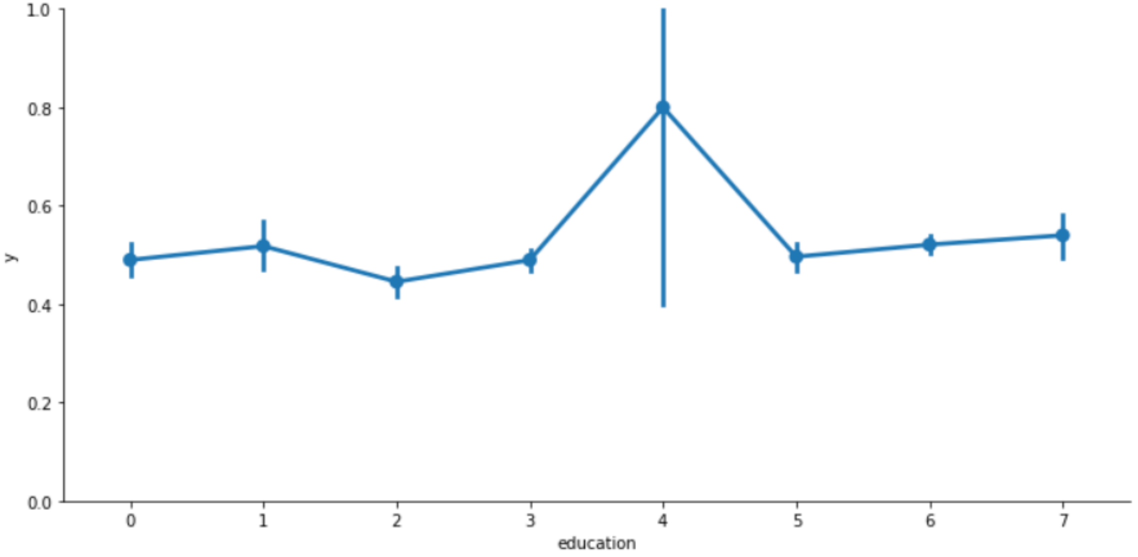





1.0
0.8
0.6
0.4
0.2
0.0
y
0
1
2
3
marital


1.0
0.8
0.6
0.4
0.2
0.0
y
0
1
2
3
4
5
6
7
8
9
month


1.0
0.8
0.6
0.4
0.2
0.0
y
0
1
default


1.0
0.8
0.6
0.4
0.2
0.0
y
0
1
2
3
4
5
6
7
education

Statistical Significance Test:

We performed Statistical significance test for every feature with the output to see if a feature is statistically significant with output.

Feature	P-Value
Age	0.551
Job	0.495
Marital	0.00604
Education	0.000896
Default	2.47e-13
Housing	0.719
Loan	0.319
Contact	6.08e-10
Month	0.839
Day of Week	0.000613
Duration	4.38e-19
Campaign	0.466
pdays	2.25e-66
previous	2.04e-19
poutcome	1.37e-56
Emp.var.rate	1.09e-10
Cons.price.idx	8.1e-05
Cons.conf.idx	1.83e-05
Euribor3m	1.24e-19
Nr.employed	3.27e-38

Conclusion After Feature Engineering

- **Duration, pdays, emp.var.rate, euribor3m and nr.employed** have good correlation with output therefore they might form a very good features compared to others.
- There is high correlation between the socio-economic variables.
- From categorical plot we can see that when value of **poutcome(Previous campaign outcome)** is success, there is 72.5% positive outcome.
- **Age, housing, month and campaign** have p-value greater than 0.05 therefore we can eliminate them as they fail to reject null hypothesis.

Models and Evaluation Metrics

Models Implemented:

Logistic Regression
Support Vector Classifier
K Nearest Neighbours
Random Forest Classifier

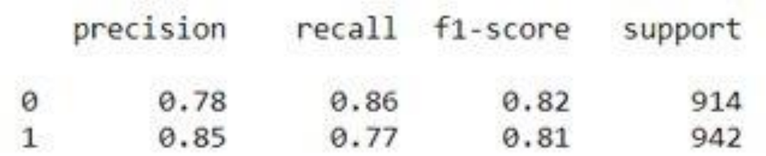
		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Since the problem we are trying to solve is a classification problem, we generated a confusion matrix to calculate F1 score to compare models

