CASE STUDY
BANK MARKETING

-CATHERINE MARIANA PHILIPS-

PROBLEM STATEMENT

The data is related with direct marketing campaigns of a banking institution. The marketing campaigns were based on phone calls

The objective is to predict if the client will subscribe (yes/no) to a term deposit, by building classification model using Machine Learning algorithms.

VARIABLE DESCRIPTION

Parameter	Description
age	Age of the clients (numeric)
job	The job type of the clients (categorical)
marital	Marital status of clients (categorical)
education	Education level of clients (categorical)
default	has credit in default? (Categorical: "yes", "no")
balance	average yearly balance, in euros (numeric)
housing	has housing loan? (Categorical: "yes", "no")
loan	has personal loan? (Categorical: "yes", "no")
contact	contact communication type (categorical: "unknown", "telephone", "cellular")
day	last contact day of the month (numeric)
month	last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
duration	last contact duration, in seconds (numeric)
campaign	number of contacts performed during this campaign and for this client (numeric, includes last contact)
pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
previous	number of contacts performed before this campaign and for this client (numeric)
poutcome	outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")
deposit	has the client subscribed a term deposit? (Categorical: "yes", "no")

- The bank marketing dataset consists of 5581 rows and 17 attributes.
- Out of these 17, 10 attributes are of categorical datatype.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	41	services	married	unknown	no	88	yes	no	cellular	11	may	105	1	336	2	failure	no
1	56	technician	married	secondary	no	1938	no	yes	cellular	26	feb	229	1	192	4	success	yes
2	30	services	single	secondary	no	245	no	yes	cellular	8	jul	187	2	-1	0	unknown	no
3	34	management	single	tertiary	no	1396	yes	no	cellular	17	jul	630	1	-1	0	unknown	no
4	29	technician	single	secondary	no	-13	yes	no	cellular	14	may	512	3	-1	0	unknown	no

SAMPLE DATA

- The required libraries are imported.
  - OS: To change the working directory
  - Numpy: To perform numerical operations
  - Pandas: To work with dataframes
- The dataset is imported.
- The dataset consists of 17 attributes out of which 10 attributes are of the categorical type.
- There are no null values in the dataset.
- Duplicate values are not present.

EXPLORATORY DATA ANALYSIS

UNDERSTANDING THE DATA
-DATA TYPES OF THE ATTRIBUTES-

```
DATA TYPES OF THE ATTRIBUTES
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5581 entries, 0 to 5580
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        5581 non-null   int64
 1   job        5581 non-null   object
 2   marital    5581 non-null   object
 3   education  5581 non-null   object
 4   default    5581 non-null   object
 5   balance    5581 non-null   int64
 6   housing    5581 non-null   object
 7   loan       5581 non-null   object
 8   contact    5581 non-null   object
 9   day        5581 non-null   int64
 10  month      5581 non-null   object
 11  duration   5581 non-null   int64
 12  campaign   5581 non-null   int64
 13  pdays      5581 non-null   int64
 14  previous   5581 non-null   int64
 15  poutcome   5581 non-null   object
 16  deposit    5581 non-null   object
dtypes: int64(7), object(10)
memory usage: 784.8+ KB
```

There are 7 attributes of int64 type and 10 attributes of the object type.

UNDERSTANDING THE DATA
-SUMMARY OF NUMERICAL DATA-

	age	balance	day	duration	campaign	pdays	previous
count	5581.000000	5581.000000	5581.000000	5581.000000	5581.000000	5581.000000	5581.000000
mean	41.169683	1514.736786	15.693603	368.175954	2.507436	52.534313	0.849669
std	11.926044	3266.534626	8.461086	344.131053	2.770717	110.754995	2.311684
min	18.000000	-3058.000000	1.000000	3.000000	1.000000	-1.000000	0.000000
25%	32.000000	110.000000	8.000000	137.000000	1.000000	-1.000000	0.000000
50%	39.000000	542.000000	15.000000	254.000000	2.000000	-1.000000	0.000000
75%	49.000000	1747.000000	22.000000	485.000000	3.000000	57.000000	1.000000
max	93.000000	81204.000000	31.000000	3284.000000	63.000000	842.000000	41.000000

- 75% of the clients are of 49 years of age.
- The yearly balance of 75% of the clients is **€1747**
- The last day when 75% of the clients were contacted was the 31st of every month.
- 50% of the clients were contacted for the first time and 75% of the clients were contacted after 842 days.

UNDERSTANDING THE DATA
-SUMMARY OF CATEGORICAL DATA-

	job	marital	education	default	housing	loan	contact	month	poutcome	deposit
count	5581	5581	5581	5581	5581	5581	5581	5581	5581	5581
unique	12	3	4	2	2	2	3	12	4	2
top	management	married	secondary	no	no	no	cellular	may	unknown	no
freq	1318	3134	2719	5497	2928	4863	4044	1407	4133	2959

- Around 1318 clients have a job in **Management**.
- 3134 clients are married.
- 2719 clients have secondary education.
- The number of clients who do not have credit in default is 5497.
- 2928 clients have taken housing loan.
- 4863 clients have taken a personal loan.
- Around 4044 clients can be contacted through cellular mode of communication.
- During the month of May, 1407 clients were contacted.
- The outcome of the previous campaign for 4133 clients seems to be unknown.
- 2959 clients have not subscribed to a term deposit

UNDERSTANDING THE DATA
-CATEGORICAL DATA: UNIQUE VALUES-

```
Job
management       1318
blue-collar       975
technician        887
admin.            661
services          452
retired           397
self-employed     206
student           182
unemployed        170
entrepreneur      160
housemaid         143
unknown            30
Name: job, dtype: int64
------------------------------
Marital Status
married     3134
single      1816
divorced     631
Name: marital, dtype: int64
------------------------------
Education
secondary    2719
tertiary     1871
primary       746
unknown       245
Name: education, dtype: int64
------------------------------
```

```
Default
no      5497
yes       84
Name: default, dtype: int64
------------------------------
Housing
no      2928
yes     2653
Name: housing, dtype: int64
------------------------------
Loan
no      4863
yes      718
Name: loan, dtype: int64
------------------------------
Contact
cellular     4044
unknown      1155
telephone     382
Name: contact, dtype: int64
------------------------------
```

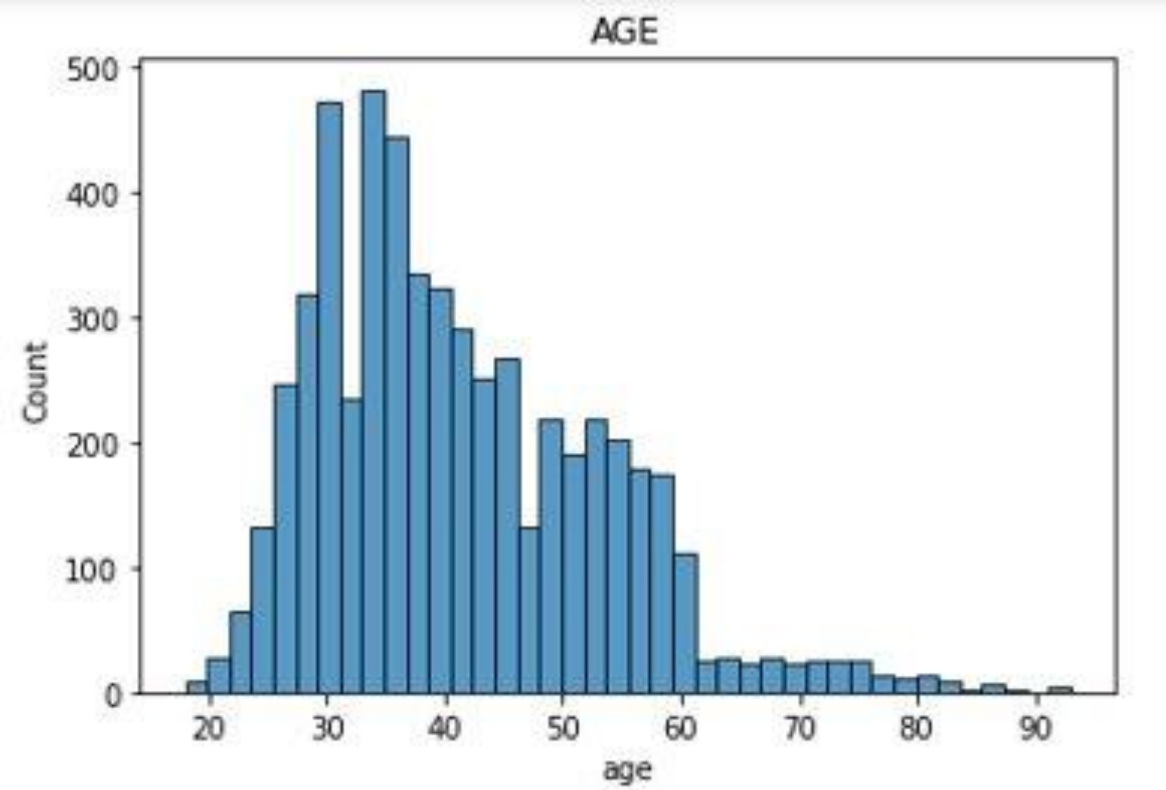
```
Month
may    1407
aug     757
jul     752
jun     627
nov     478
apr     450
feb     383
oct     190
jan     180
sep     168
mar     129
dec      60
Name: month, dtype: int64
------------------------------
P_Outcome
unknown    4133
failure     632
success     539
other       277
Name: poutcome, dtype: int64
------------------------------
Deposit
no     2959
yes    2622
Name: deposit, dtype: int64
```

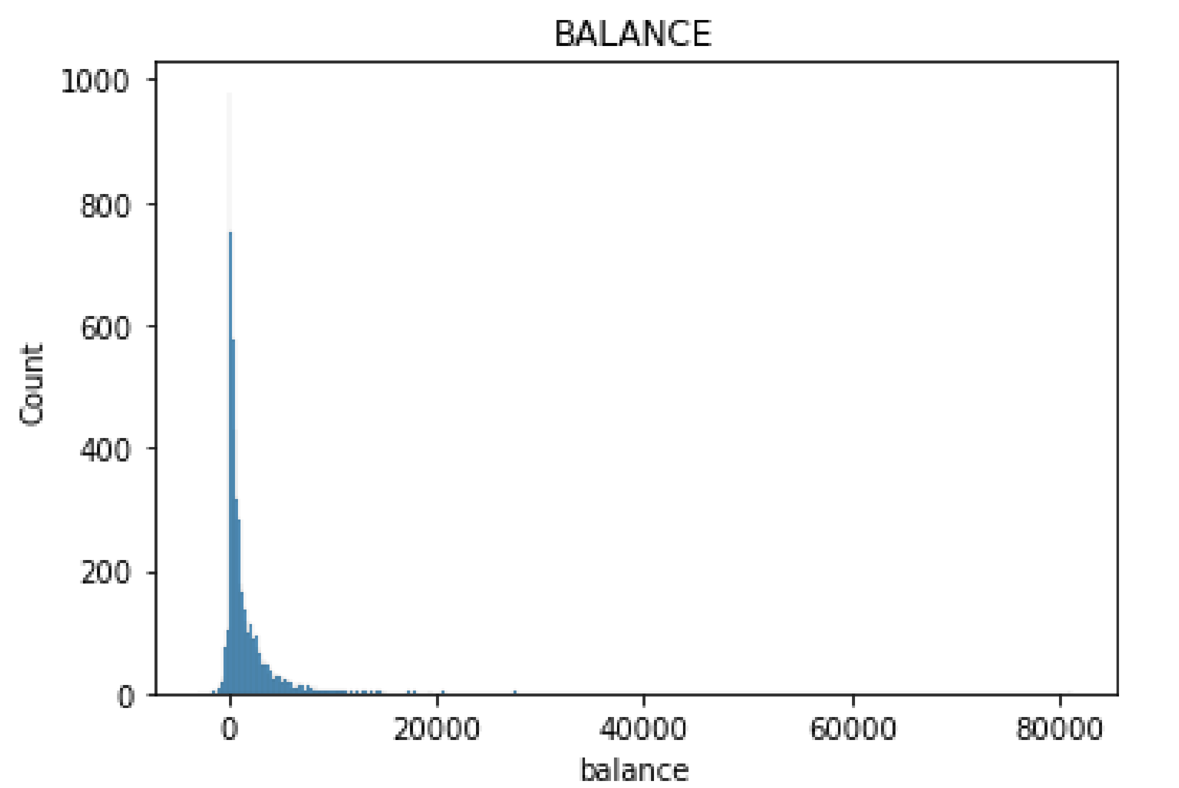
DATA VISUALIZATION

The libraries "seaborn", "plotly" and "matplotlib" are imported to help us in plotting the various graphs.

