Predicting Deposits Subscribe by Telephonic Marketing

Kurnia Anwar Ra’if

Sept, 16th 2021

My Profile

Physicist l Earth Scientist l Data Enthusiast l Software Engineer l Data Scientist

An undergraduate Physics, ITB - July 2020

Kurnia Anwar Ra’if

https://github.com/anwarraif

https://www.linkedin.com/in/kurnia-anwar-ra-if/

kurniaanwarraif@gmail.com

Today's Presentation

STEP 1 : Data Understanding

Deep and dive the bank data

STEP 2 : Exploratory Data Analysis

Analyze Categorical and Numerical Data

STEP 3 : Data Cleaning and Data Preprocessing

STEP 4 : Predicting Deposits Subscribe using supervised ML Model :
Logistic Regression, Naïve Bayes, KNN,and SVM

Data Understanding

Data Understanding : Source of Data

The data source is from Kaggle : Banking Dataset – Marketing Target

Maybank, the largest financial institution in Malaysia that started in 1970. It provides different types of financial and banking services to the public which include deposit and investment, loan financing, wealth management, banking and so on. Due to an intense competition among the financial institutions or banks, Maybank has provided a broad range of products and services to attract the customers. During the marketing of those services and products, Maybank has spent a huge amount of money.

The purpose is to predict the customers will subscribed deposit or not.

Data Understanding

There are 16 Columns and Subscribed Column as a target prediction

Column Name	Description
age	Age of customers
Job	Type of Job
Marital	Marital Status
Education	Type of education customers
Default	Has credit or not ?
Balance	Average yearly balance, in euros
Housing	Has housing loan ?
Loan	Has personal loan ?
Contact	Type of contact communication
Day	Last contact day of the month
Month	Last contact mount of year
Duration	Last contact duration, in seconds
Campaign	number of contacts performed during this campaign and for this client
pdays	number of days that passed by after the client was last contacted from a previous campaign
previous	number of contacts performed before this campaign and for this client
poutcome	outcome of the previous marketing campaign
Subscribed	has the client subscribed a term deposit?

Exploratory Data Analyst (EDA)

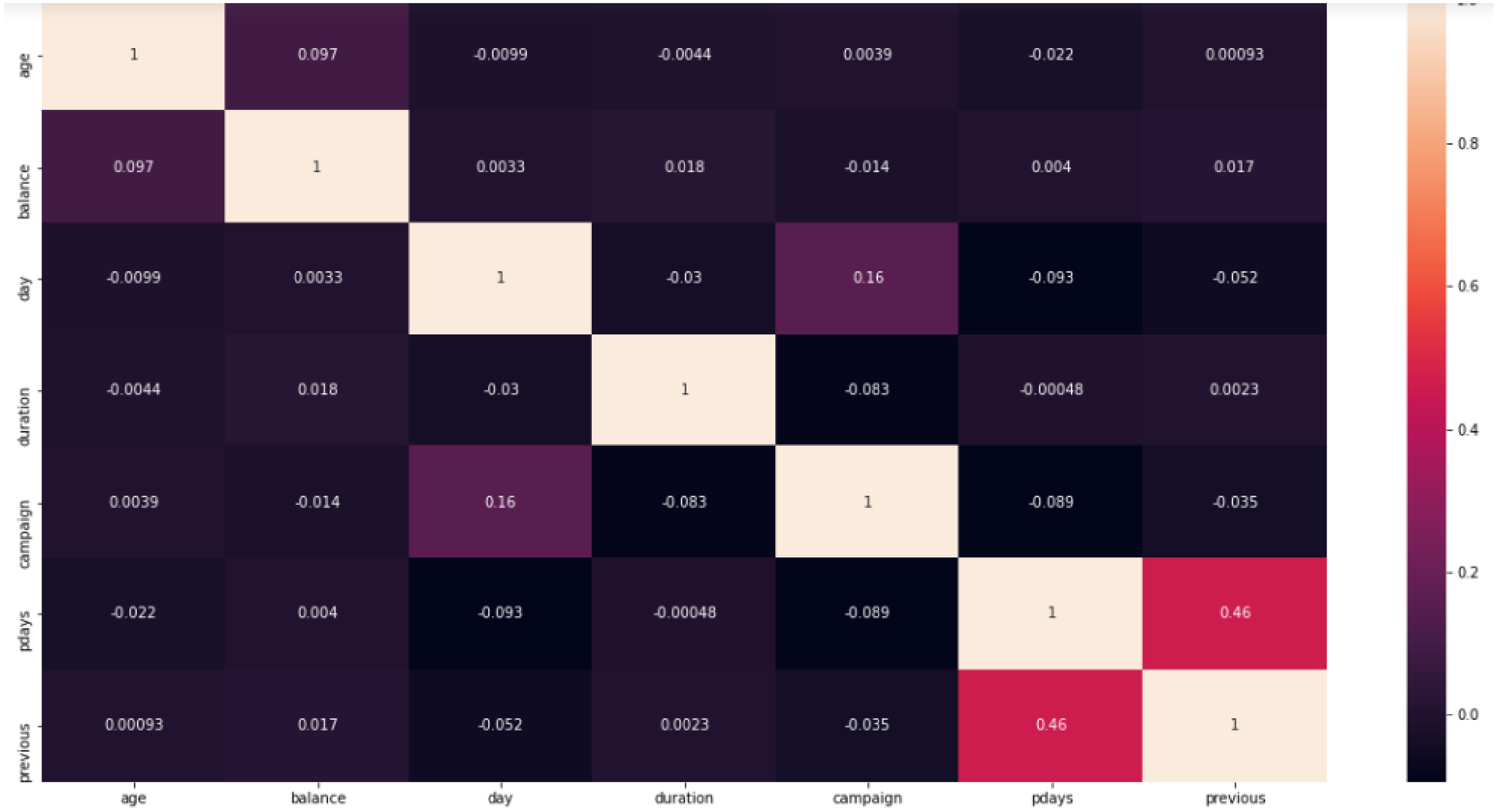
Data Overview

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	subscribed
58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
33	services	married	secondary	no	-333	yes	no	cellular	30	jul	329	5	-1	0	unknown	no
57	self-employed	married	tertiary	yes	-3313	yes	yes	unknown	9	may	153	1	-1	0	unknown	no
57	technician	married	secondary	no	295	no	no	cellular	19	aug	151	11	-1	0	unknown	no
28	blue-collar	married	secondary	no	1137	no	no	cellular	6	feb	129	4	211	3	other	no
44	entrepreneur	single	tertiary	no	1136	yes	yes	cellular	3	apr	345	2	249	7	other	no

There are :