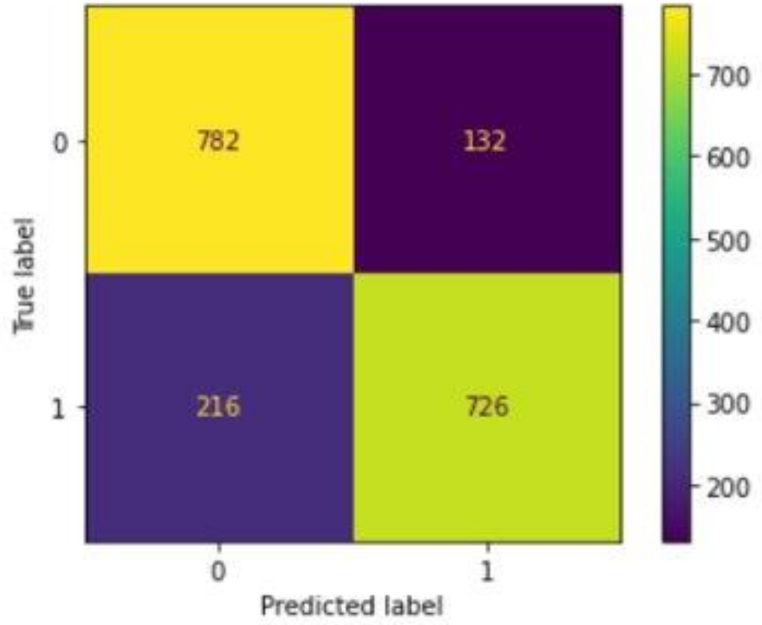
$$\text{F1 Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision + Recall}}$$

Logistic Regression

The accuracy for this model is: 81.25%

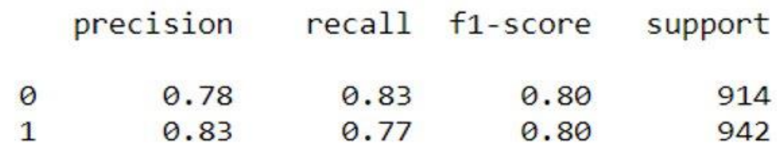


```
     precision    recall  f1-score   support

0         0.78      0.86      0.82       914
1         0.85      0.77      0.81       942
```