Tableau is also used for better visualization

DISTRIBUTIONS

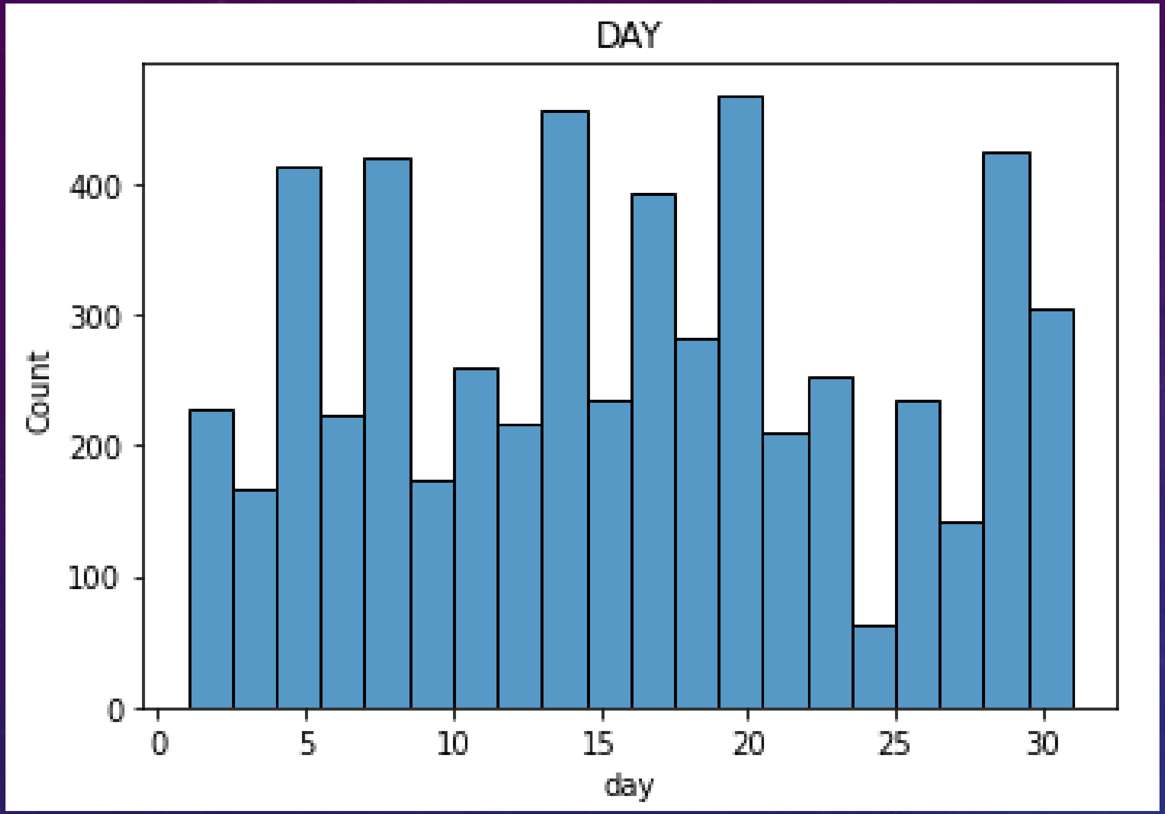


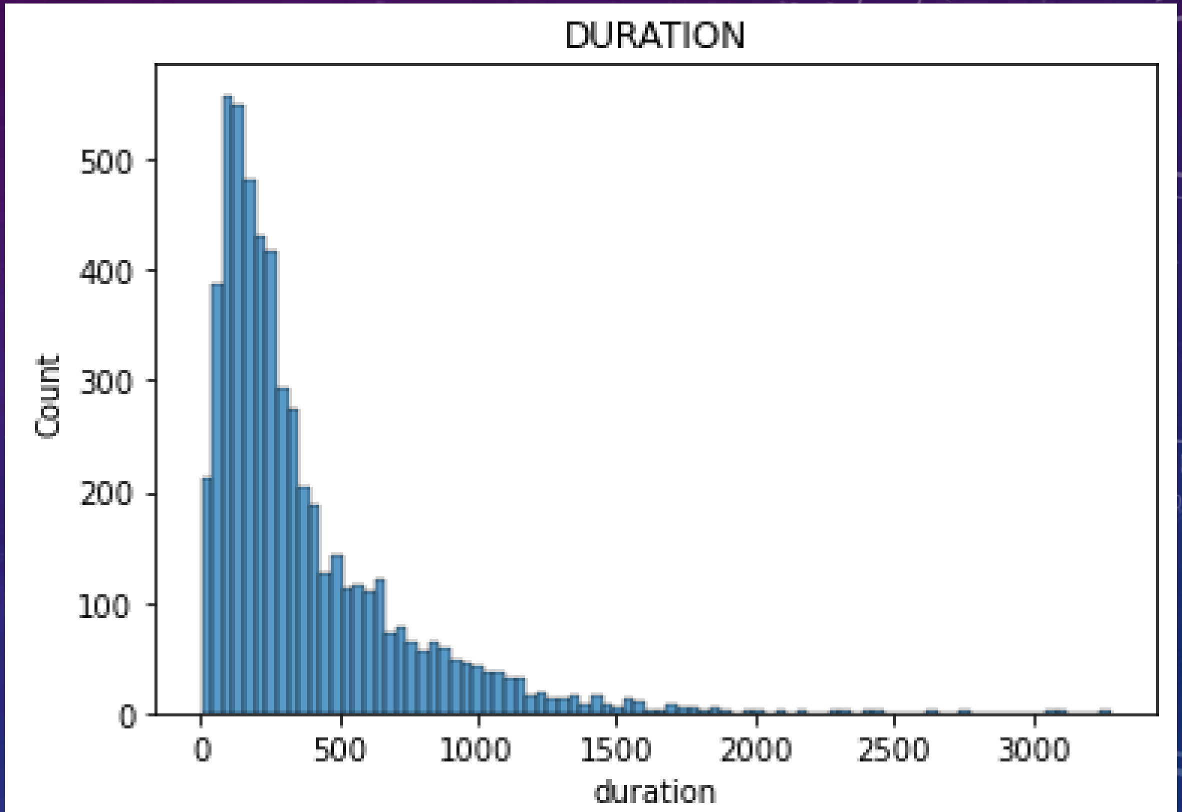


The distribution is slightly right skewed. **(median<mean)**

The distribution is right skewed. **(median<mean)**

DISTRIBUTIONS

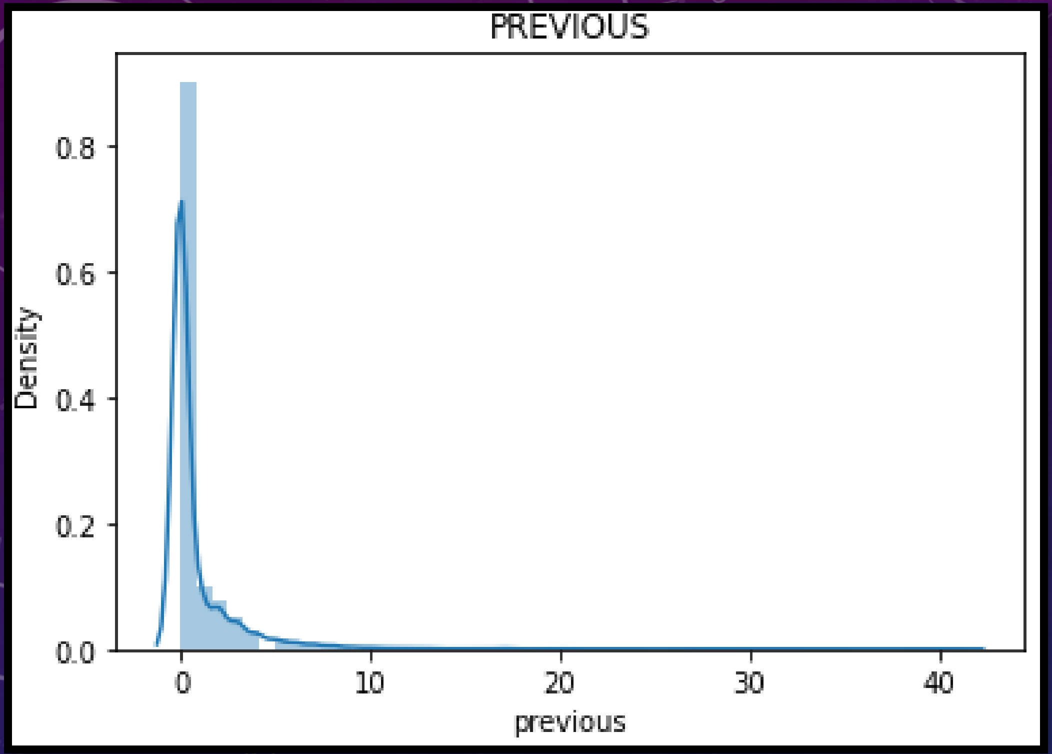


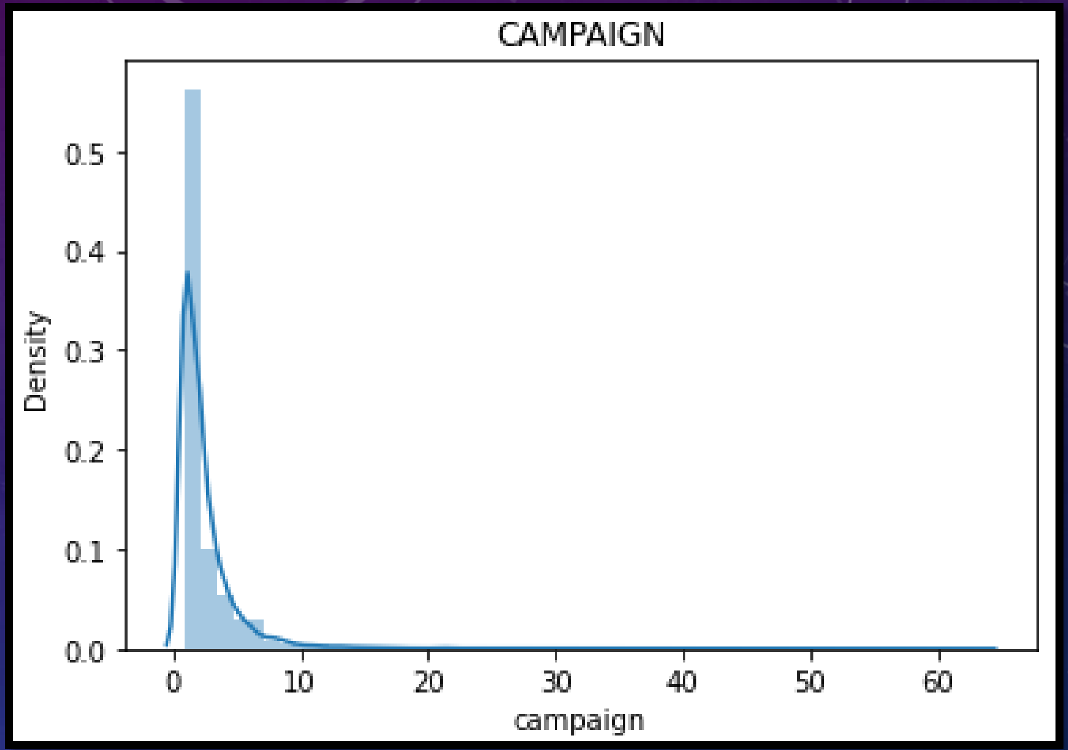


The distribution is normally distributed. (**mean=median**)

The distribution is right skewed. (**median<mean**)