A. Categorical Data : 9 Columns
B. Numerical Data : 7 Columns

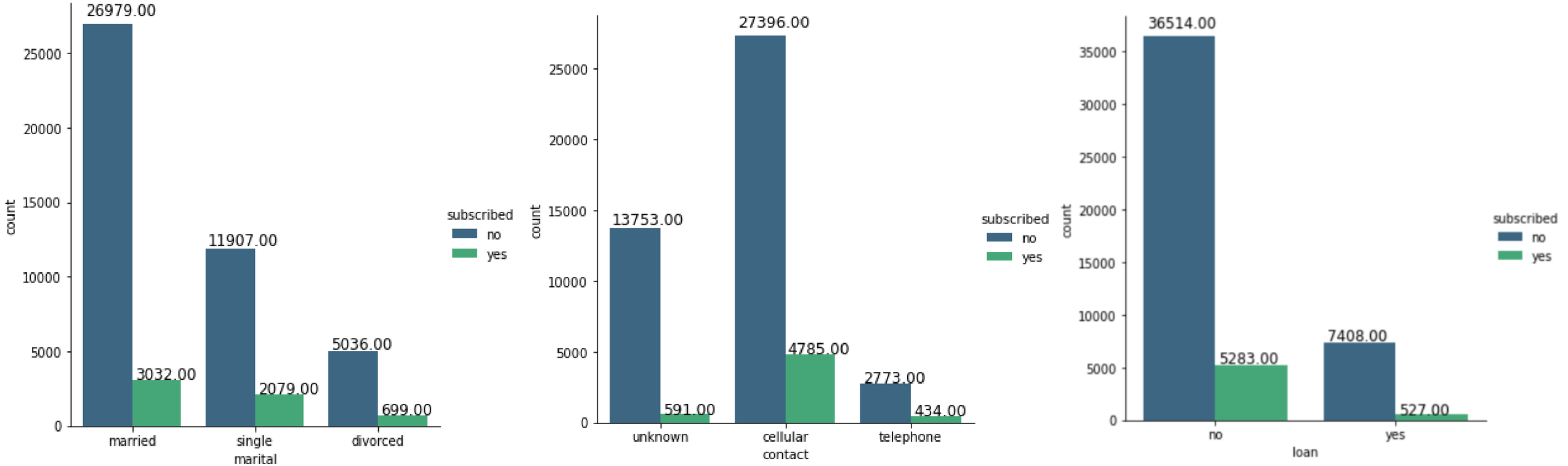
Correlation for each columns



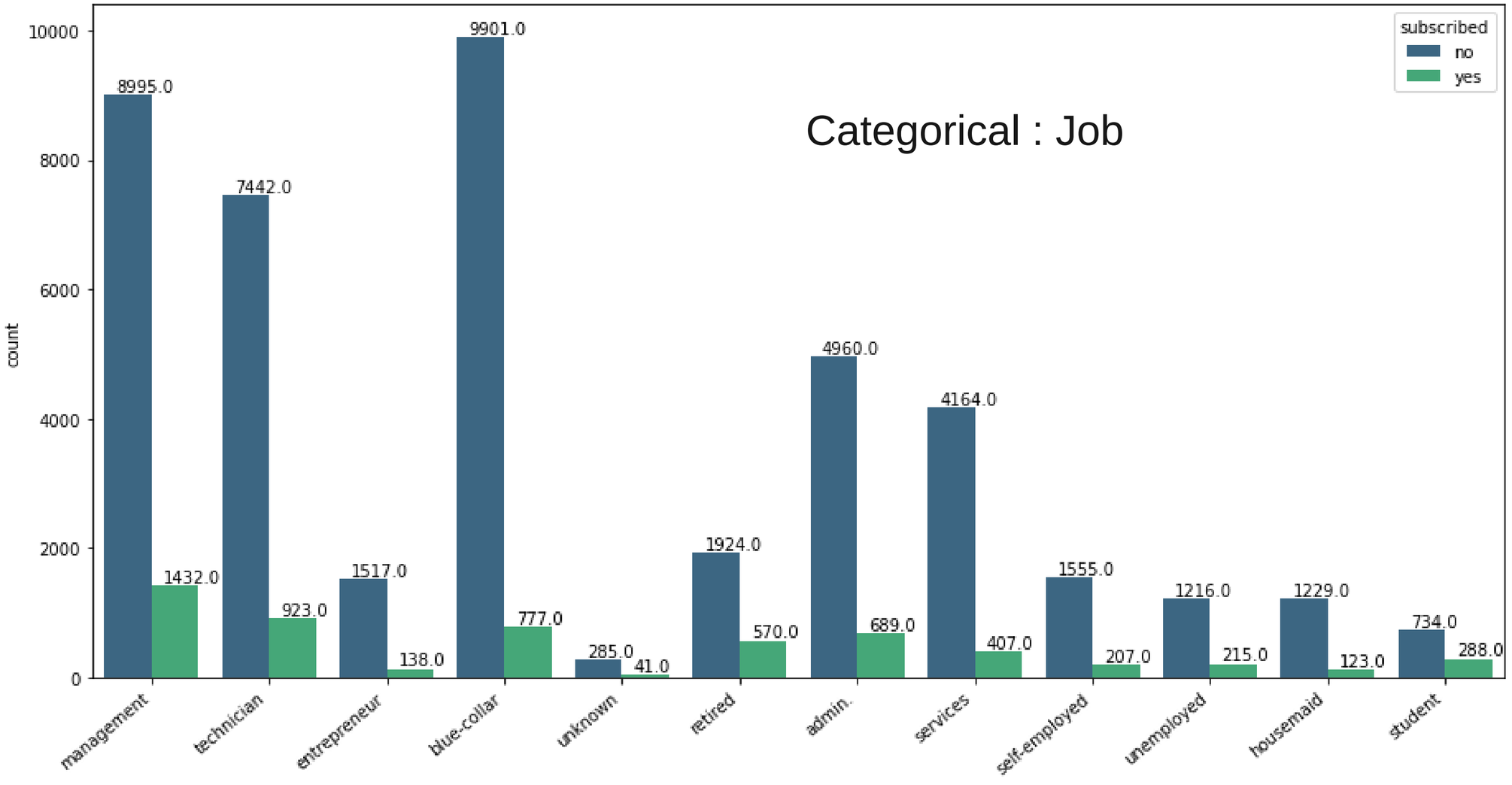
From the graph above, it seems like nothing highly correlated as most of the values is below 0.5. There is no correlation between each variables, No features are removed.

