Name- Abhishek Deswal
Internship Program- Data Science with Machine Learning and Python
Batch- Jun 2022- Aug 2022
Certificate ID- TCRIB3R153
Date of submission- 09-08-2022

Technical Coding Research Innovation, Navi Mumbai,
Maharashtra, India-410206

(Product subscription on Bank Institution using ML and Python)

A Case-Study Submitted for the requirement of
Technical Coding Research Innovation

For the Internship Project work done during
DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM

by

Abhishek Deswal(TCRIB3R153)

Rutuja Doiphode
CO-FOUNDER &CEO
TCR innovation.

Abstract - The purpose of this paper is to investigate the status of Portuguese banking institution. The study has two objectives: one is to identify and measure the factors of clients perceived as important in deciding to patronize a Portuguese bank and other is to draw a client profile for Portuguese banks.

Index Terms –

- Problem Statement.
- Introduction to dataset.
- Exploratory data analysis.
- Training and testing of data.
- Model selection and building.
- Conclusion

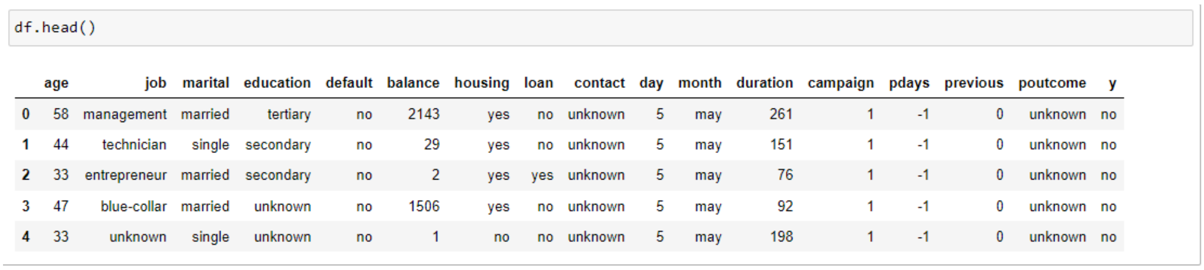
I. Problem Statement

The given 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 (Col -21).

II. Introduction to dataset.

We are using the given data set which is named as bank-full dataset, and the given dataset is in csv file format which is in unstructured format, the given data set contains 21 columns they are age, job, martial, duration, campaign, pdays, previous, outcome, and we need to predict the target variable. This dataset includes more than 15 features and more than 45211 rows, the dataset contains both numerical and categorical data. We are using NUMPY, PANDAS, MATPLOTLIB, SKLEARN libraries of python. the dataset is portrayed below:

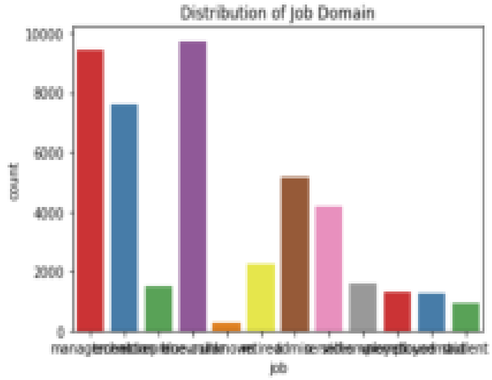


```
df.head()
```

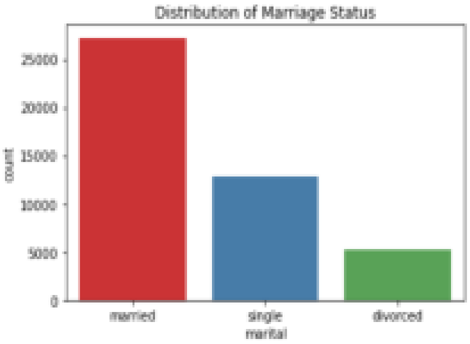
	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