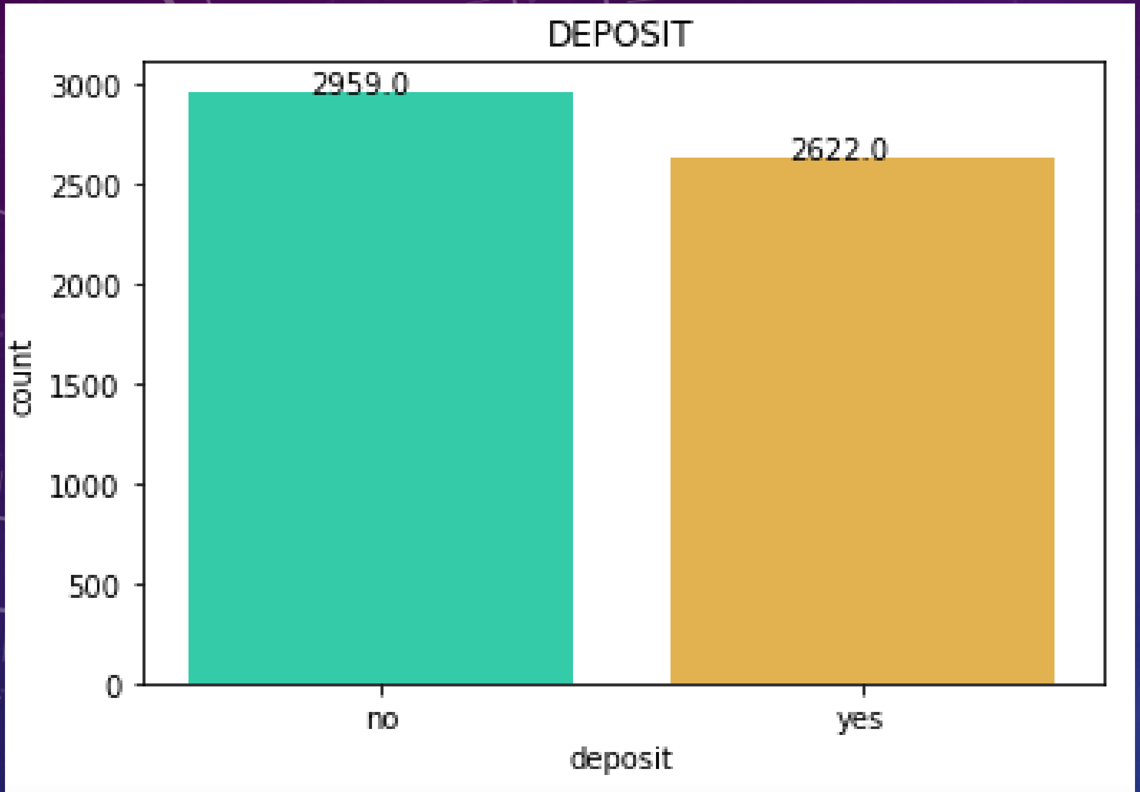
DISTRIBUTIONS





The distribution is right skewed for both "previous" and "campaign" (**median<mean**)

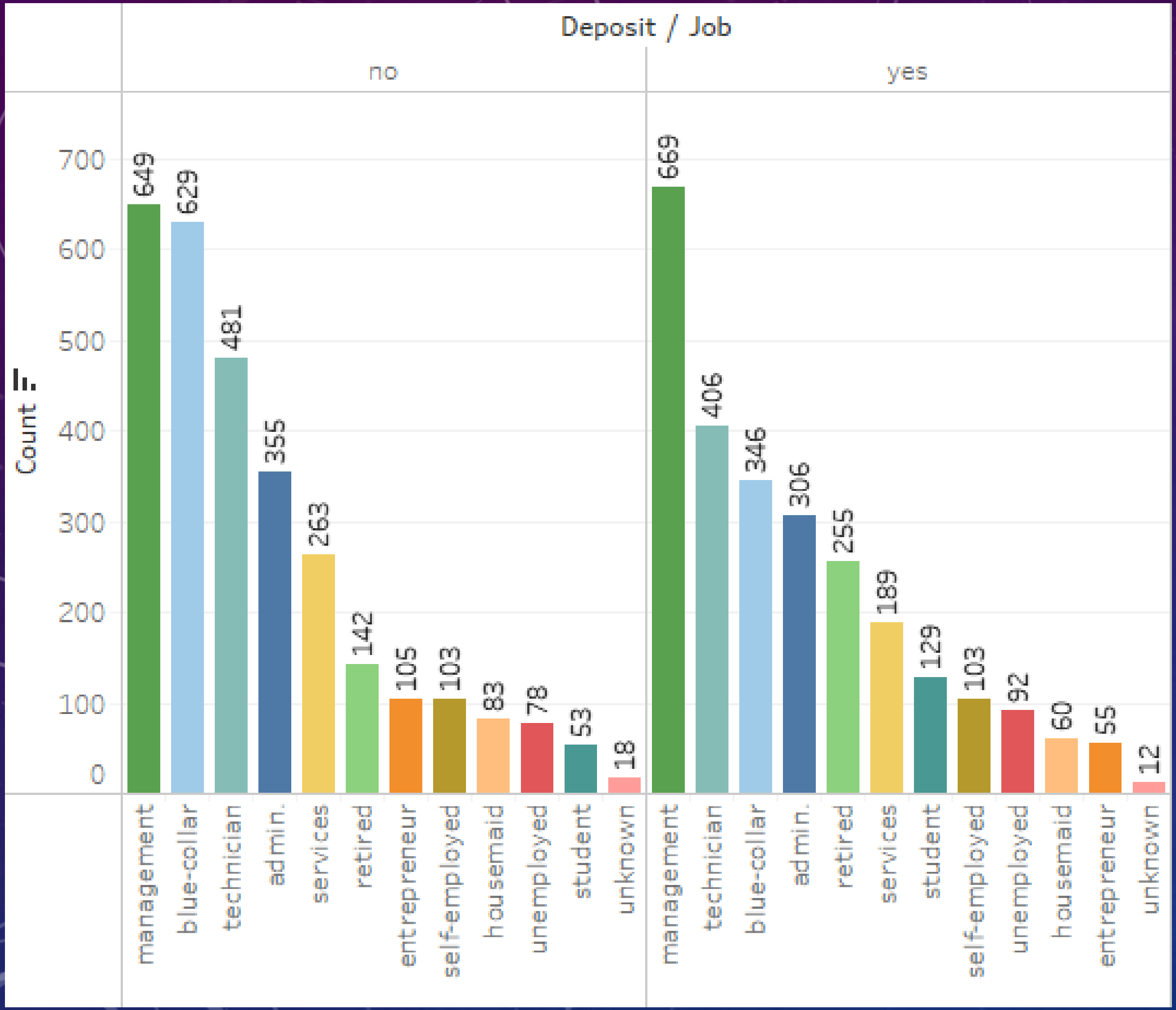
TARGET VARIALBE: DEPOSIT



- We can see that there is only a difference of 337 between the categories in the target variable.

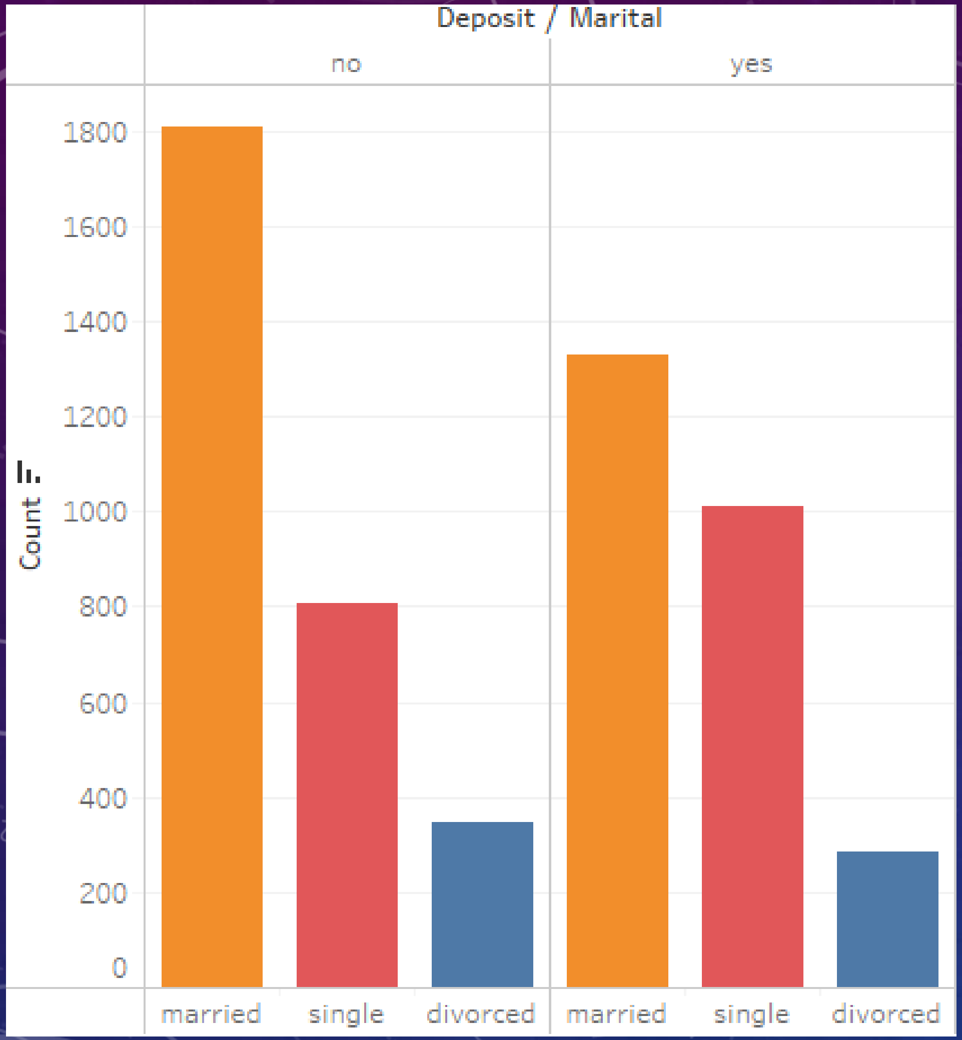
- Hence the **data is balanced**

Job and Deposit



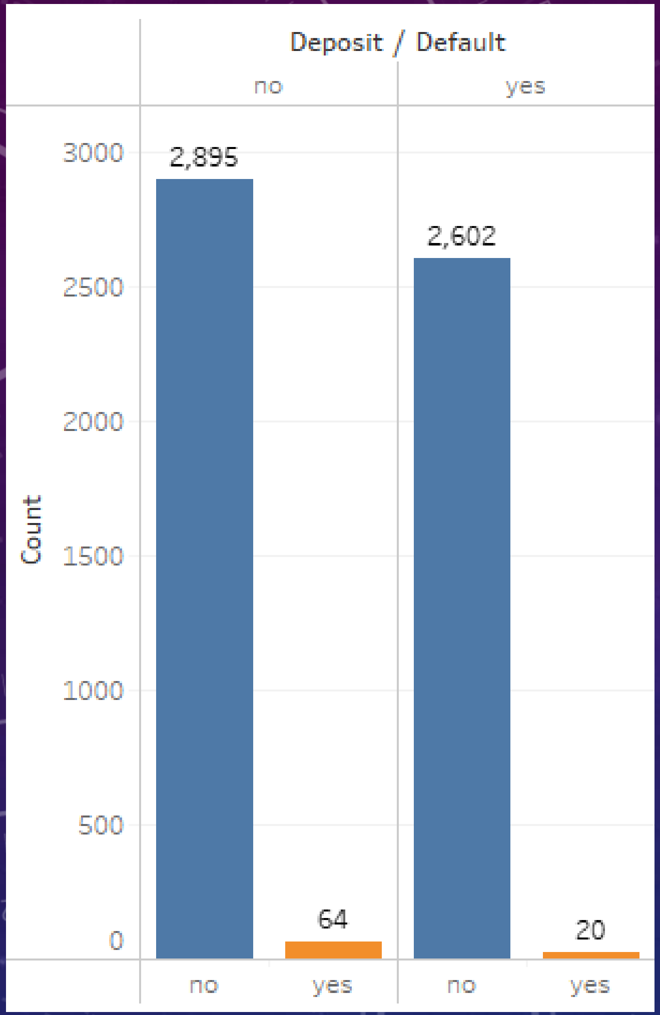
Customers who have a job in management have a higher rate of subscribing to term deposit, but they are also the highest when it comes to not subscribing.