EDA | Categorical Data

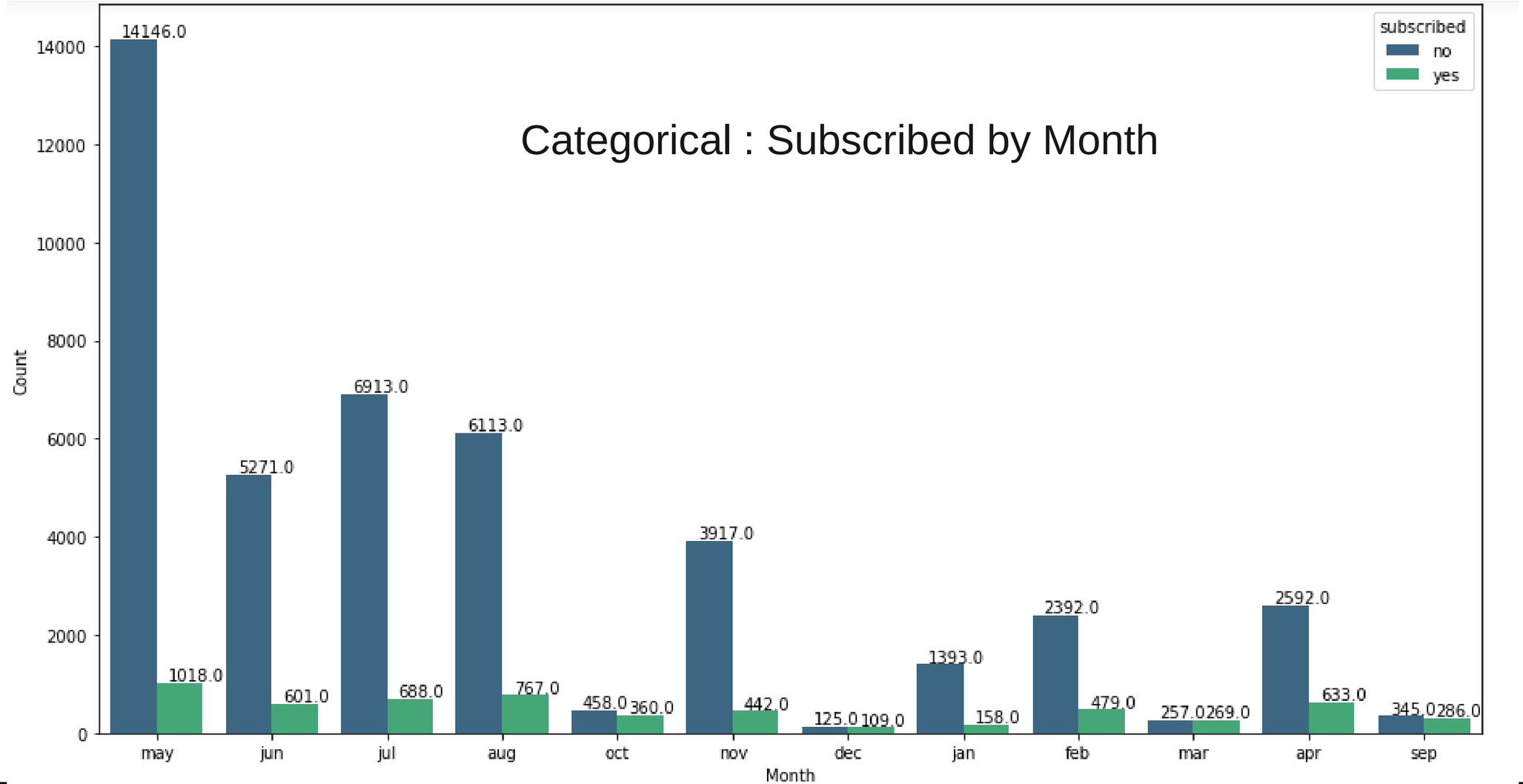
Categorical : Marital, Contact, Loan



EDA | Categorical Data

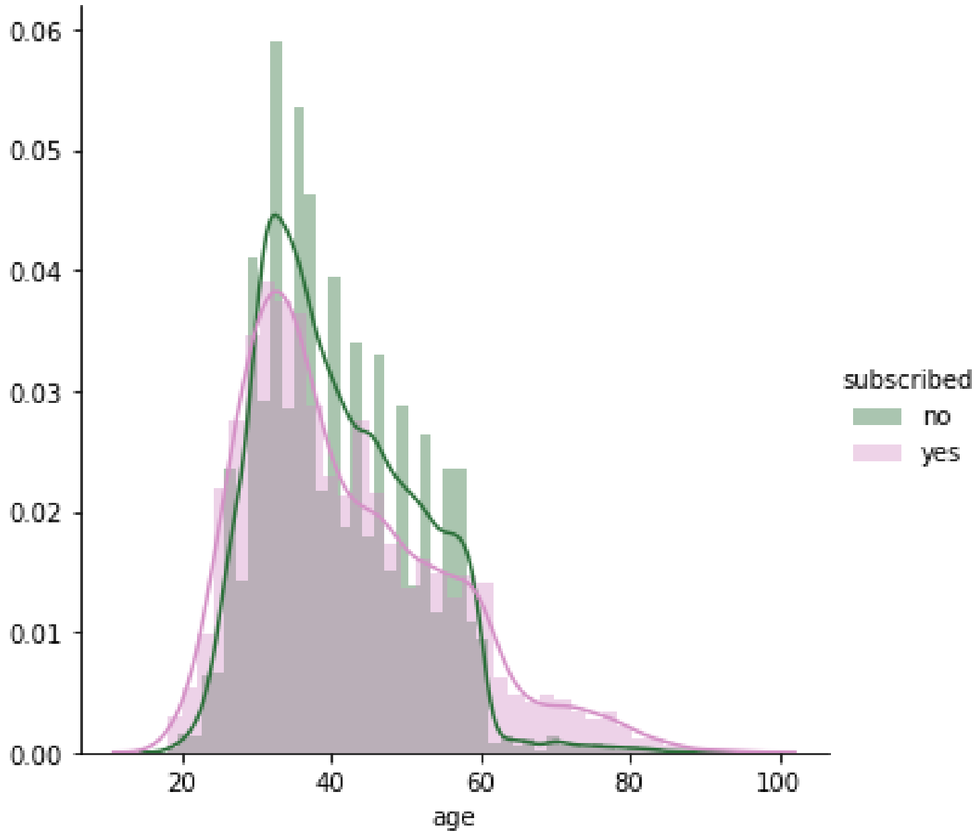


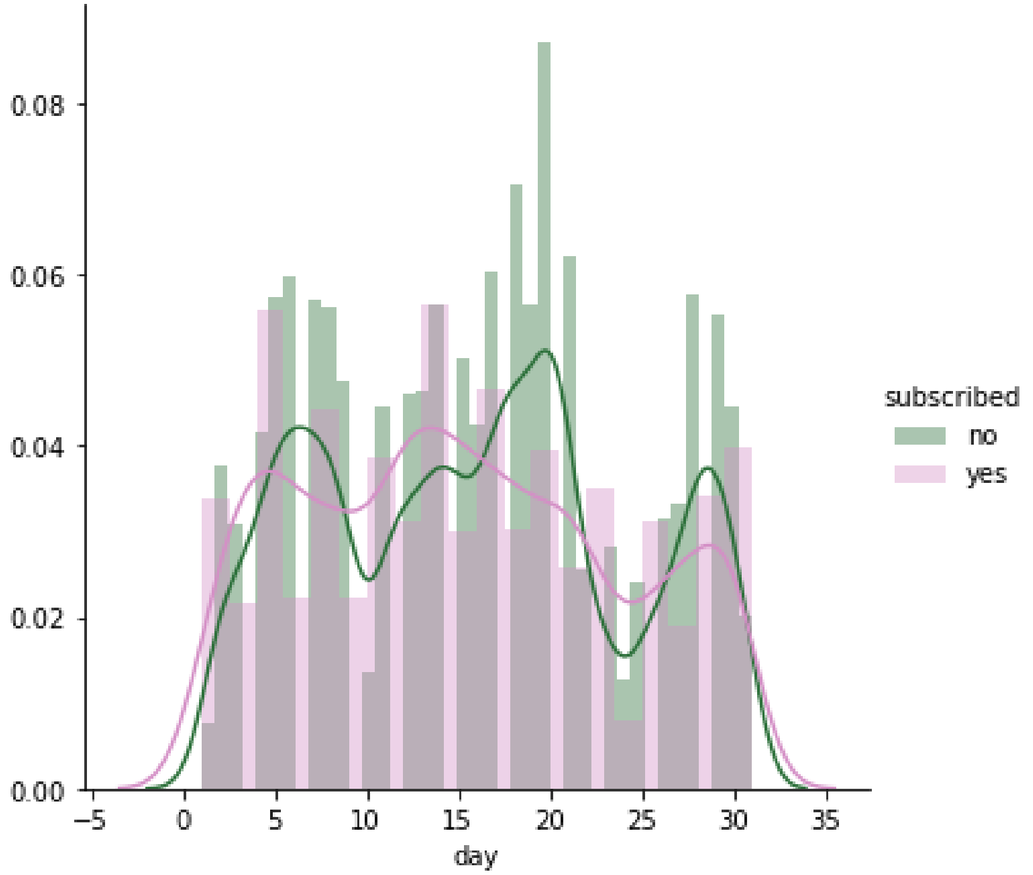
EDA | Numerical Data



EDA | Numerical Data

Numerical : Age and Day





Data Cleaning and Data Preprocessing

Data Cleaning

Drop Duplicated

```
Before drop duplicated data

(39785, 17)

After drop duplicated data

(36895, 17)
```

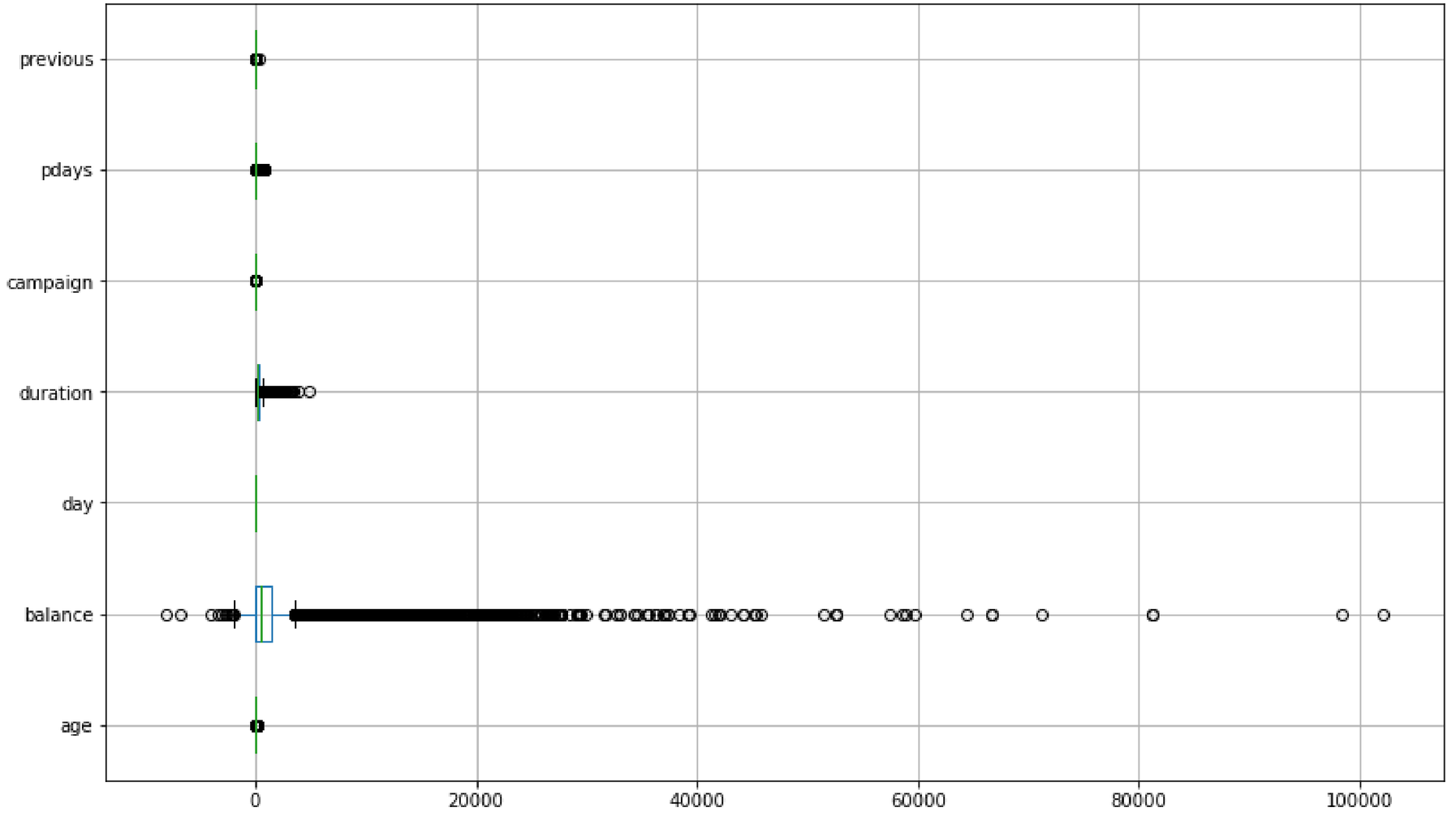
```
age           0
job           0
marital       0
education     0
default       0
balance       0
housing       0
loan          0
contact       0
day           0
month         0
duration      0
campaign      0
pdays         0
previous      0
poutcome      0
subscribed    0
```

Missing Value Check

Column Name	Missing Value	Imputation
Job	Unknown	Mode
Education	Unknown	Mode
Contact	Unknown	Mode
Poutcome	Unknown	Mode

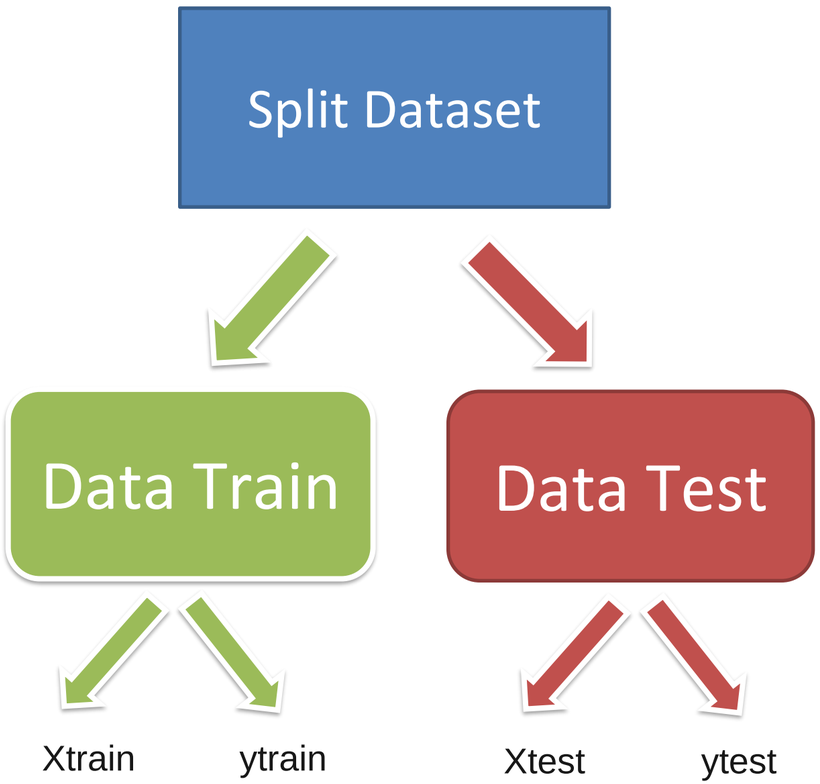
Column Name	% Missing Value
Job	0,613
Education	3,825
Contact	35,868
Poutcome	100