III. Exploratory Data Analysis

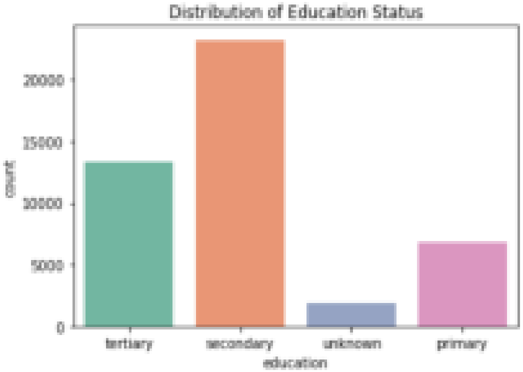
EDA is an approach to analyse the data using visual techniques. It is used to discover patterns or to check assumptions with the help of statistical summary and graphical representation. Using these techniques, we can find if there is any missing values or null values which are present in the dataset. By applying this approach, we can detect the values which can effect on our prediction results.



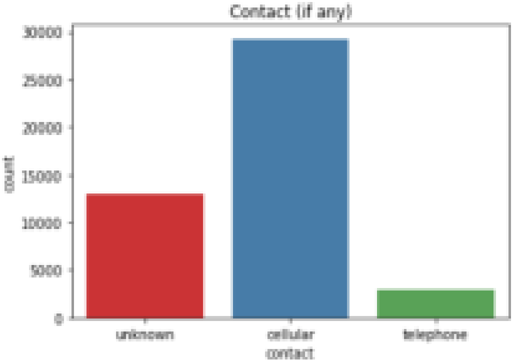
Distribution of Job Domain



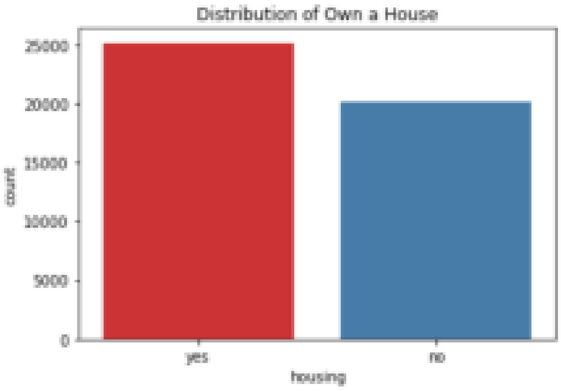
Distribution of Marriage Status



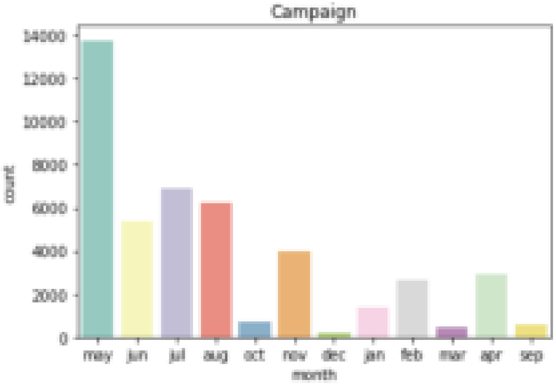
Distribution of Education Status



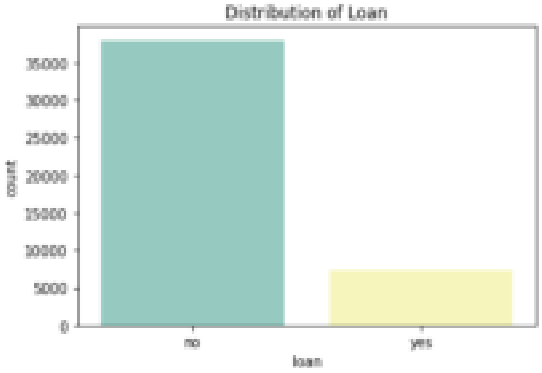
Contact (if any)