Marital and Deposit



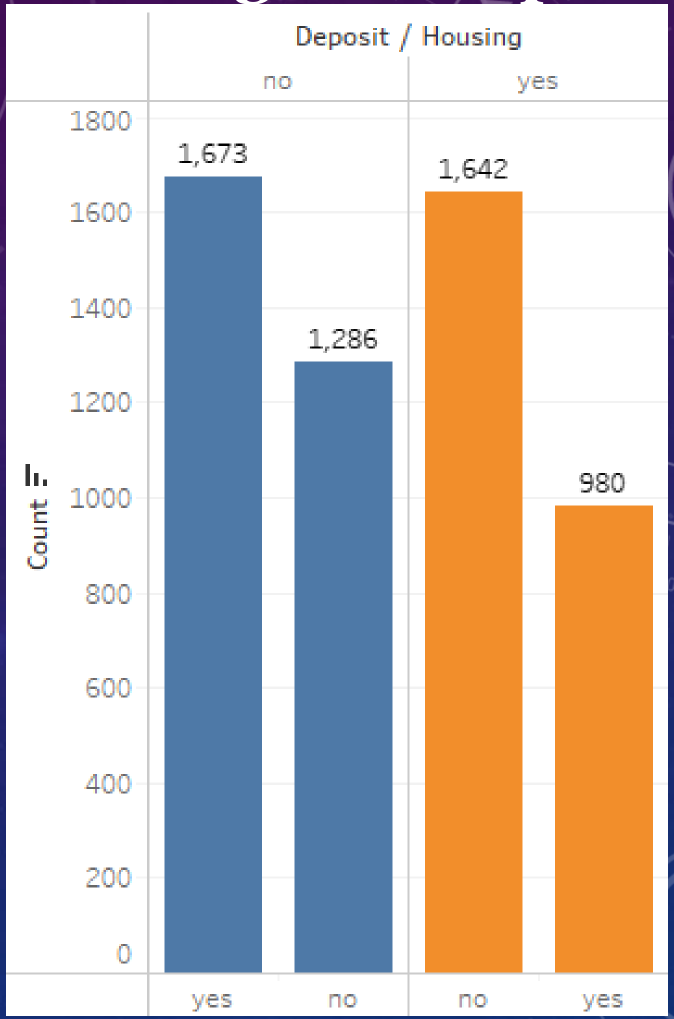
Majority of the customers are married followed by single and divorced.

Default and Deposit



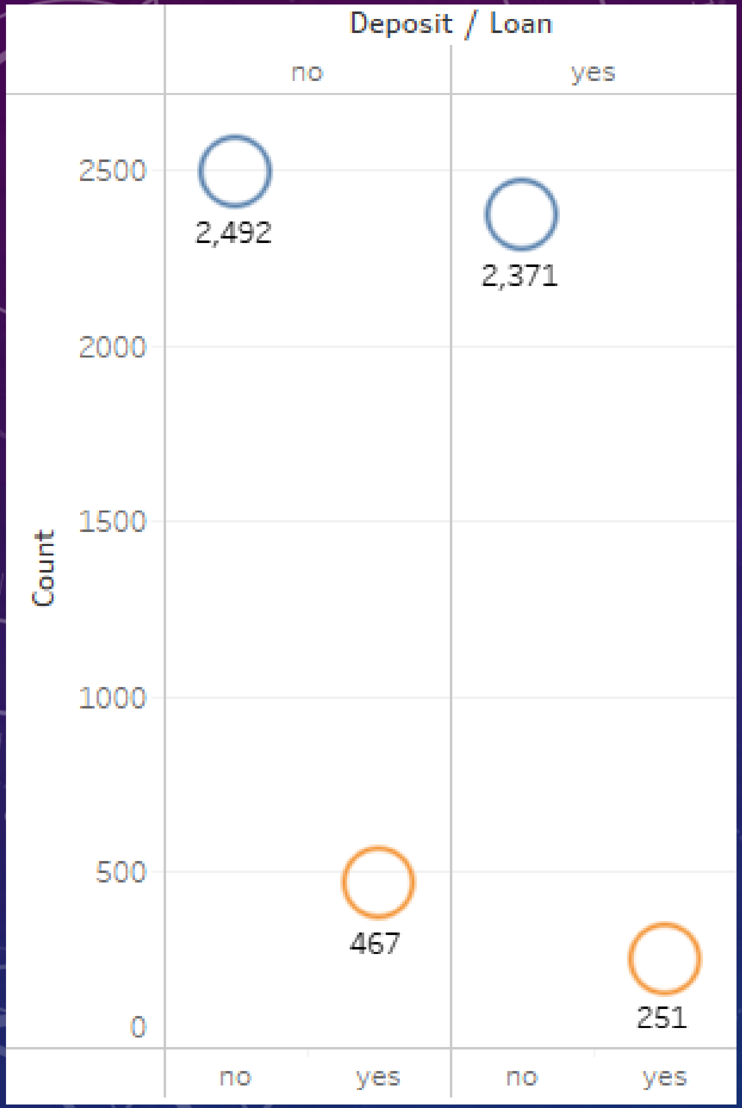
There is a uniform distribution among customers who do not have credit in default and so this feature contributes less in predicting if a customer will subscribe to a term deposit or not.

Housing and Deposit



Customers who have housing loan has lower rate of subscribing to term deposit.

Personal Loan and Deposit

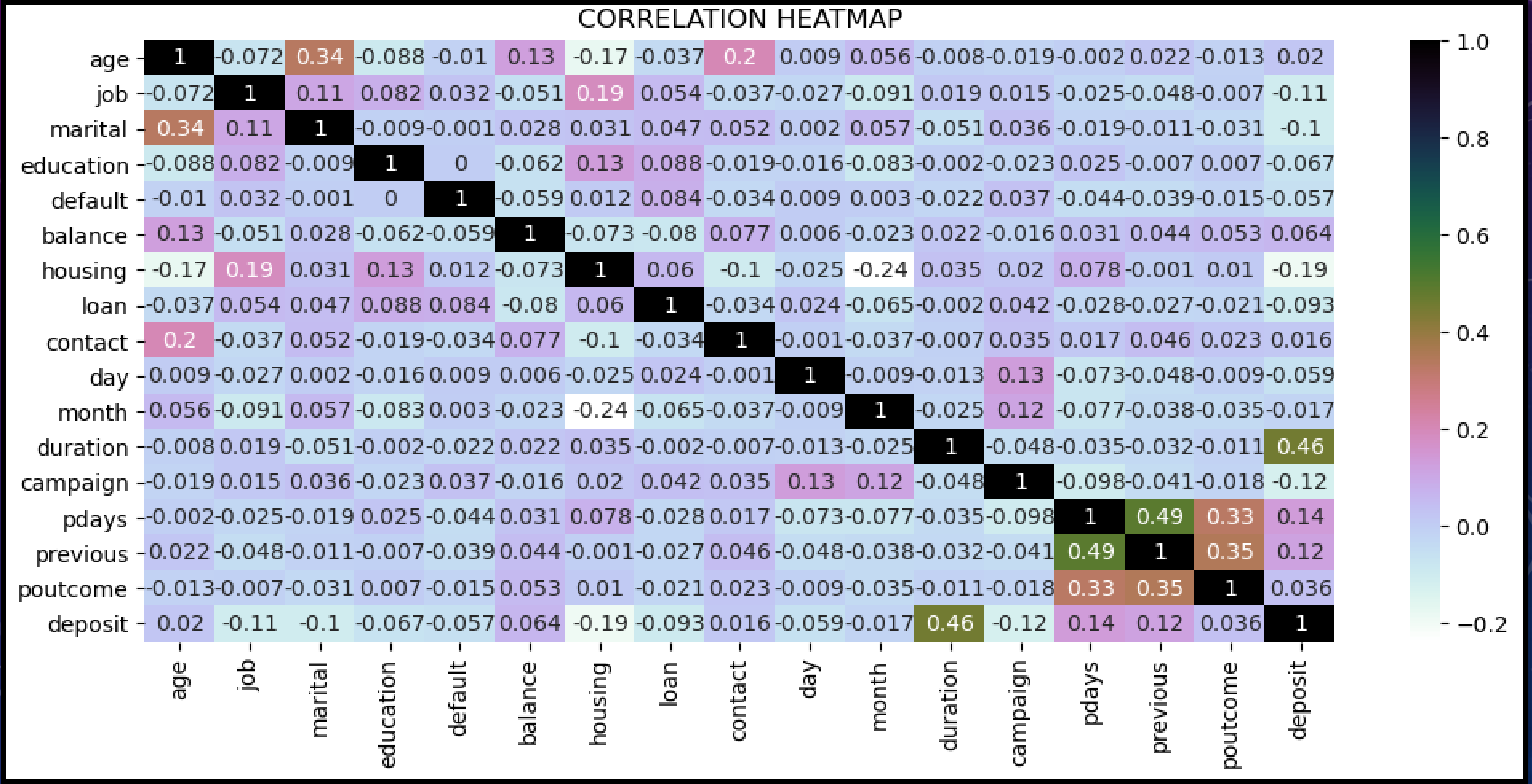


People who have not taken loan are more in number. Therefore people who have not taken loan are more likely to subscribe to term deposit.

CORRELATION

- Dummy encoding is done on the categorical variables.
- The correlation is found for the target variable and predictor variables.