Outliers Analysis



There are 14073 Outliers in this dataset, these outliers will be removed

Data Pre-Processing



Encoding Preparation

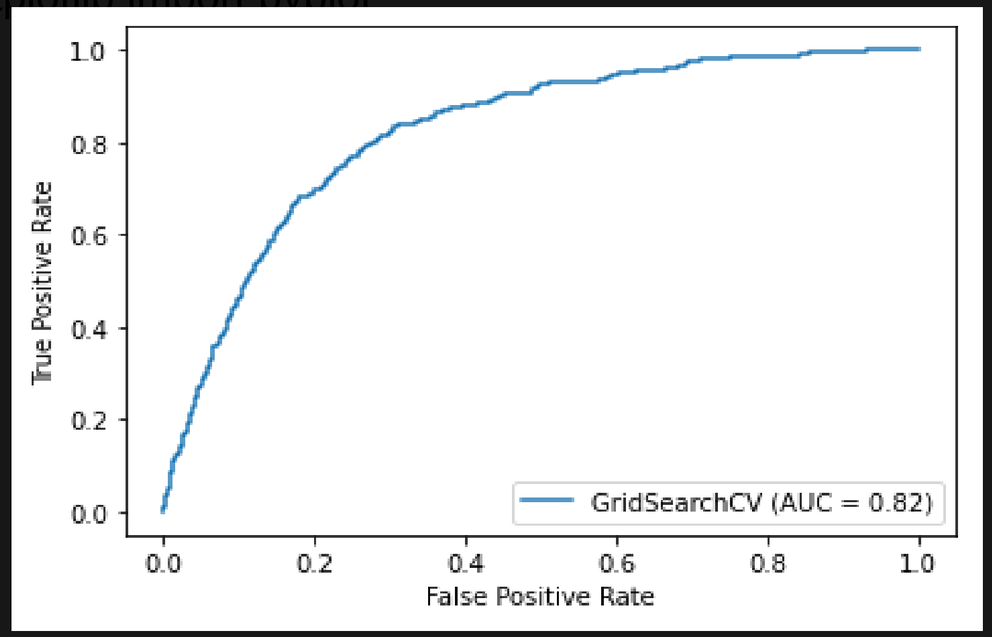
Job,marital,education,contact ,default,housing,loan,month, subscribed

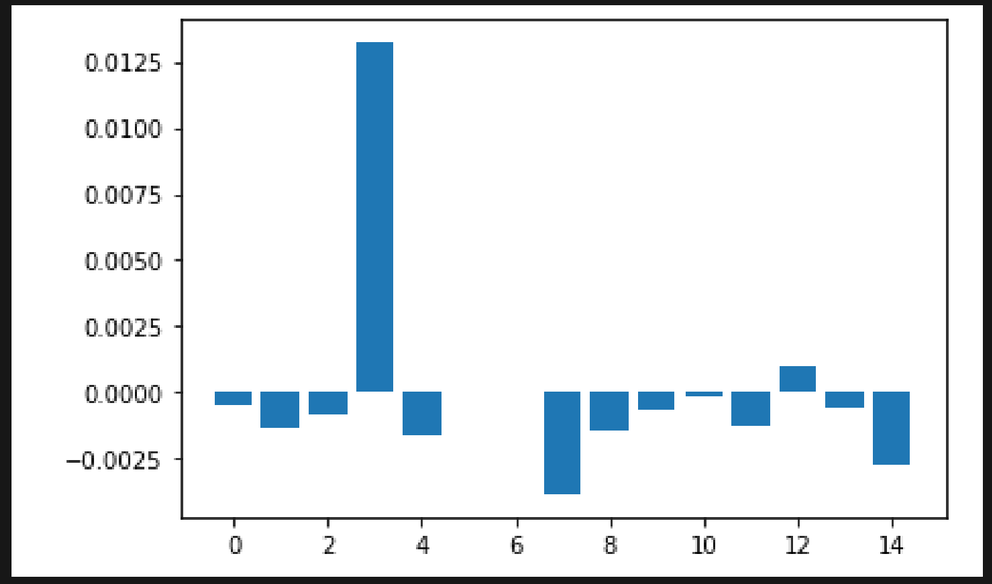
Standar Scaller fit_transform() both Xtrain & Xtest

Machine Learning Modelling

from matplotlib import pyplot

04

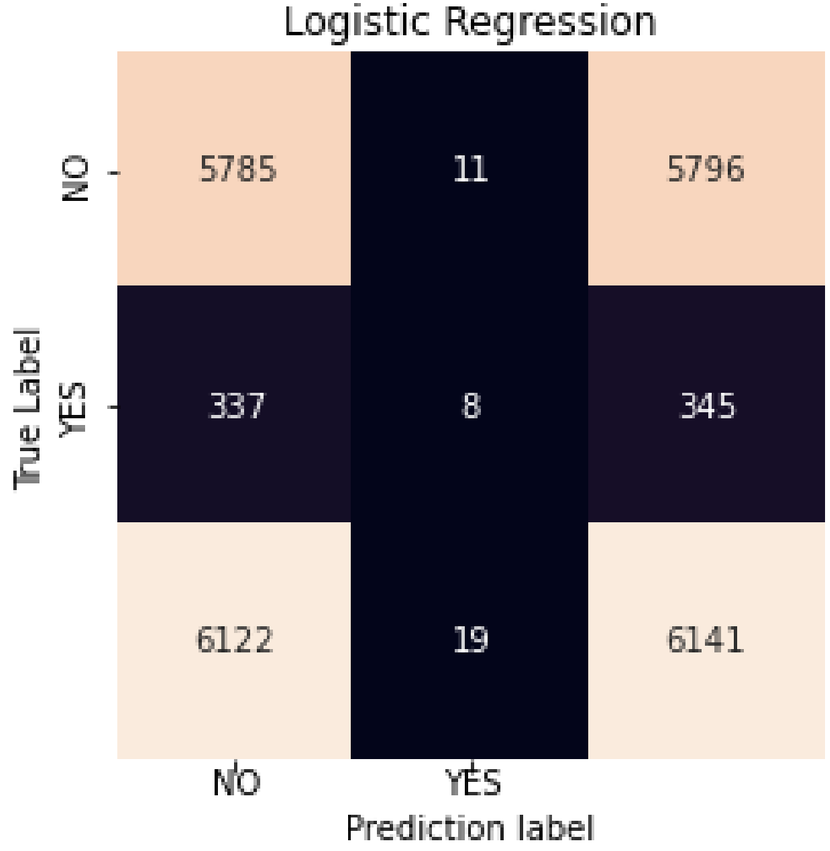




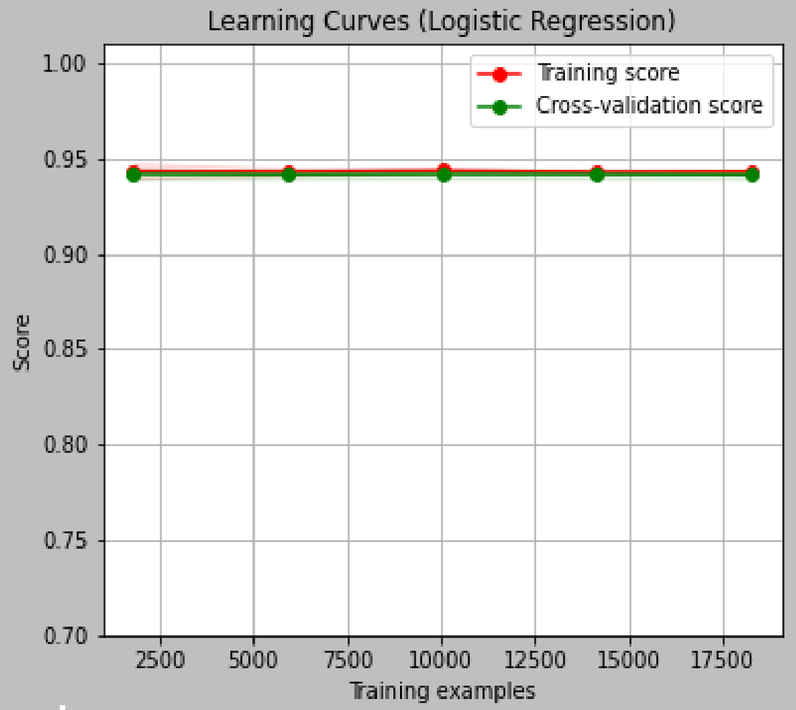
Age is the most features important of this model followed by campaign