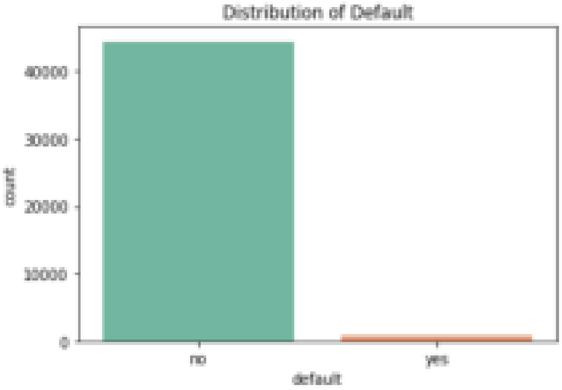
Distribution of Own a House



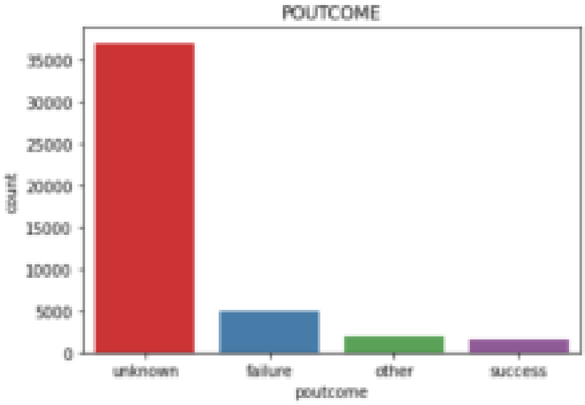
Campaign



Distribution of Loan



Distribution of Default



• **Default**: The default variable has no effect on the client subscribing to the term deposit. As we can see, without entry the client took the term deposit and the client with credit does not take the term deposit. So, we will skip this feature.

• **Marital**: About 60% of the clients approached were married.

• **Education**: Clients with university and secondary education were approached more than others and also have a higher success rate. (In terms of term deposit number)

• **Housing**: A housing loan does not have a big impact on the number of term deposits purchased.

• **Loan**: We approach 84% of clients who do not have a personal loan. Contact: About 64% of calls originate from the mobile network.

• **Month**: In May approximately 33% were contacted and in January, February we have no data or no one was contacted. The success rate was almost the same in June, July and August.

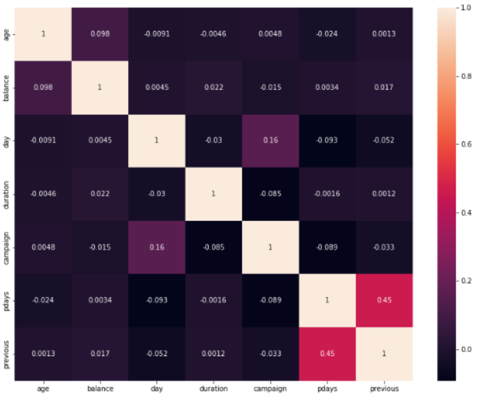
• **day of week**: We have collected values for 5 days. There is no significant difference in the number of clients approached and the number of registered persons.

• **poutcome**: If the client took a term deposit last time, there is a greater chance that the client will subscribe to it again.

IV. Correlation

Statistics from categorical variables (based on univariate analysis):

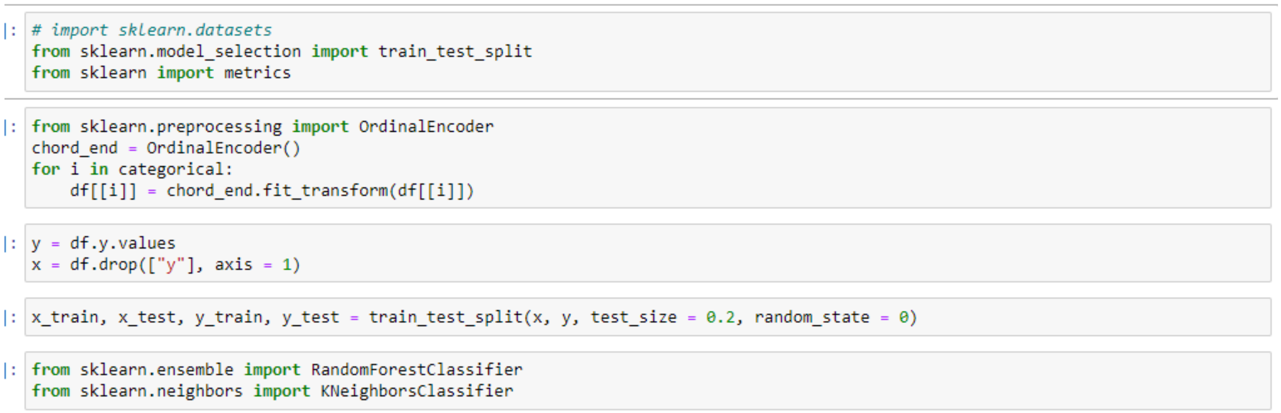
Correlation is an indication about the changes between two variables. We can plot correlation matrix to show which variable is having a high or low correlation in respect to another variable. The correlation matrix of the numerical data is obtained as follows:



V. TRAINING AND TESTING

The data is than spilt into test dataset and train dataset and then the train data is used to make different models and the test data is used to test these trained models and the one with best accuracy is selected. The ratio of the Train to Test dataset should be of approximately 80:20 or 70:30.

Models Used for the Evaluation:



```
# import sklearn.datasets
from sklearn.model_selection import train_test_split
from sklearn import metrics
```

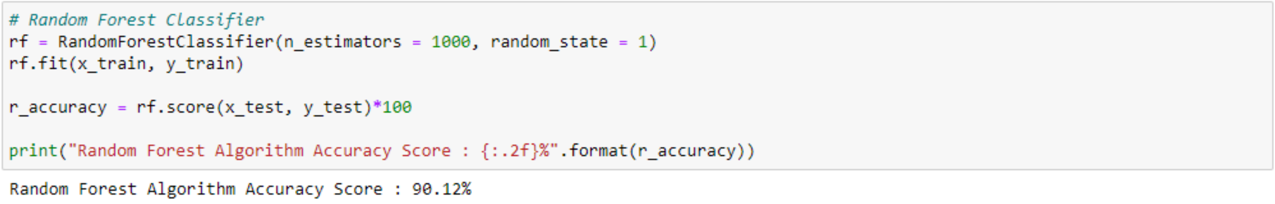
```
from sklearn.preprocessing import OrdinalEncoder
chord_end = OrdinalEncoder()
for i in categorical:
    df[[i]] = chord_end.fit_transform(df[[i]])
```

```
y = df.y.values
x = df.drop(["y"], axis = 1)
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 0)
```

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
```

Random Forest Classifier:



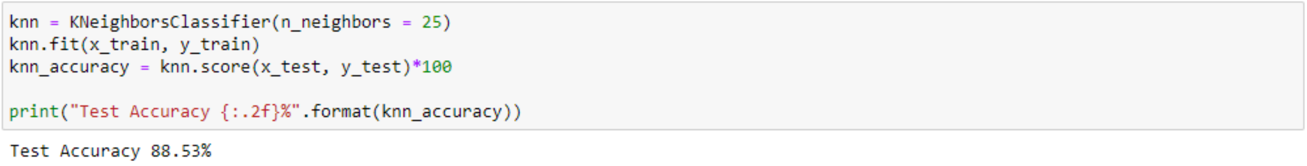
```
# Random Forest Classifier
rf = RandomForestClassifier(n_estimators = 1000, random_state = 1)
rf.fit(x_train, y_train)

r_accuracy = rf.score(x_test, y_test)*100

print("Random Forest Algorithm Accuracy Score : {:.2f}%".format(r_accuracy))
```

```
Random Forest Algorithm Accuracy Score : 90.12%
```

K – Nearest Neighbour:



```
knn = KNeighborsClassifier(n_neighbors = 25)
knn.fit(x_train, y_train)
knn_accuracy = knn.score(x_test, y_test)*100

print("Test Accuracy {:.2f}%".format(knn_accuracy))
```

```
Test Accuracy 88.53%
```

VI. CONCLUSION

After applying Random Forest classification algorithm, machine learning model is able to predict the results with 90.12 % accuracy score.