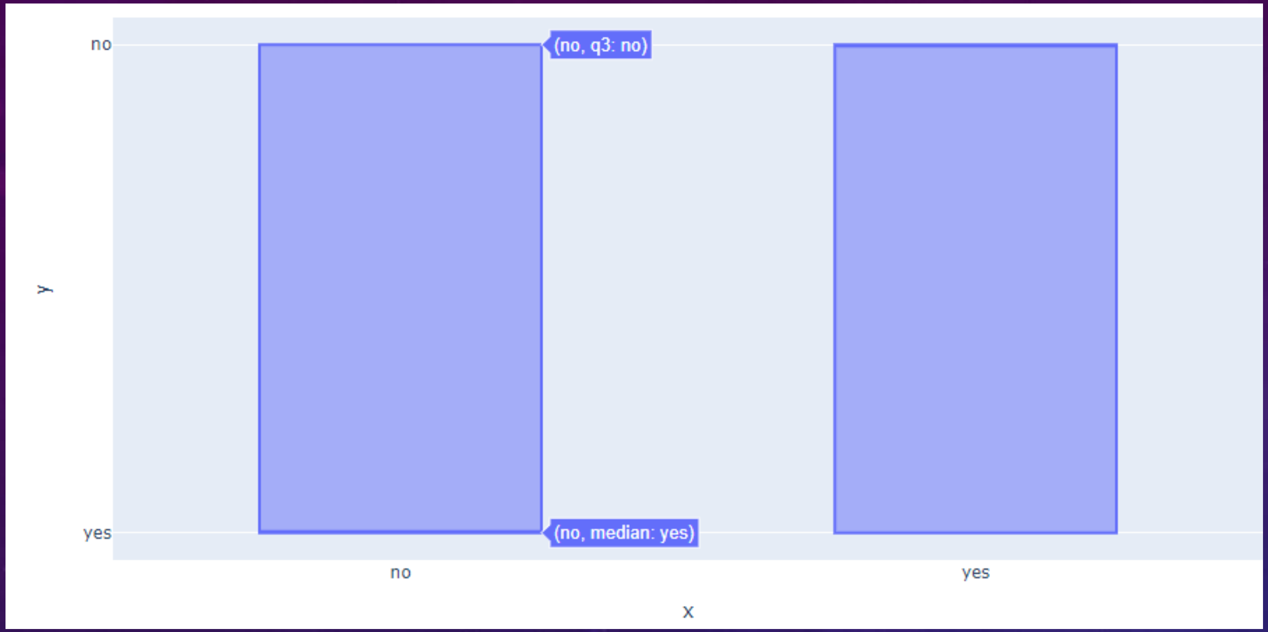
CORRELATION

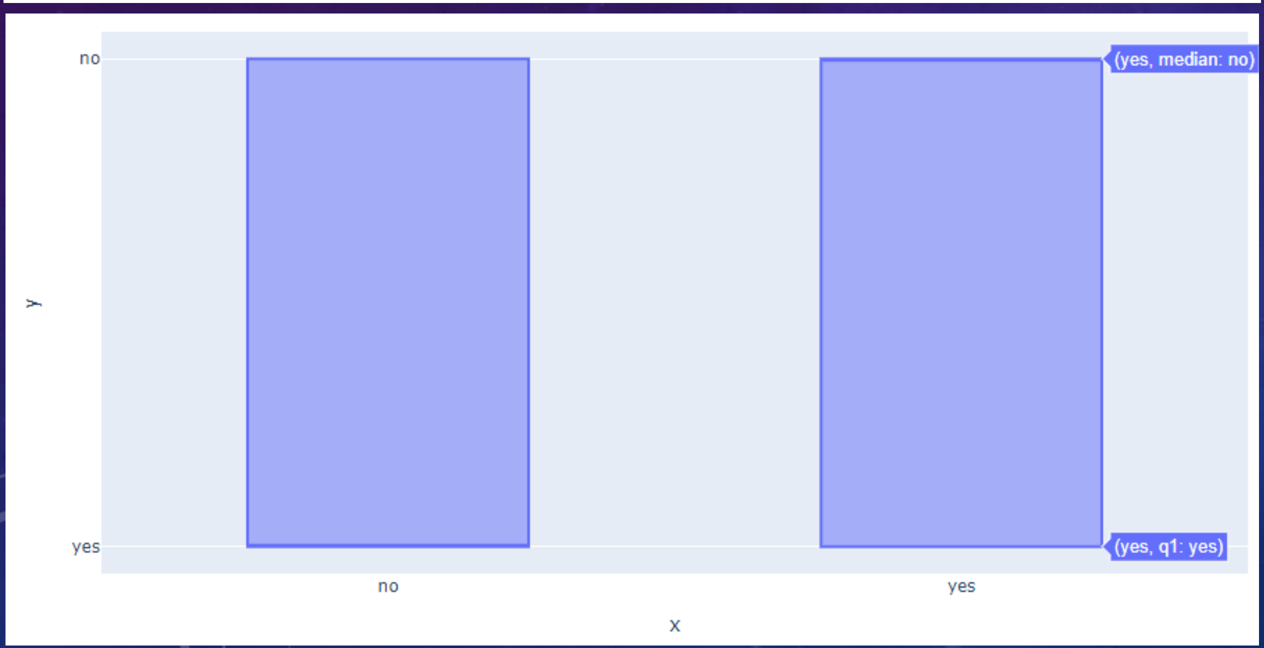


CORRELATION

- The variables job, campaign, housing are weakly and inversely correlated with the target variable Deposit.
- -The variables pdays and previous are weakly correlated with the target variable Deposit.
- -Duration has a strong correlation with Deposit

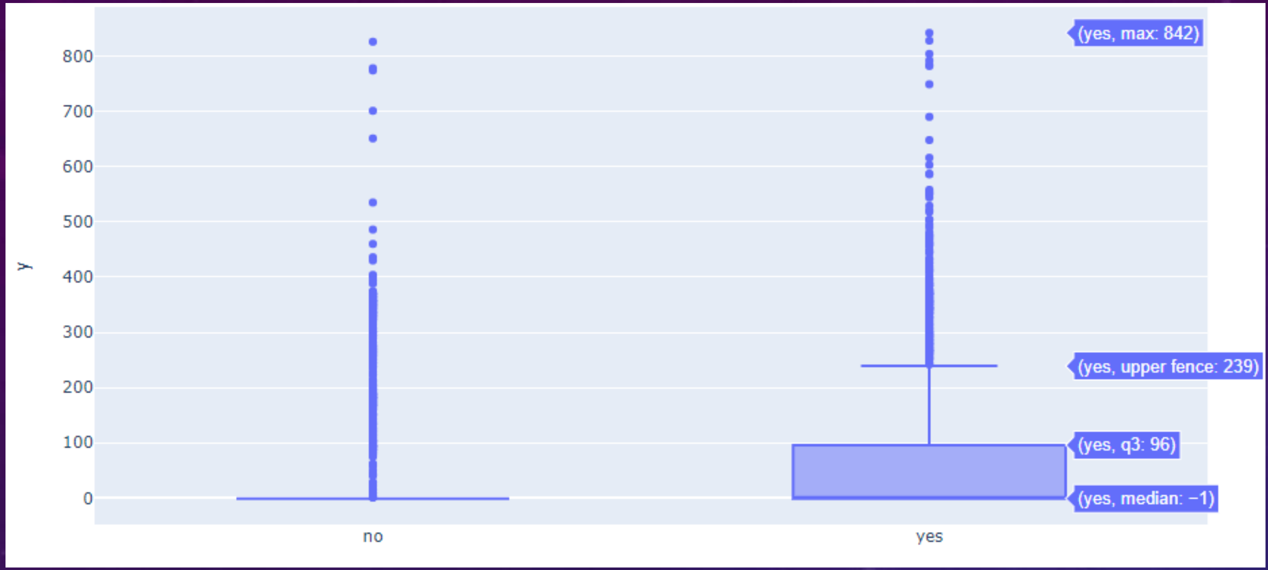
SIGNIFICANCE OF HOUSING ON THE TARGET VARIABLE

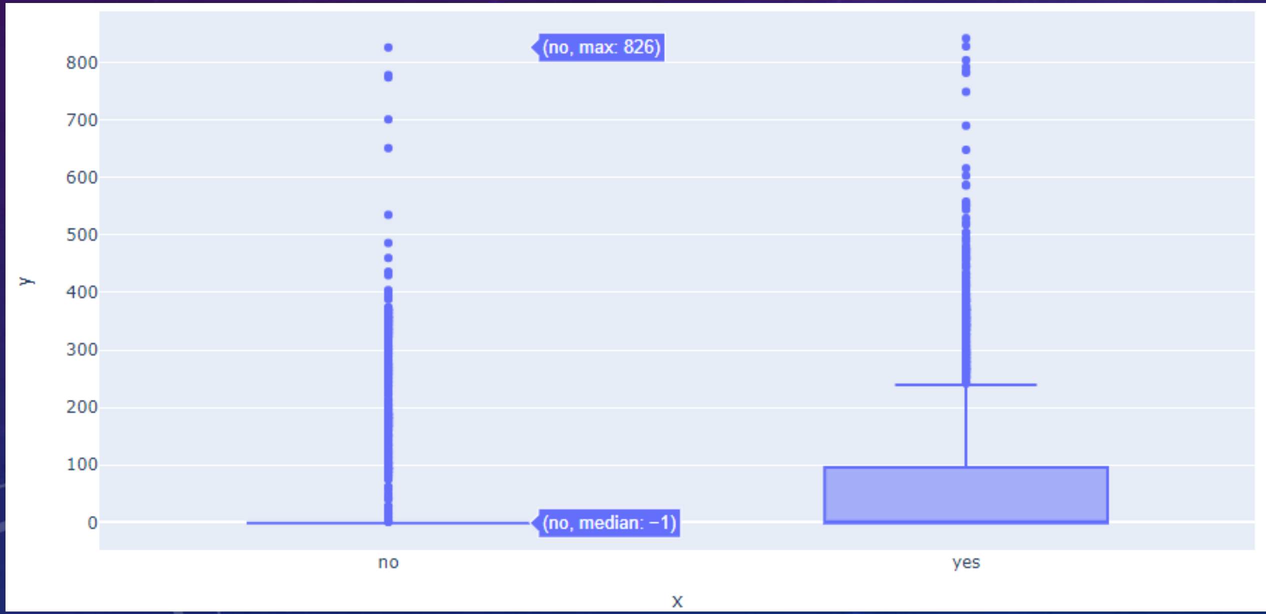




From the above plot we can see that, the housing(housing loan) of a customer can be useful for predicting the target variable Deposit since there is a difference in the medians i.e the medians do not overlap in the boxplot.

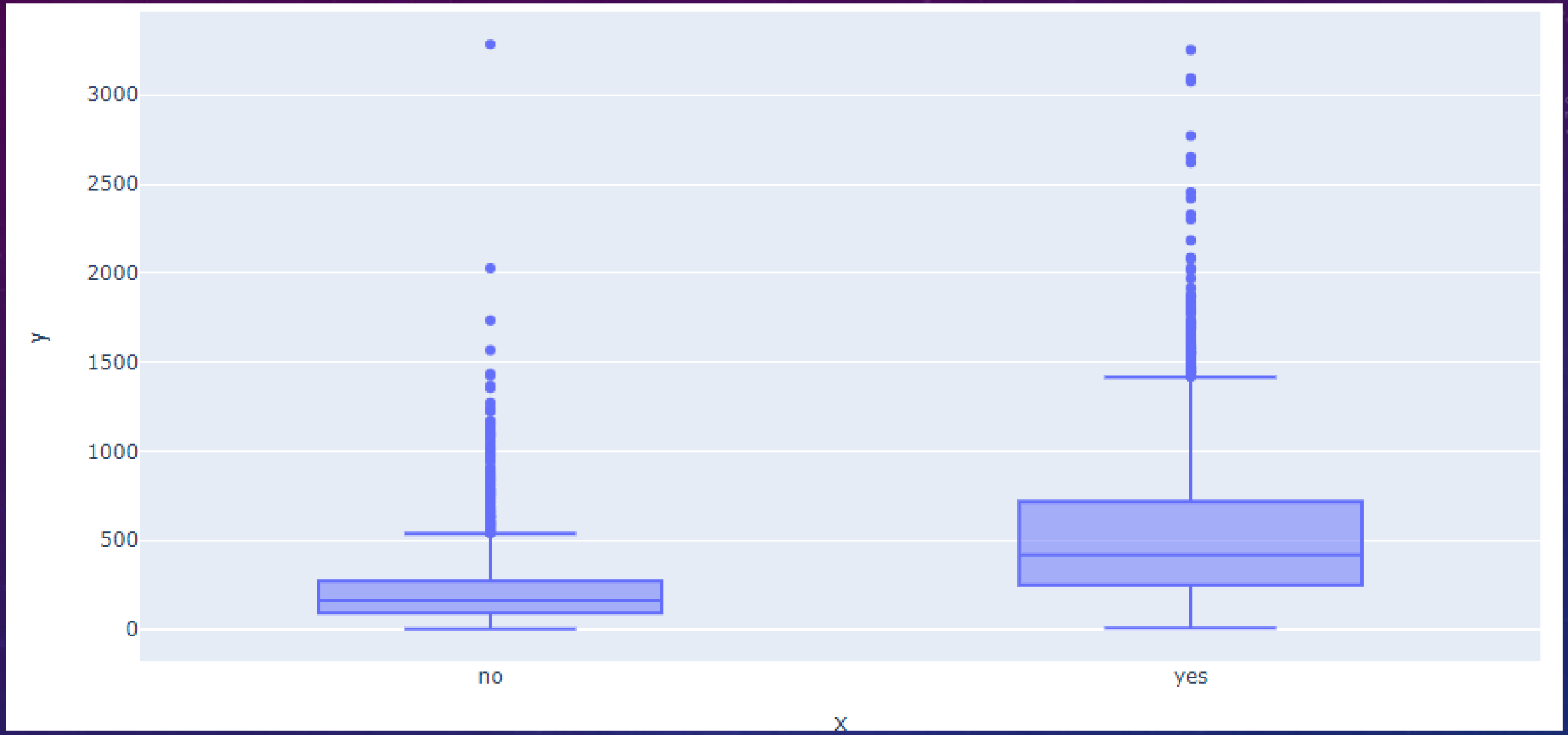
SIGNIFICANCE OF PDAYS ON THE TARGET VARIABLE





Pdays can be useful for predicting the target variable Deposit since there is a difference in the medians i.e the medians do not overlap in the boxplot.