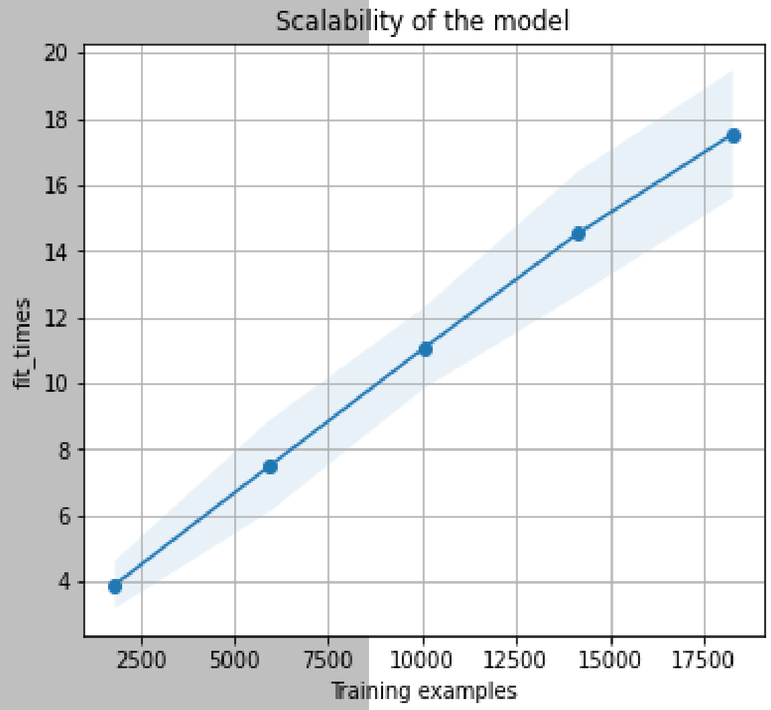
Logistic Regression

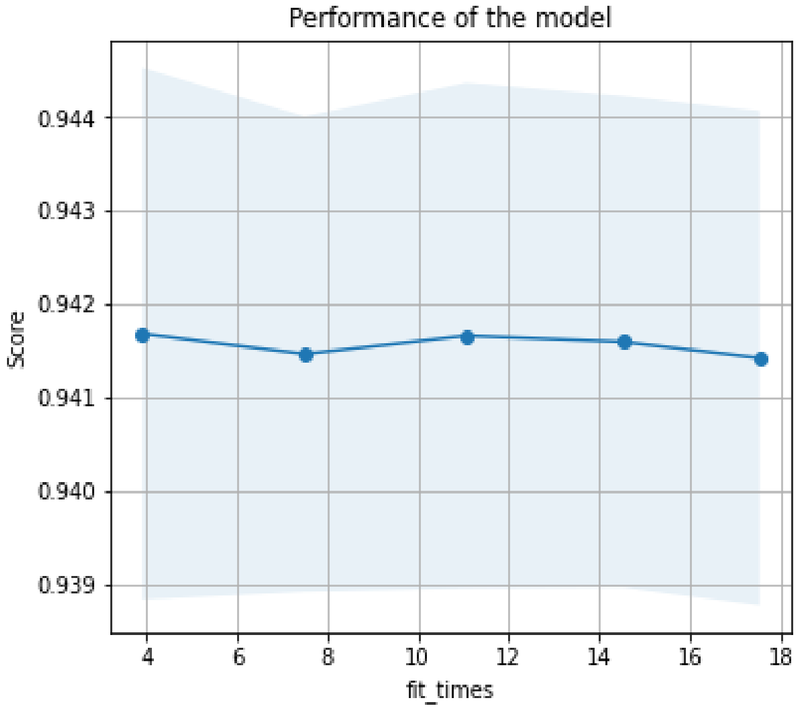
Test Accuracy : 94%
Train Accuracy : 94%

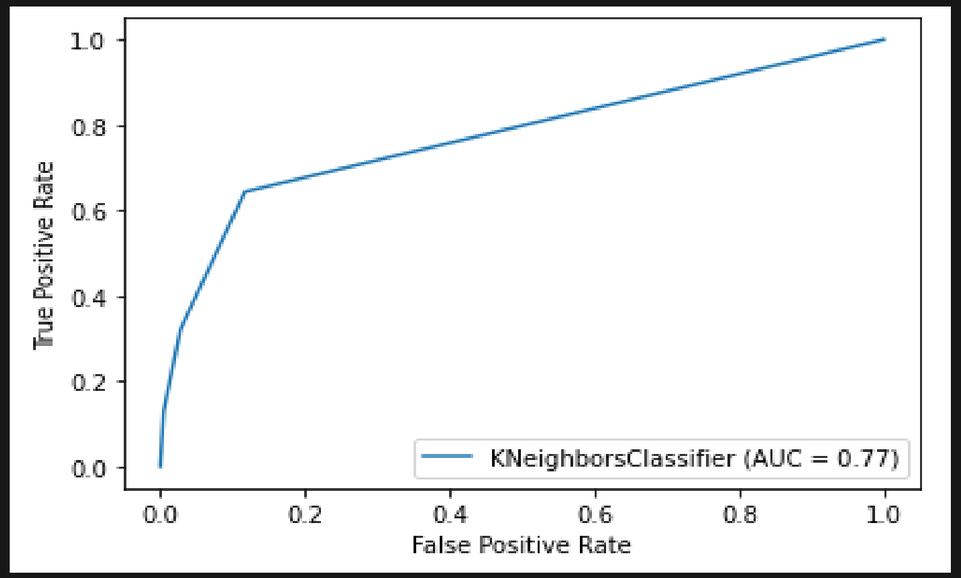


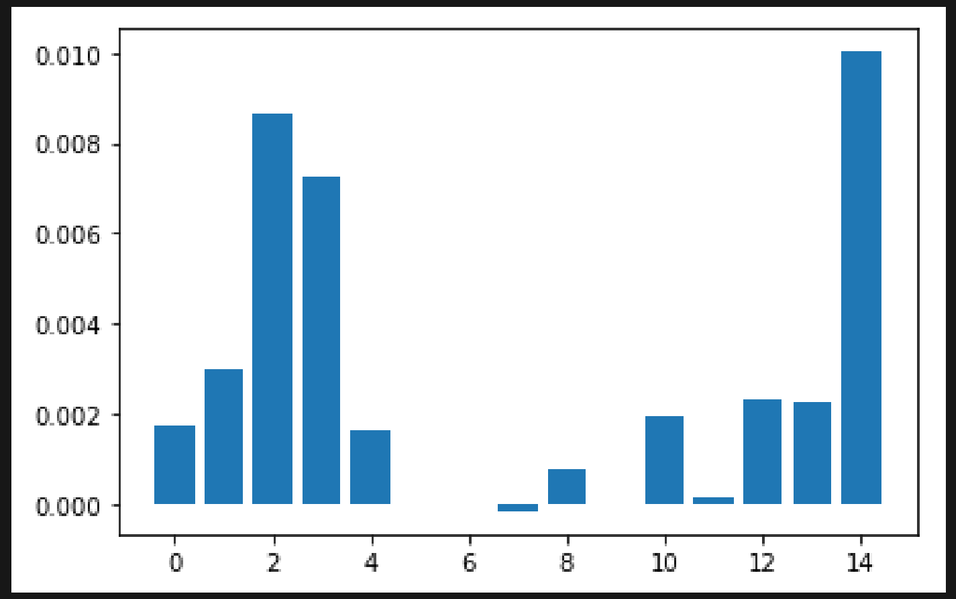
Logistic Regression







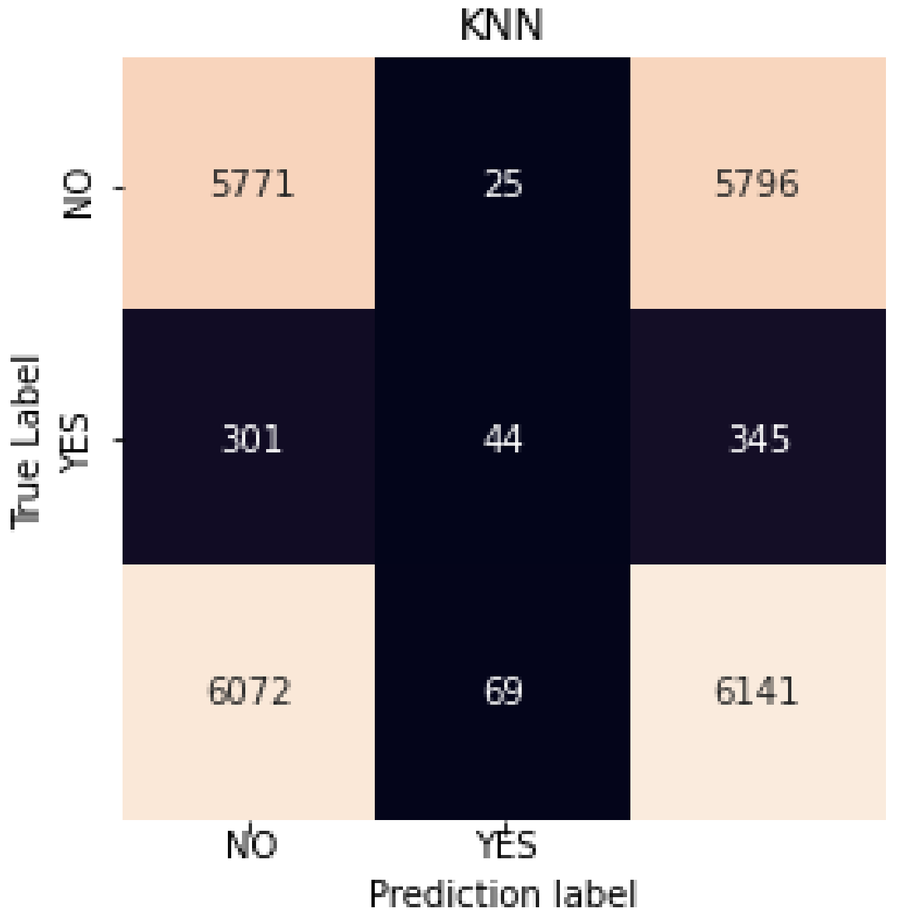




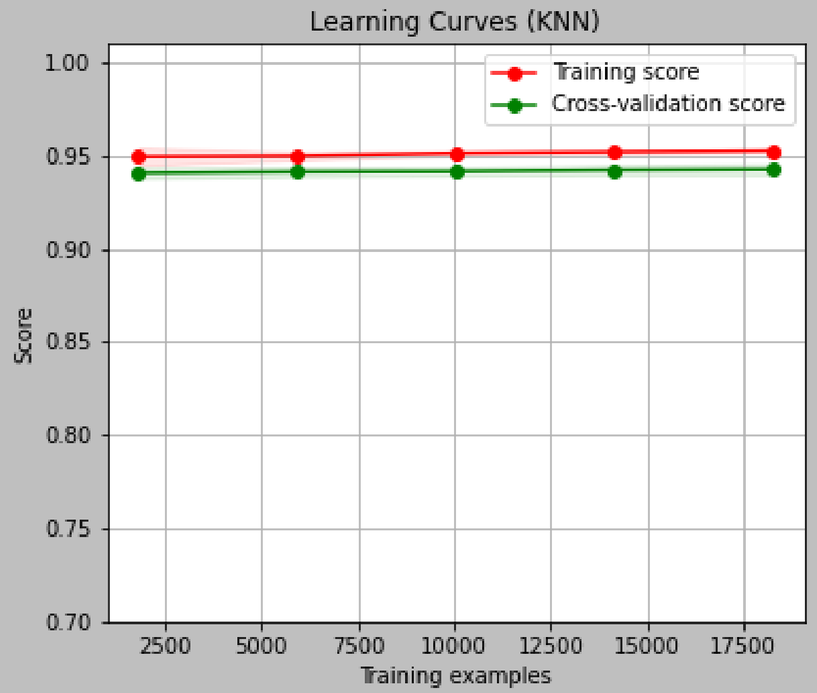
previous is the most features important of this model followed by marital and education.