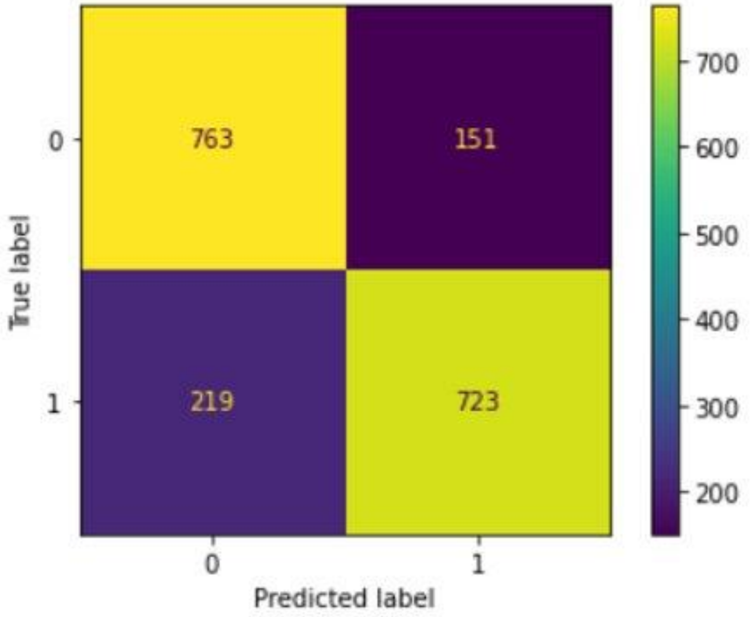


Support Vector Classifier

The accuracy for this model is: 80.06%

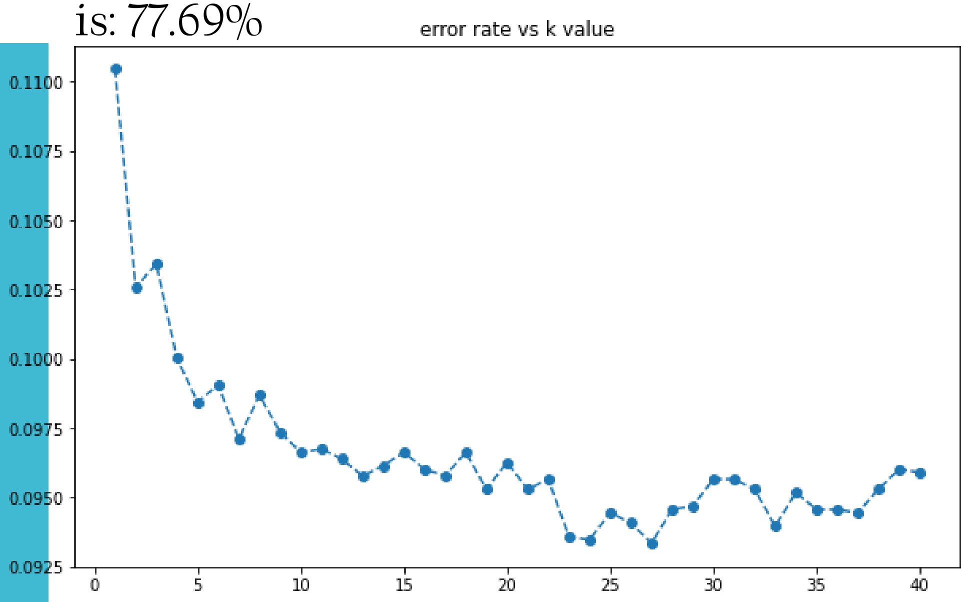


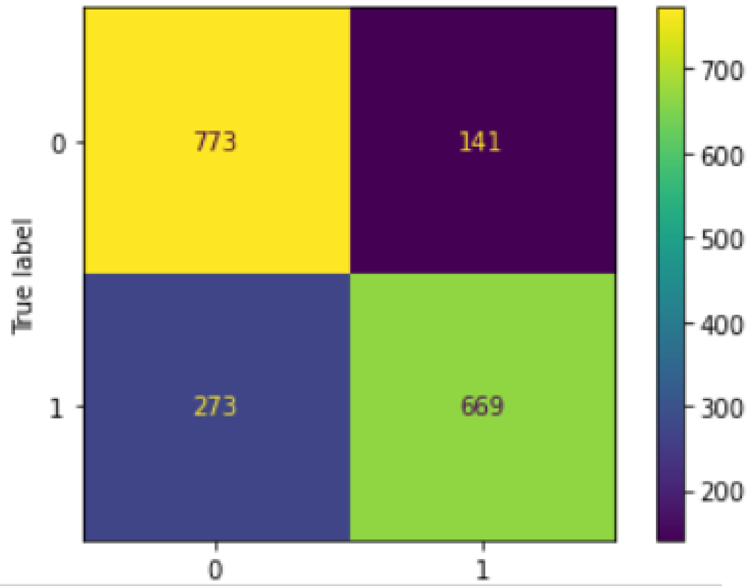
```
     precision    recall  f1-score   support

0         0.78      0.83      0.80       914
1         0.83      0.77      0.80       942
```

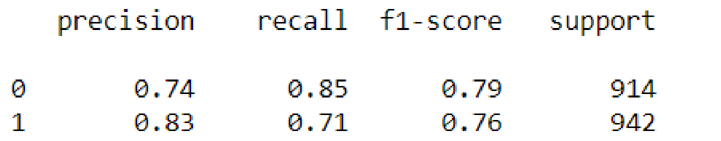


K- Nearest Neighbors (k=27)

The accuracy of this model is: 77.69%



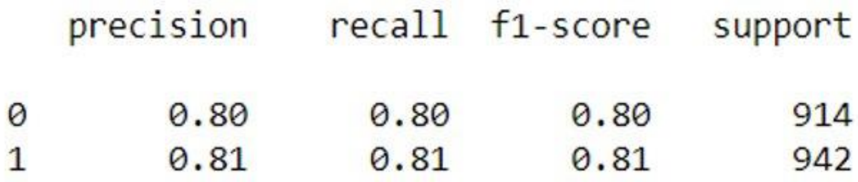




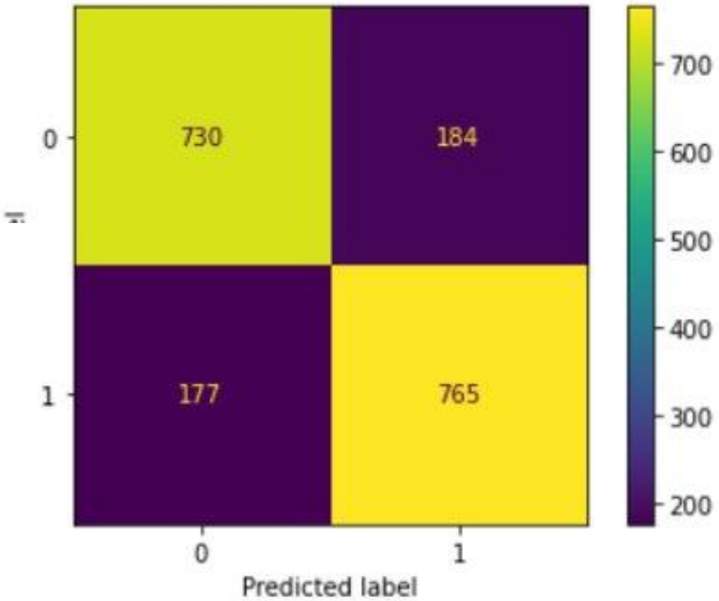
```
          precision    recall  f1-score   support

0              0.74      0.85      0.79       914
1              0.83      0.71      0.76       942
```

Random Forest Classifier

The accuracy for this model is: 80.54%



```
      precision    recall  f1-score    support

0          0.80      0.80      0.80        914
1          0.81      0.81      0.81        942
```



Conclusion after Machine Learning

- Using our evaluation metrics, we can conclude that the logistic regression model was the best fit model on our data set giving an overall F1 score of 0.82

- Questions?