SIGNIFICANCE OF DURATION ON THE TARGET VARIABLE



From the above plot we can see that, duaration can be useful for predicting the target variable Deposit since there is a difference in the medians i.e the medians do not overlap in the boxplot

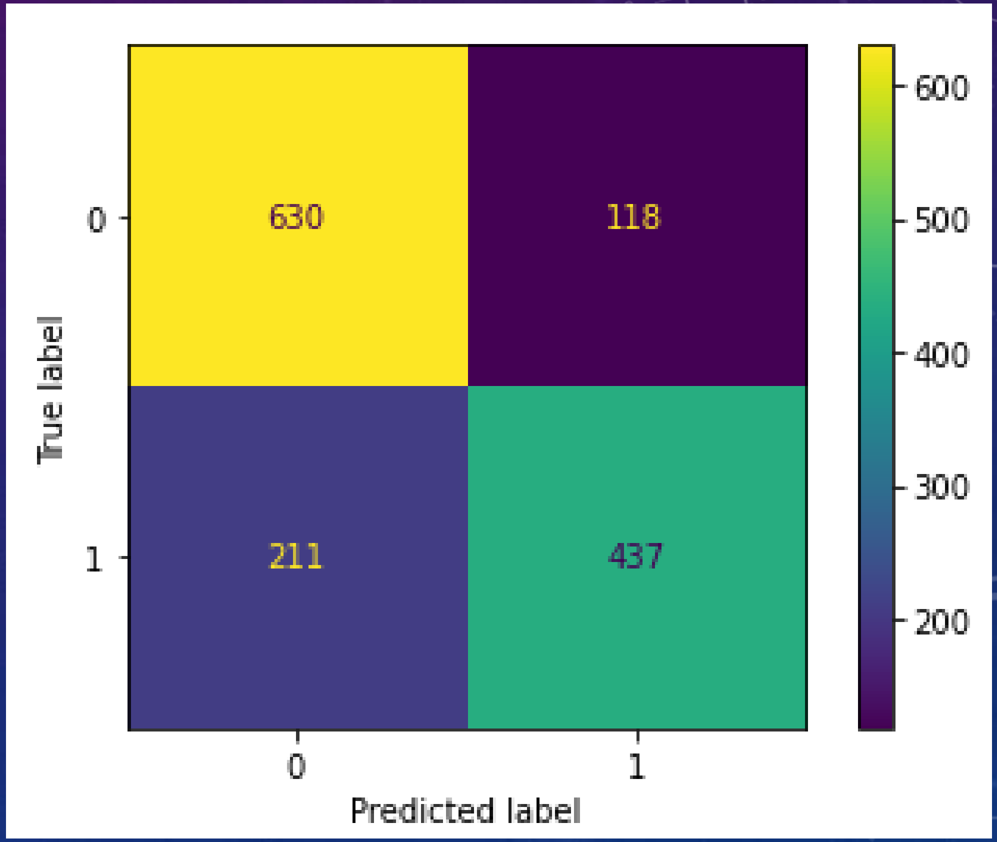
MODEL BUILDING- PRE REQUISITES

- The **sklearn** library is being imported to perform **standardization, splitting of data** and to find the **confusion matrix** and check the **accuracy score**
- We also import tree, RandomForestClassifier, LogisticRegression, KNNClassifier to build the models
- Even though there is a weak correlation, based on the boxplot, we are considering the attributes 'Housing', 'Duration', 'Campaign', 'Pdays' and 'Previous' for building the model.

LOGISTIC REGRESSION

Testing Accuracy : 0.7643266475644699

Misclassified samples: 329

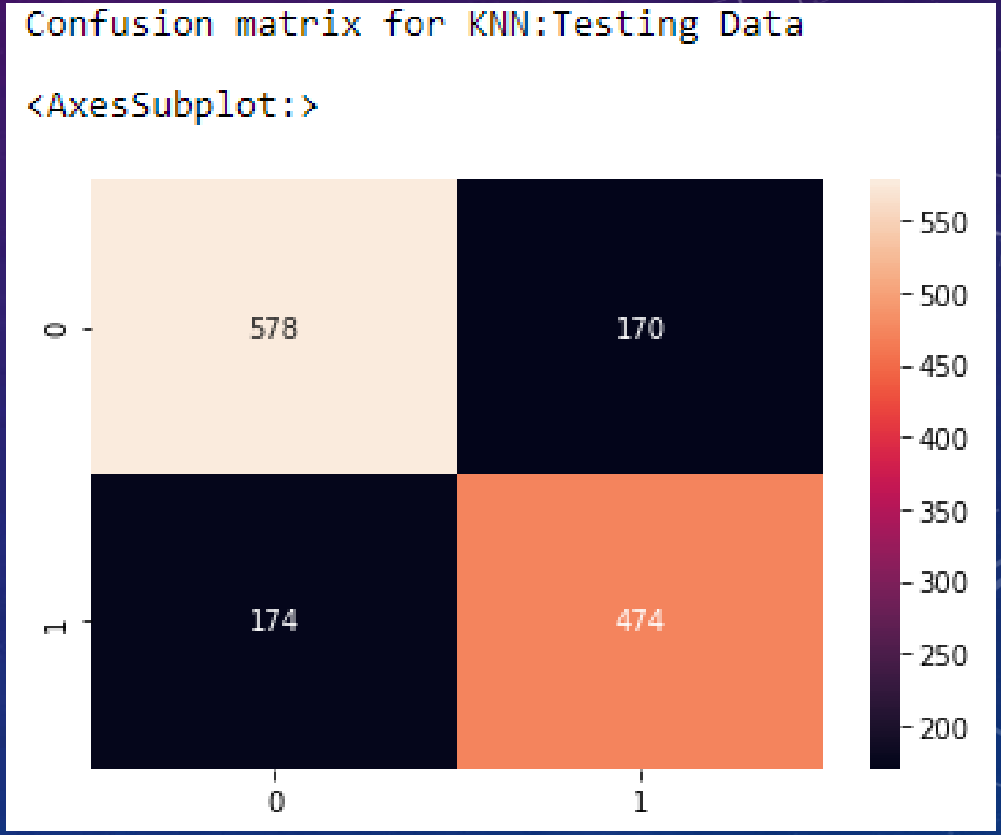


CONFUSION MATRIX

K NEAREST NEIGHBOURS

Testing Accuracy : 0.7535816618911175

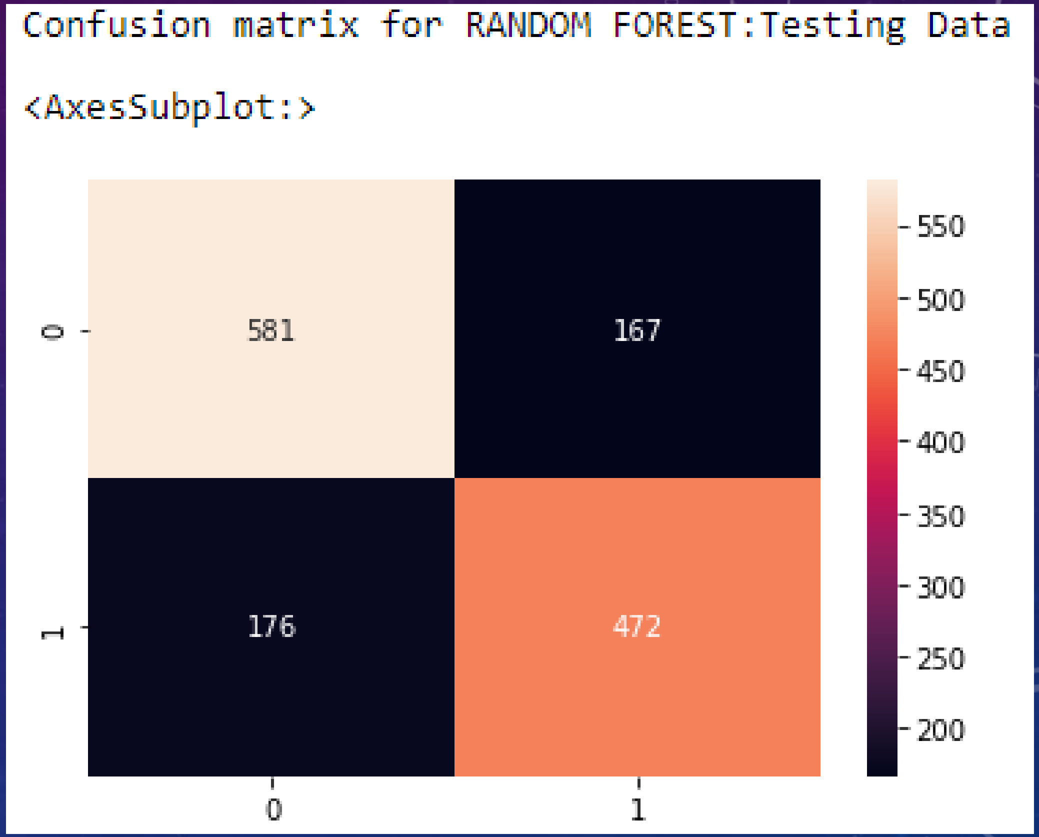
Misclassified samples: 344



CONFUSION MATRIX

RANDOM FOREST

Testing Accuracy : 0.75
Misclassified samples: 349



CONFUSION MATRIX

MODEL	ACCURACY SCORE	MISCLASSIFIED SAMPLES
Logistic Regression	0.7643266475644699	329
K Nearest Neighbours	0.7535816618911175	344
Random Forest	0.75	349

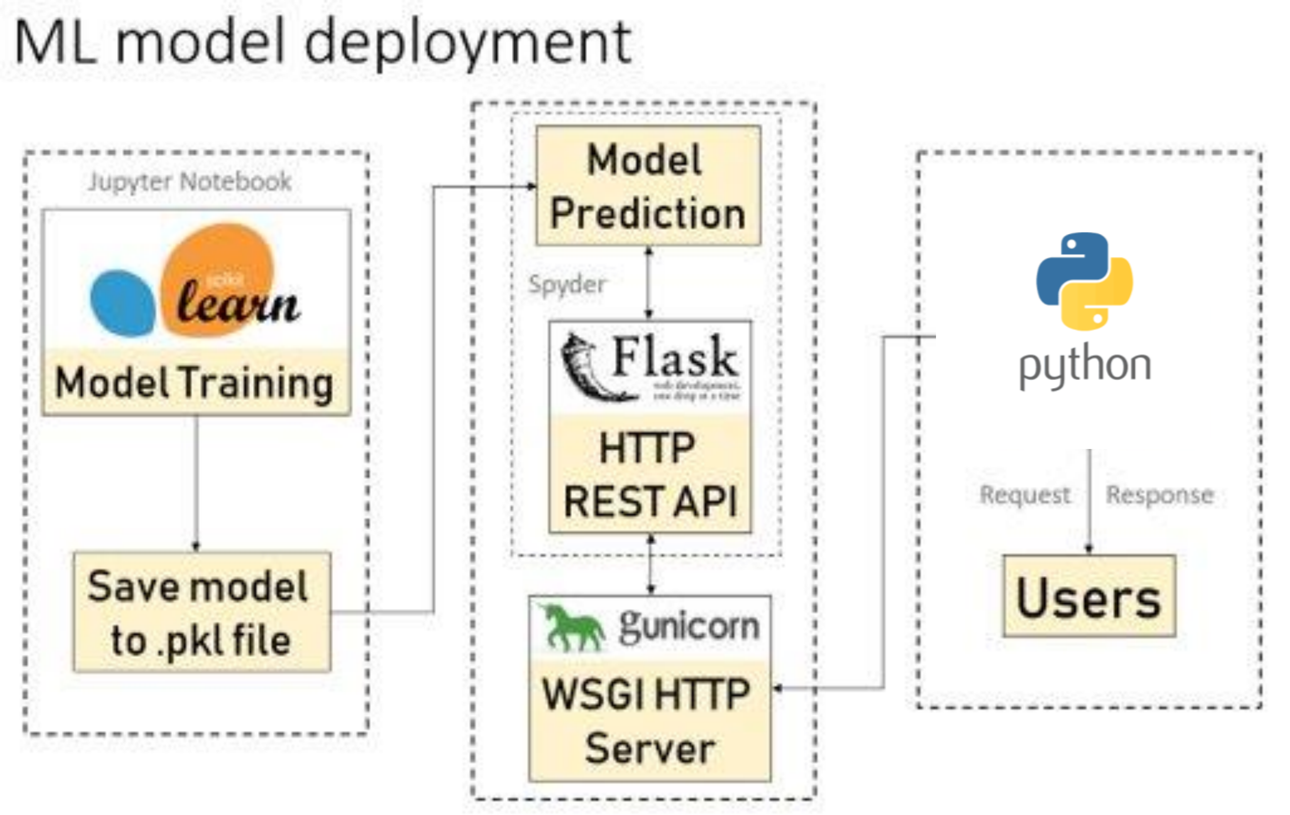
- Logistic Regression has the best accuracy score and also has the least number of misclassified samples.
- The Logistic Regression model will be deployed using HTML and Flask server

DEPLOYING THE MODEL USING HTML AND FLASK SERVER

Deploying a machine learning model

- Deployment is **the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data**.

- It is one of the last stages in the machine learning life cycle and can be one of the most complex.



What is a Flask Server?