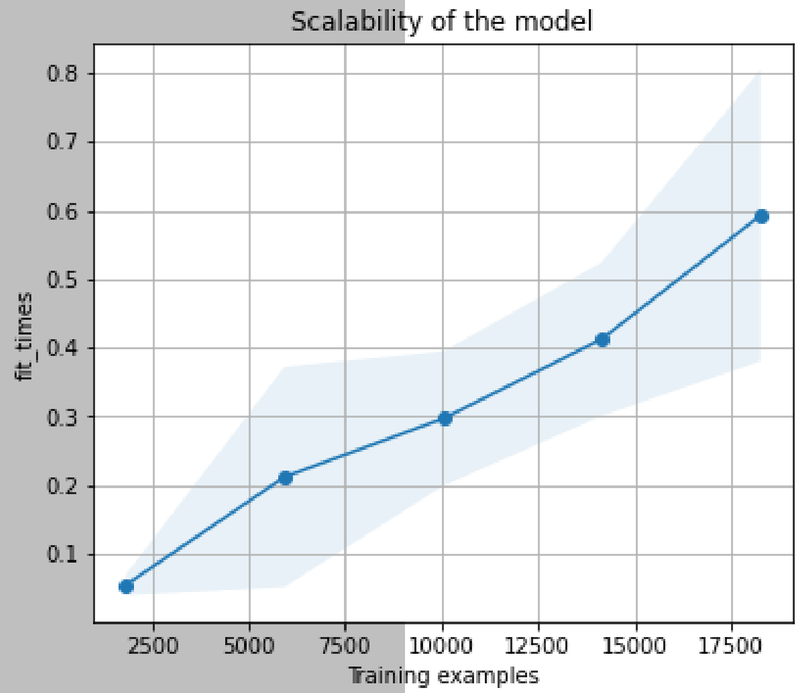
K-Nearest Neighbours

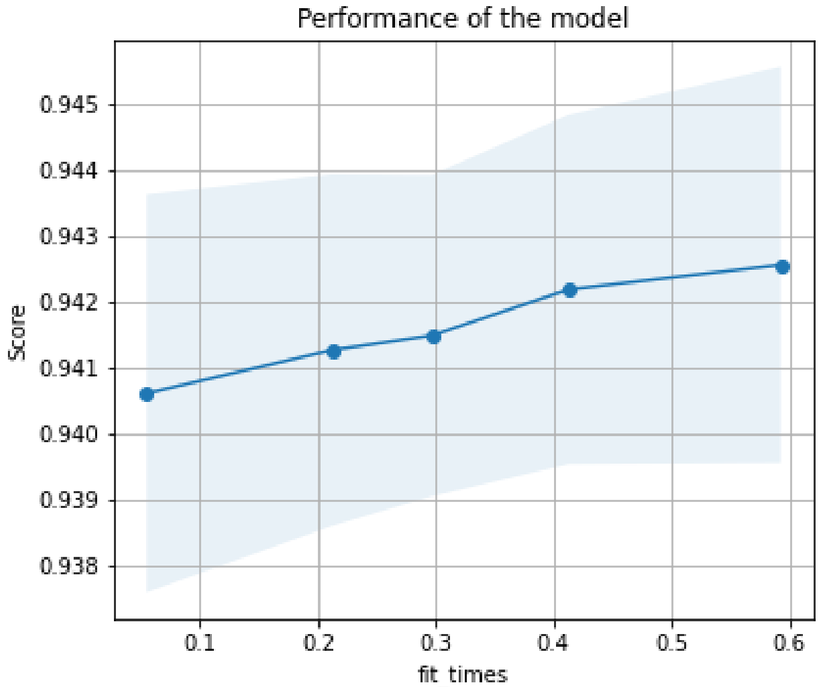
Test Accuracy : 94%
Train Accuracy : 95%