- **Flask is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python easier.**

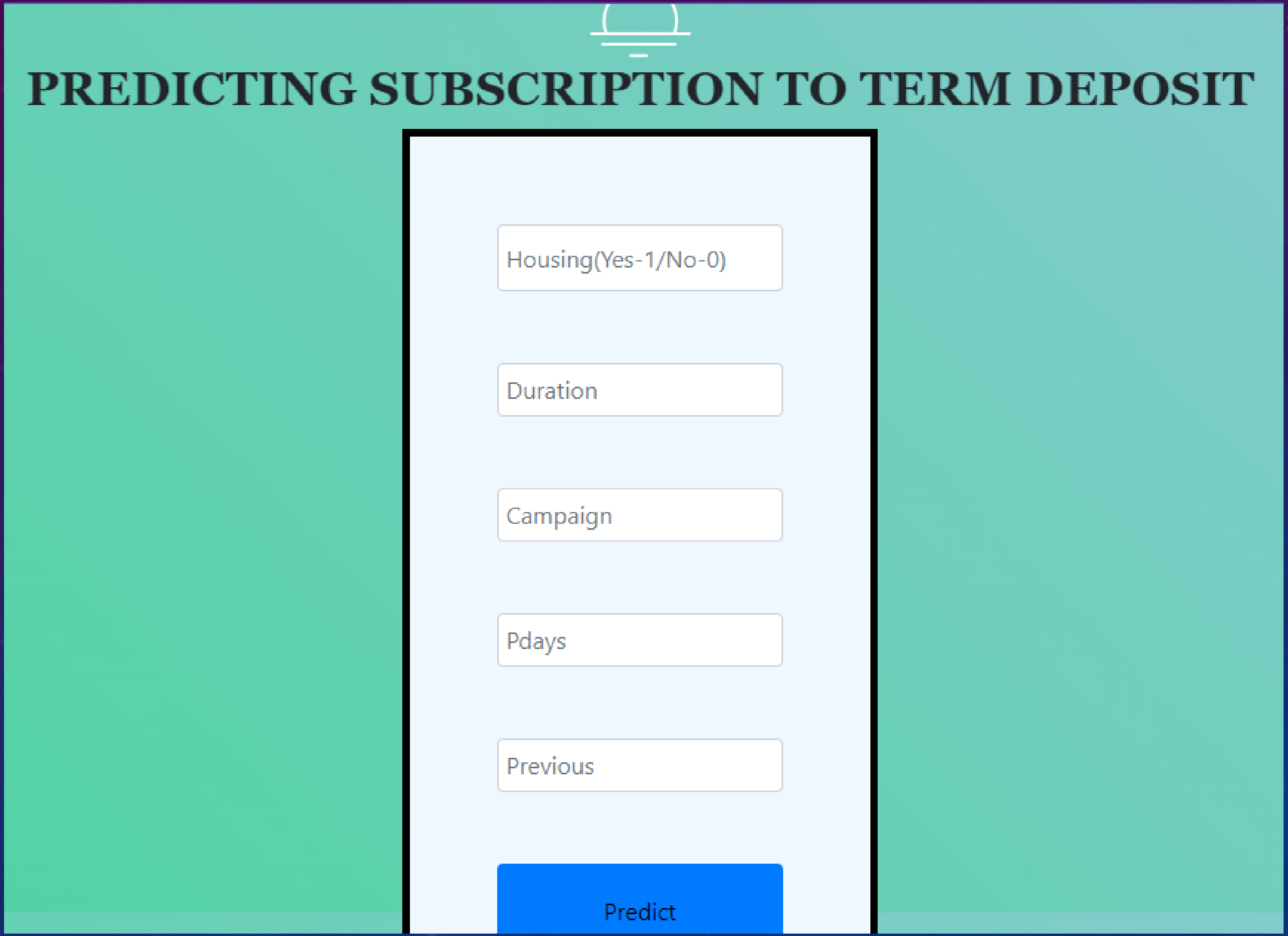
- It gives developers flexibility and is a more accessible framework for new developers since you can build a web application quickly using only a single Python file

- A pickle file of the Logistic Regression model is created.

- We use PyCharm to deploy the model using Flask server.

- HTML and CSS are used to create a form through which input data is given and prediction is being done

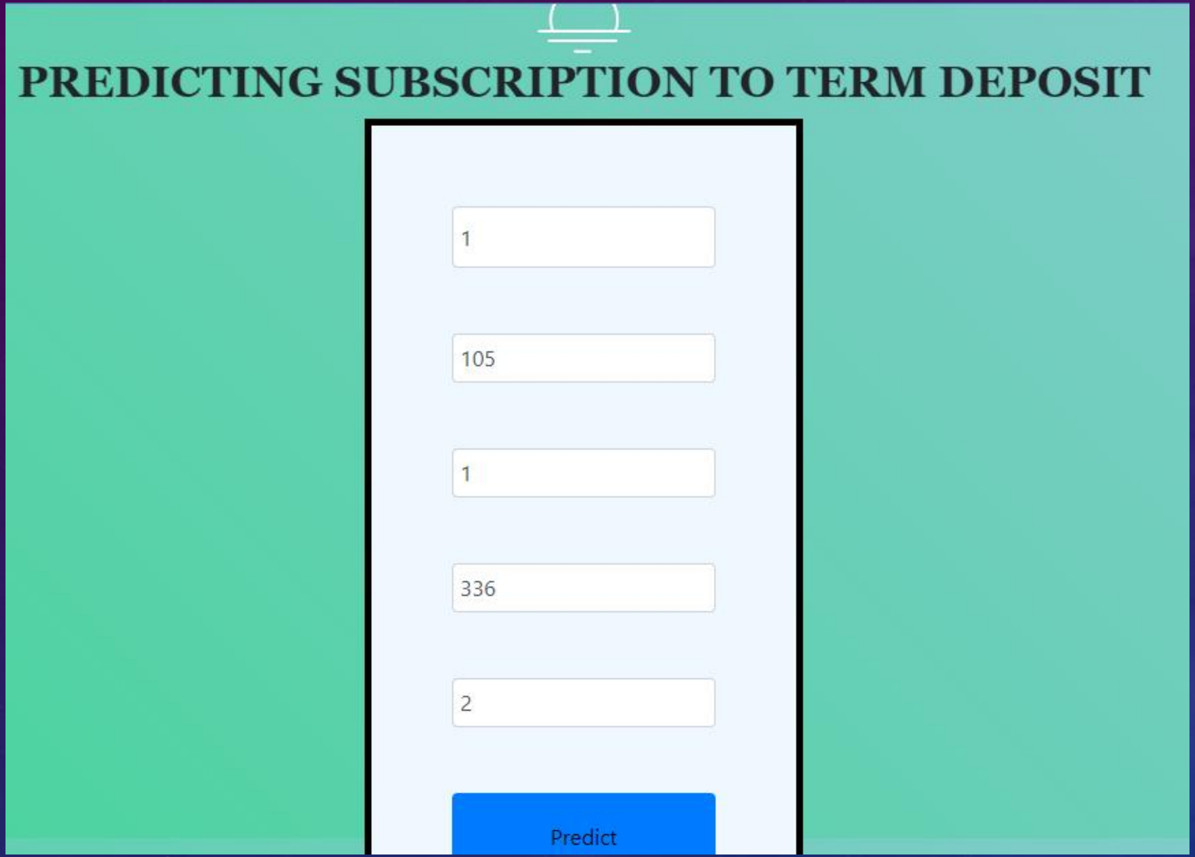
HTML PAGE

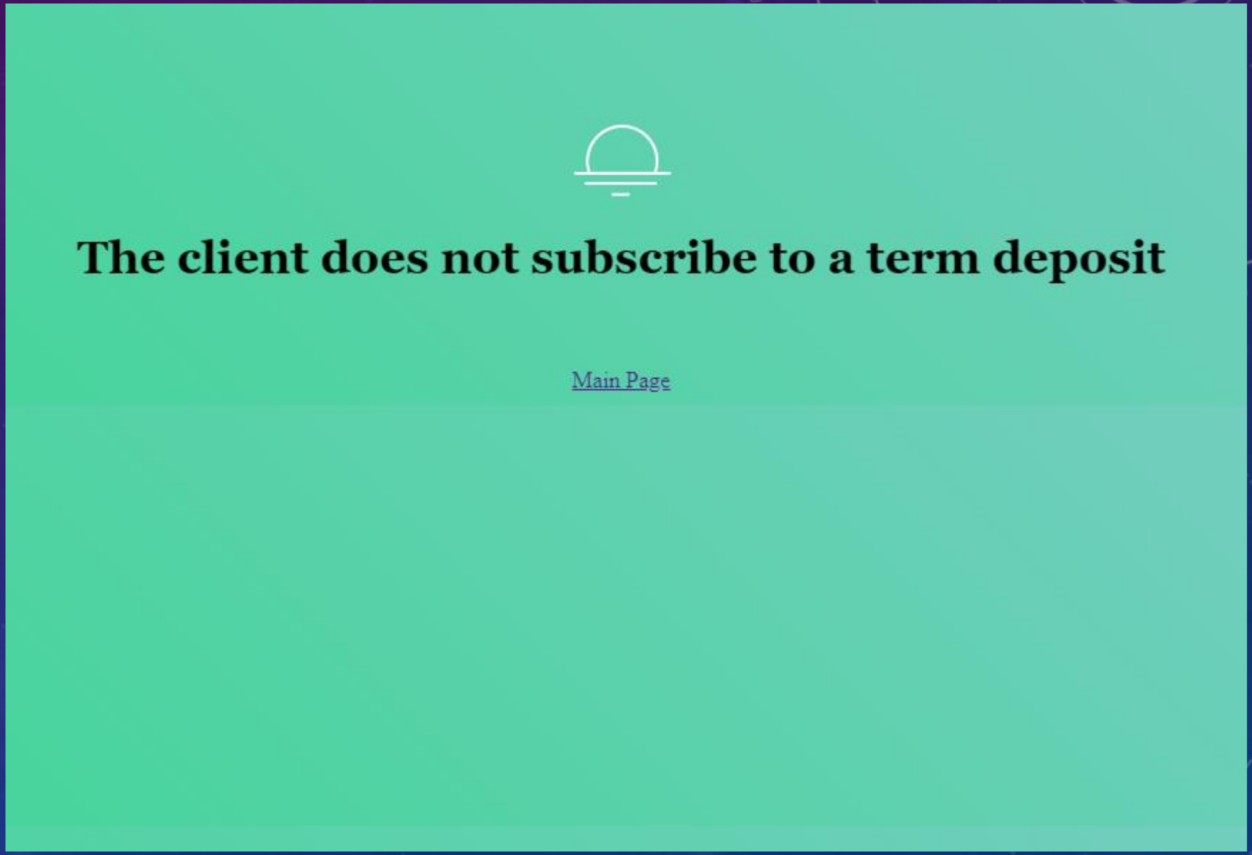

PREDICTING SUBSCRIPTION TO TERM DEPOSIT
Housing(Yes-1/No-0)
Duration
Campaign
Pdays
Previous
Predict

In the Jupyter notebook, while giving the values ([[1,105,1,336,2]])) for prediction
we get the predicted value as 0 i.e

Client does not subscribe to a term deposit.