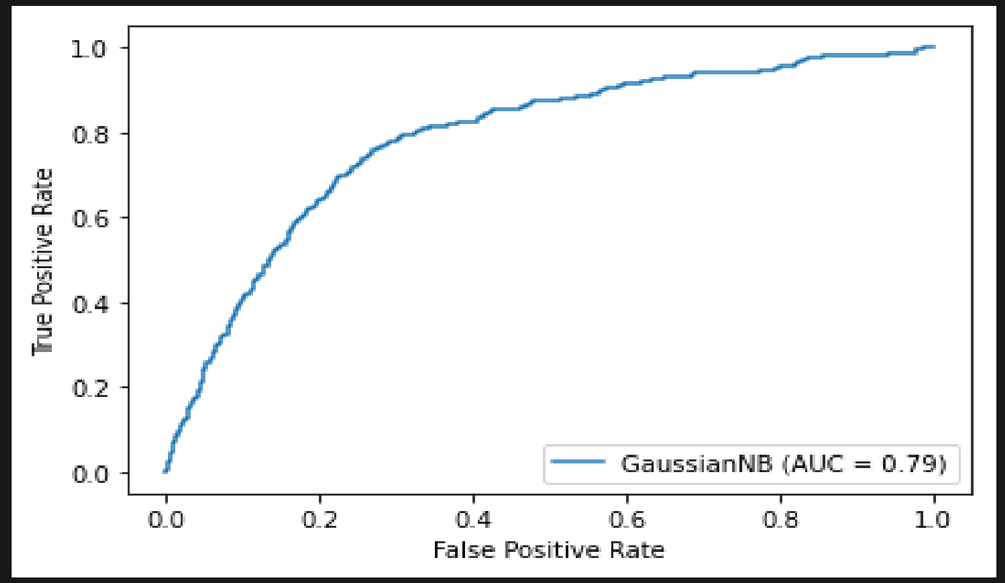


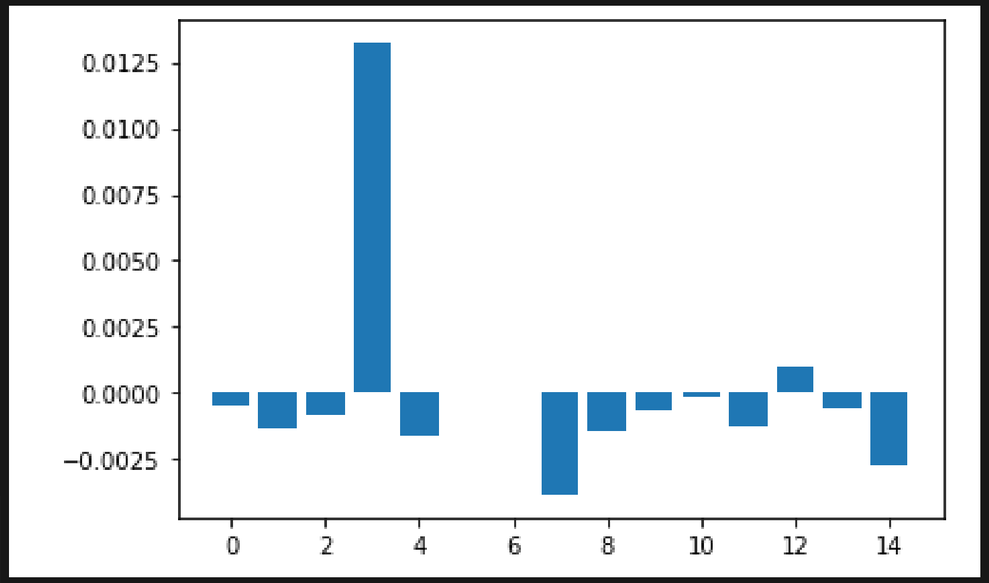
K-Nearest Neighbours







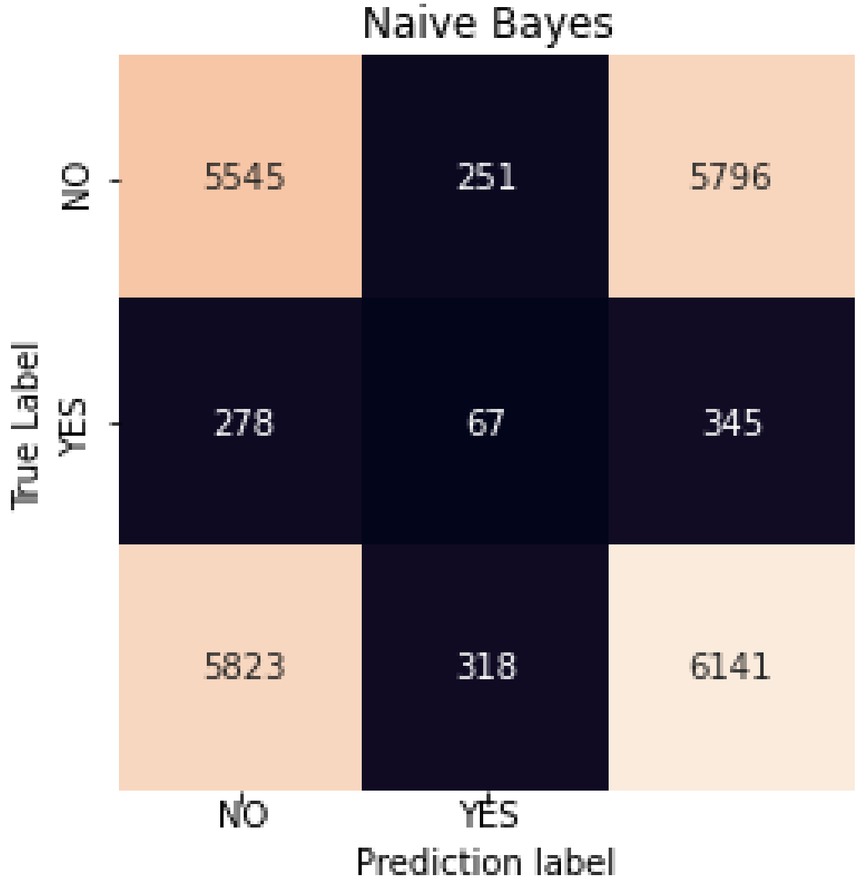




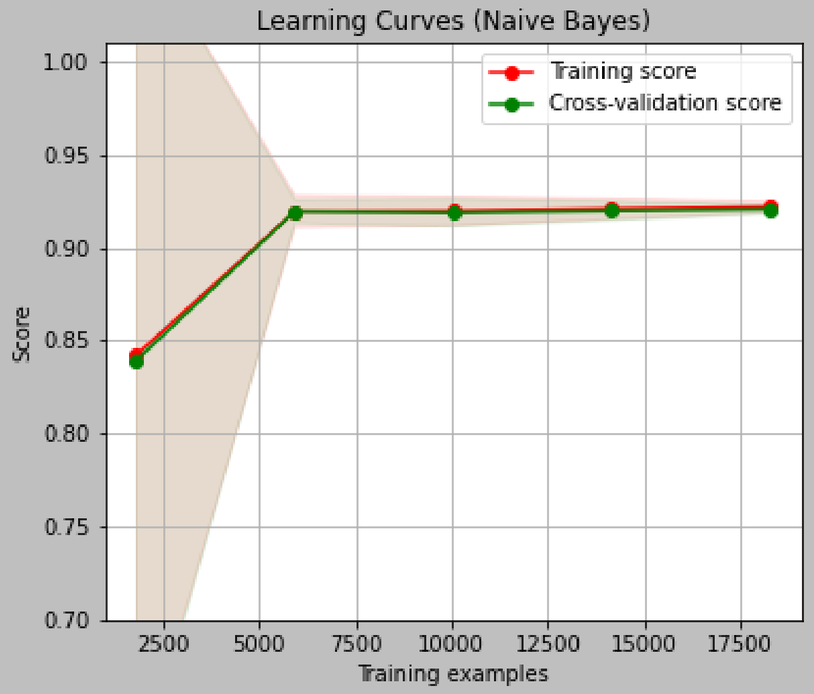
Age is the most features important of this model followed by campaign

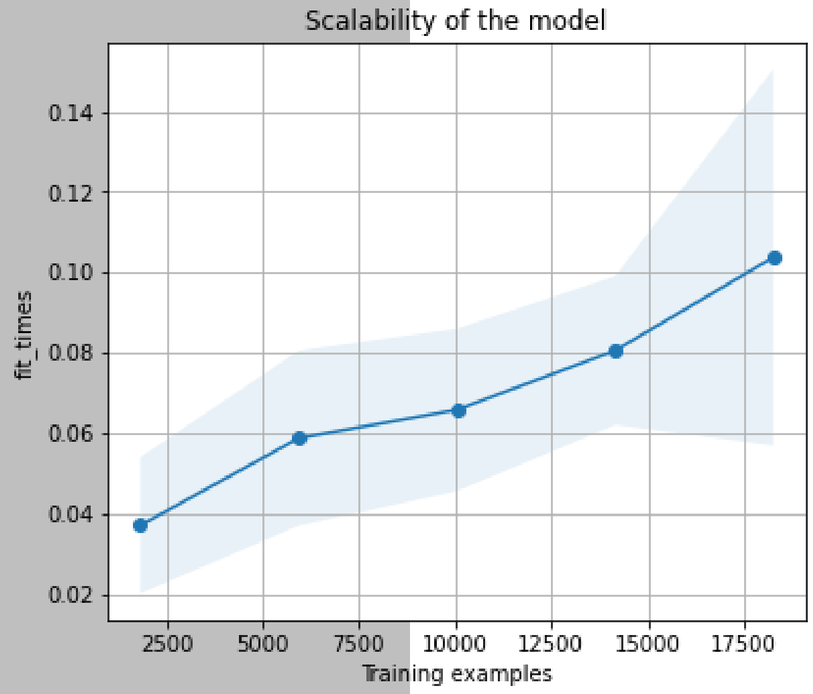
Naïve Bayes

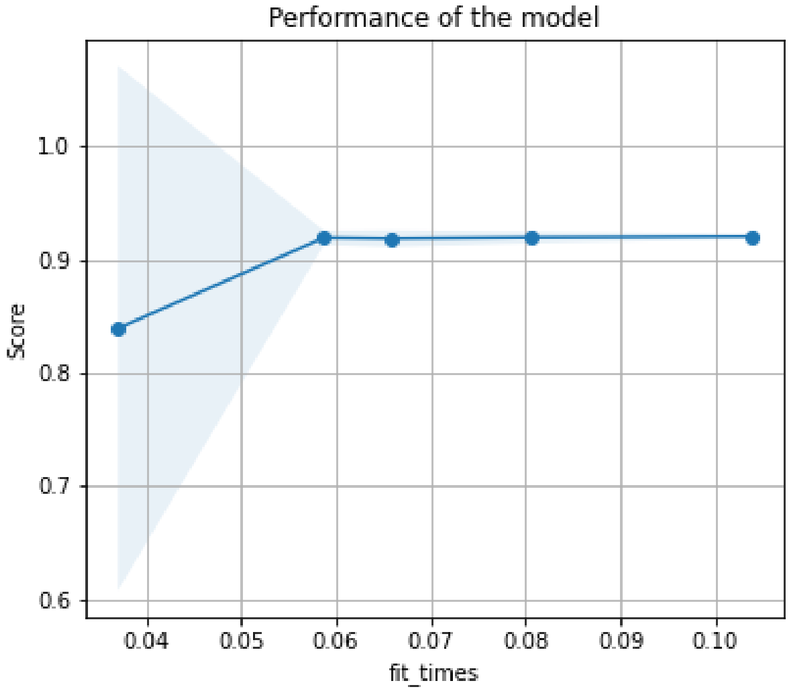
Test Accuracy : 91%
Train Accuracy : 92%

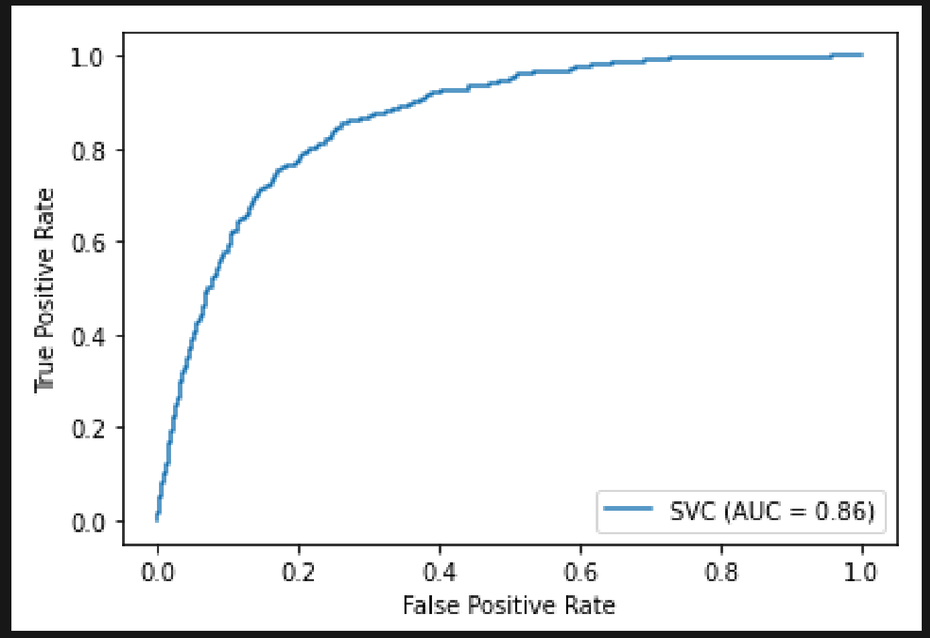


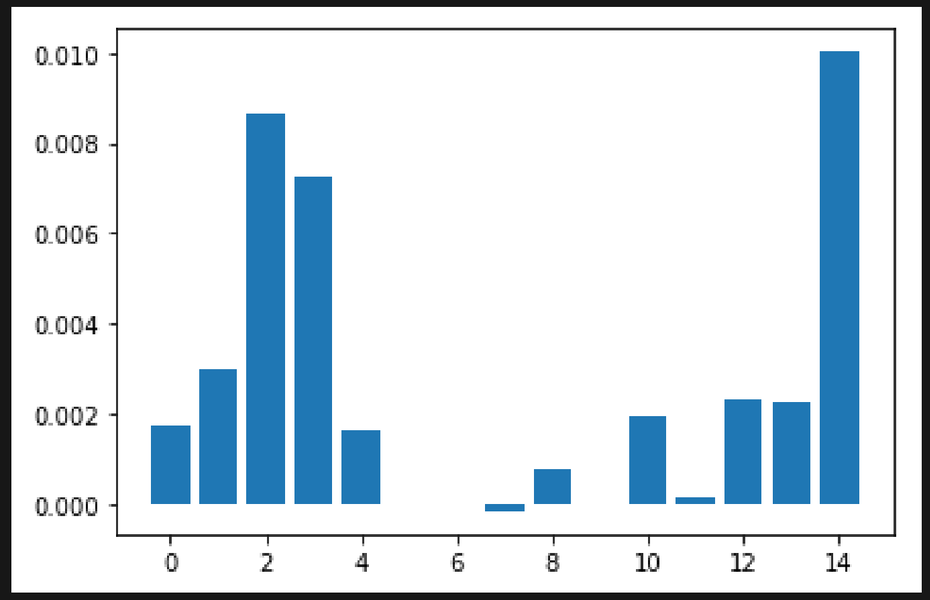
Naïve Bayes


Learning Curves (Naive Bayes)
Training score
Cross-validation score
Score
Training examples
1.00
0.95
0.90
0.85
0.80
0.75
0.70
2500
5000
7500
10000
12500
15000
17500


Scalability of the model
fit_times
Training examples
0.14
0.12
0.10
0.08
0.06
0.04
0.02
2500
5000
7500
10000
12500
15000
17500


Performance of the model
Score
fit_times
1.0
0.9
0.8
0.7
0.6
0.04
0.05
0.06
0.07
0.08
0.09
0.10

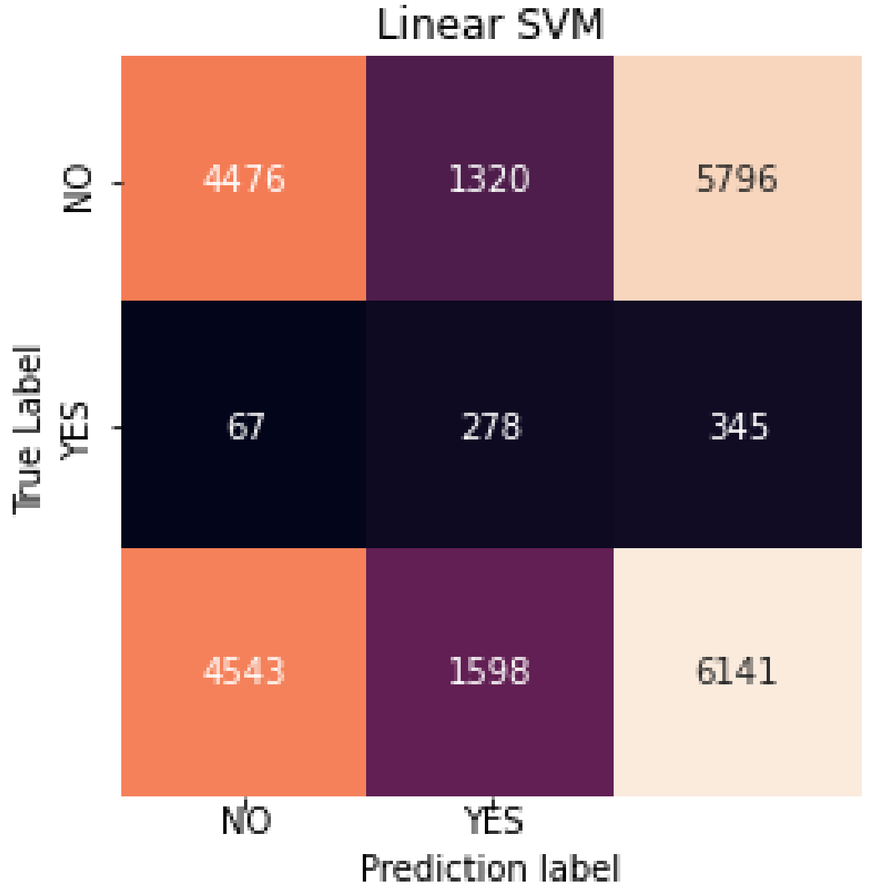




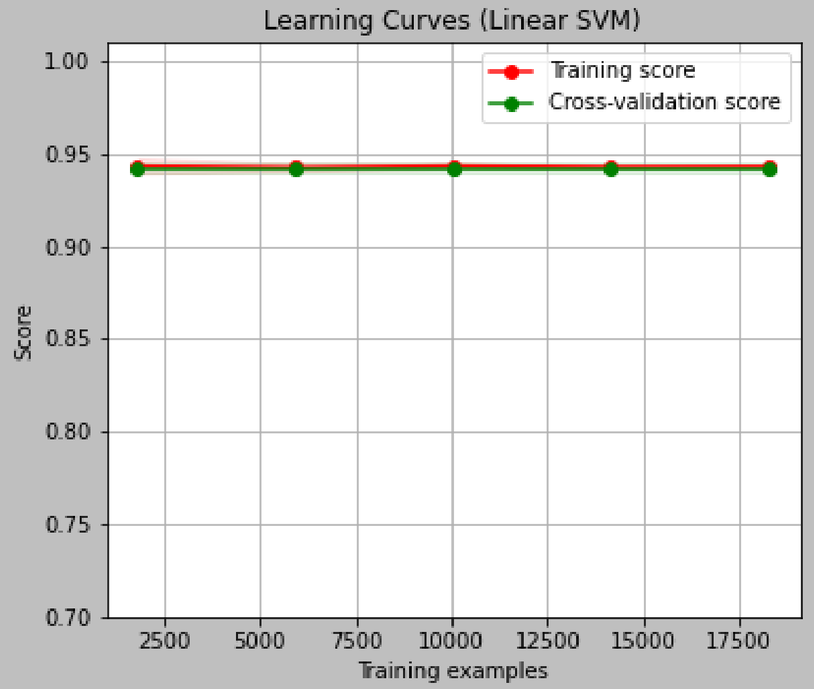
previous is the most features important of this model followed by marital and education.

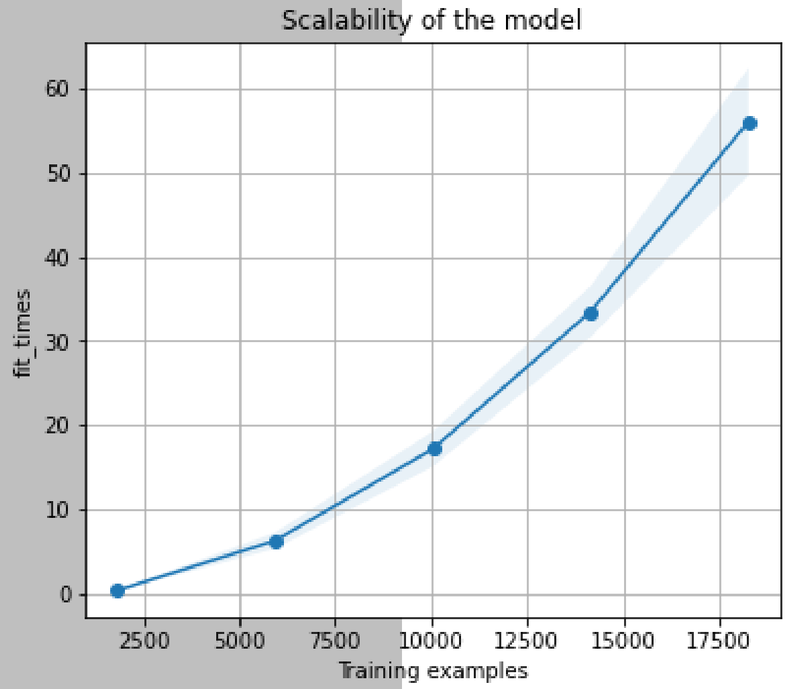
Support Vector Machine

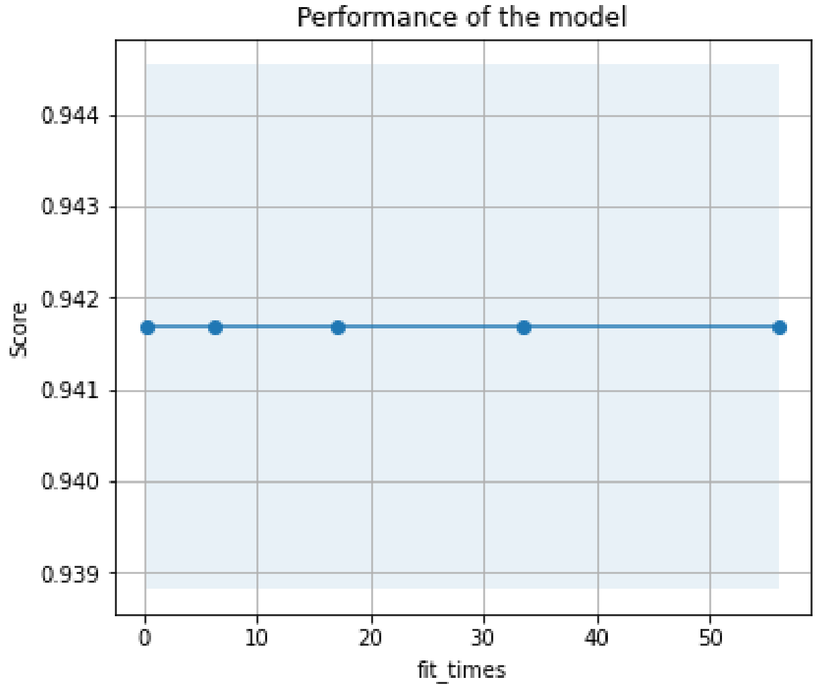
Test Accuracy : 94%
Train Accuracy : 94%



Support Vector Machine







Conclusion :

A. SVM and logistic regression have the highest accuracy among all but AUC graph in SVM has highest value than all of models.

B. For learning curve in Cross Validation Score, KNN has overfitting but another model are well fitting.

C. From scalability of model, Naive Bayes has the fastest fit times but SVM has the slowest fit times.

D. From the performance model graph, SVM has the constant performance score while the fit times increase

E. SVM can be used as the baseline model to predict deposits subscribed by telephonic marketing

F. SVM and KNN has the same importance features (previous,marital,education) and also for Logistic Regression and Naïve Bayes has the same importance features (Age and campaign)

Q & A Session

Special Thanks to :

And All Mentor Dibimbing + Mbak Rini & Mbak Miranda

Thank You ☺