While giving the input in the HTML page, we are getting the same prediction



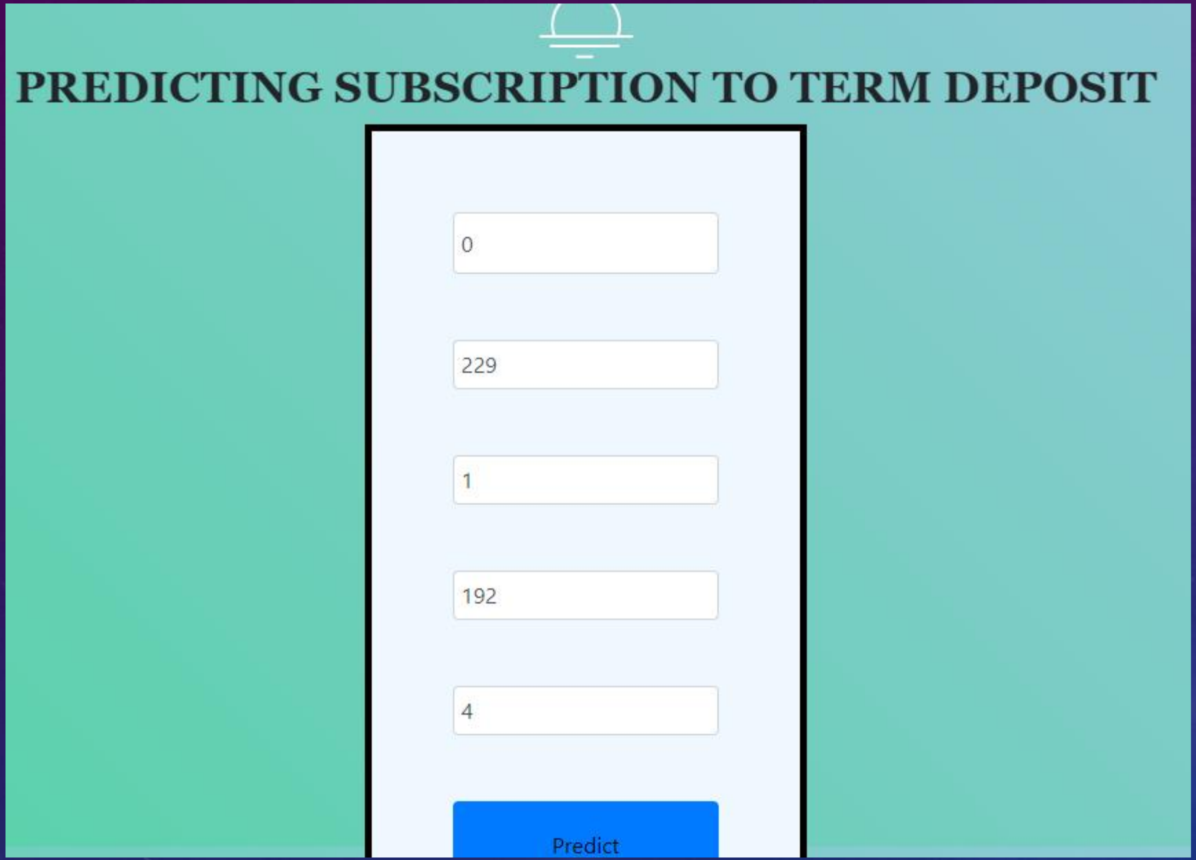


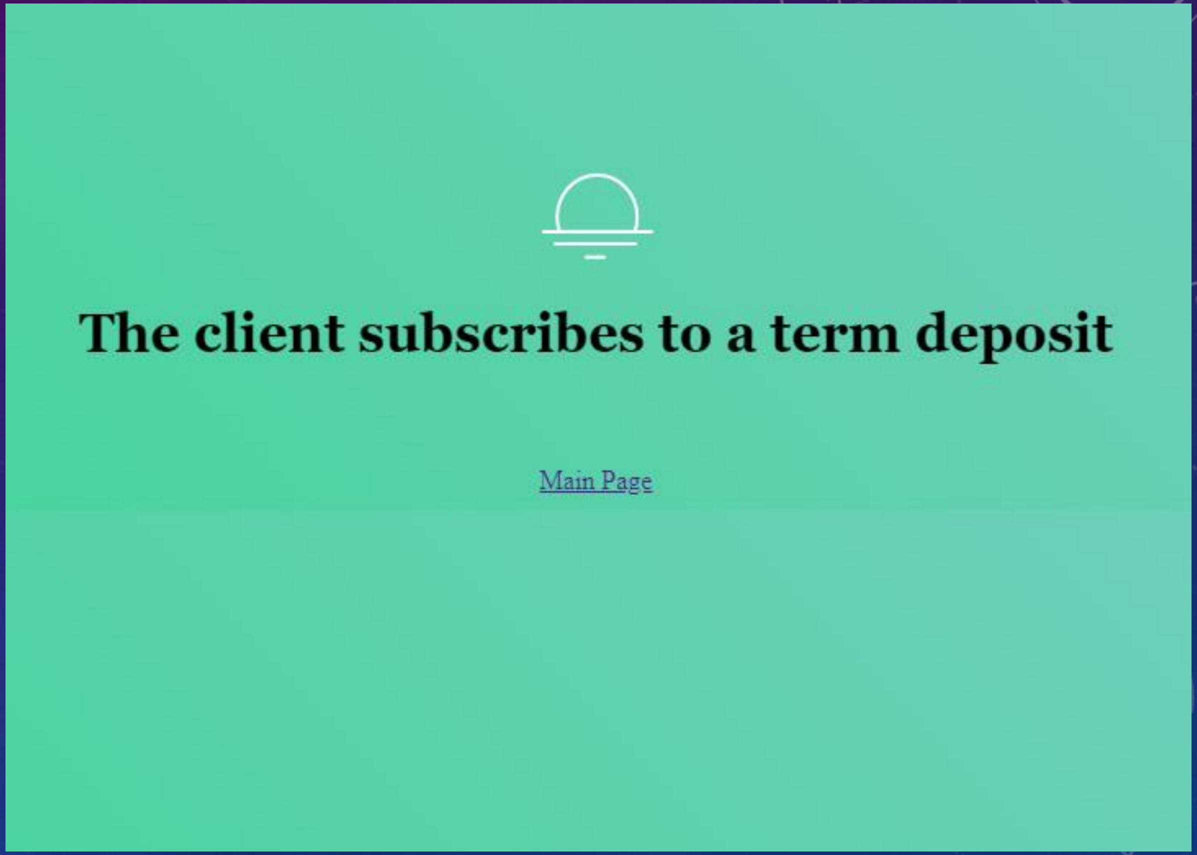
	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	41	services	married	unknown	no	88	yes	no	cellular	11	may	105	1	336	2	failure	no

In the Jupyter notebook, while giving the values ([[0,229,1,192,4]])) for prediction we get the predicted value as 1 i.e

Client subscribes to a term deposit.

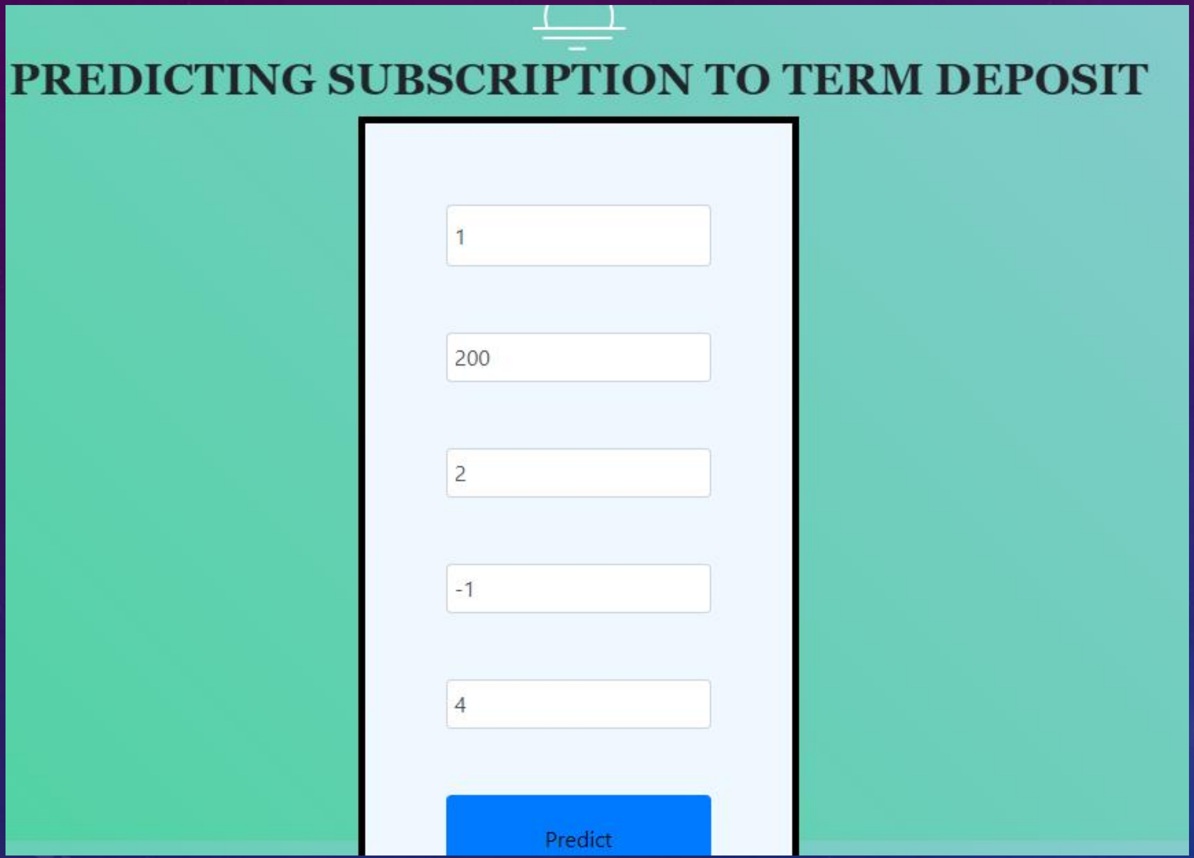
While giving the input in the HTML page, we are getting the same prediction

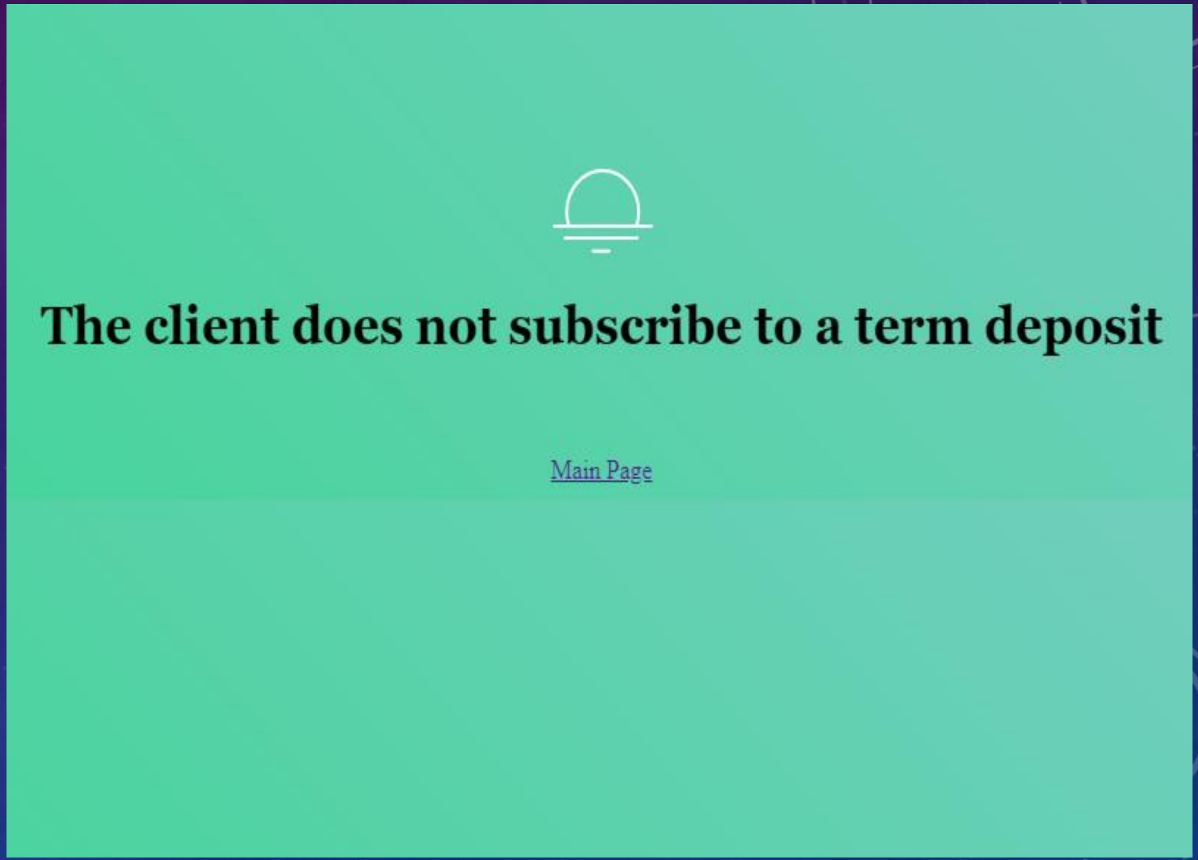




1	56	technician	married	secondary	no	1938	no	yes	cellular	26	feb	229	1	192	4	success	yes

New data is given as input and we get a prediction with 76% accuracy





THANK YOU