Name- Shashank Shivaji Patil
Internship Program- Data Science with Machine Learning and Python
Batch- Apr 2022-Jun 2022
Certificate ID- TCRIB3R84
Date of submission- 19-07-2022

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

Shashank Shivaji Patil (TCRIB3R84)

Name- Shashank Shivaji Patil
Internship Program- Data Science with Machine Learning and Python
Batch- Apr 2022-Jun 2022
Certificate ID- TCRIB3R84
Date of submission- 19-07-2022

Rutuja Doiphode
CO-FOUNDER &CEO
TCR innovation.

BANK DATASET

***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 perceive as important in deciding patronize a Portuguese bank and other Is to draw a client profile for Portuguese bank.**

**Index Terms –**

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

## 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, poutcome, 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:

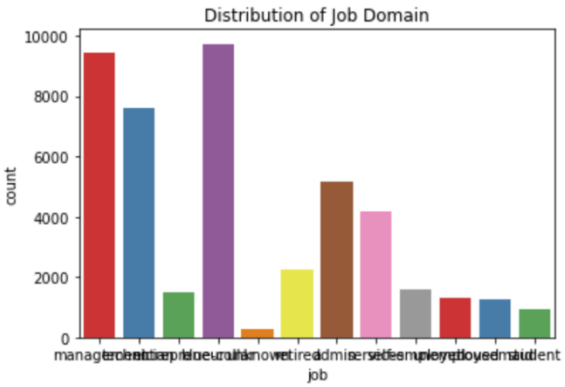
dataset

| | 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 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 45206 | 51 | technician | married | tertiary | no | 825 | no | no | cellular | 17 | nov | 977 | 3 | -1 | 0 | unknown | yes |
| 45207 | 71 | retired | divorced | primary | no | 1729 | no | no | cellular | 17 | nov | 456 | 2 | -1 | 0 | unknown | yes |
| 45208 | 72 | retired | married | secondary | no | 5715 | no | no | cellular | 17 | nov | 1127 | 5 | 184 | 3 | success | yes |
| 45209 | 57 | blue-collar | married | secondary | no | 668 | no | no | telephone | 17 | nov | 508 | 4 | -1 | 0 | unknown | no |
| 45210 | 37 | entrepreneur | married | secondary | no | 2971 | no | no | cellular | 17 | nov | 361 | 2 | 188 | 11 | other | no |

45211 rows × 17 columns

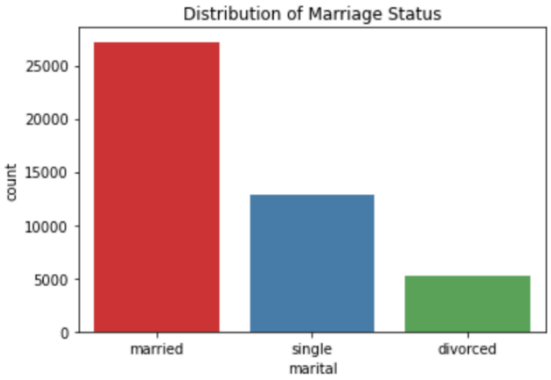
## III. EXPLORATORY DATA ANALYSIS.

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 affect on our prediction results.

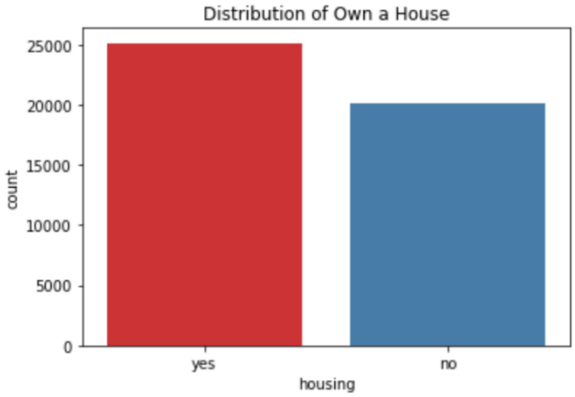



JOB Description

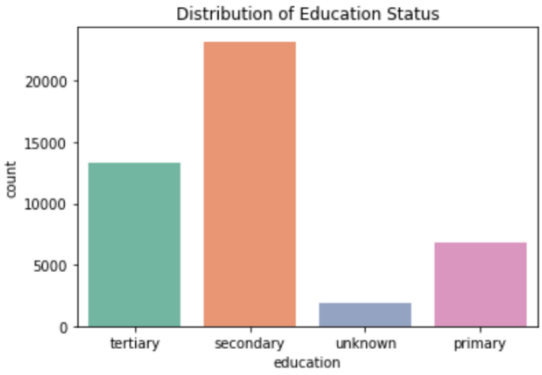
Name- Shashank Shivaji Patil
Internship Program- Data Science with Machine Learning and Python
Batch- Apr 2022-Jun 2022
Certificate ID- TCRIB3R84
Date of submission- 19-07-2022



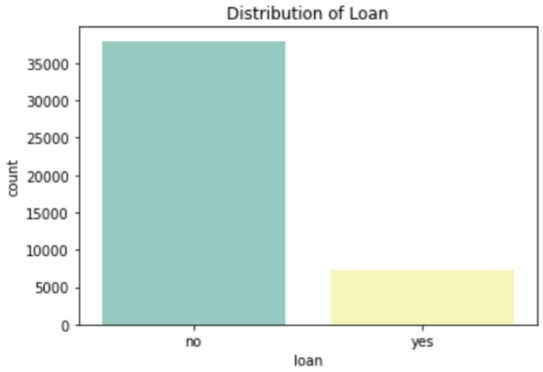

Marital Status



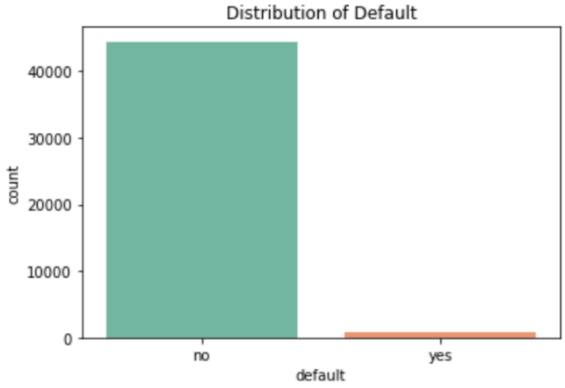

HOUSING



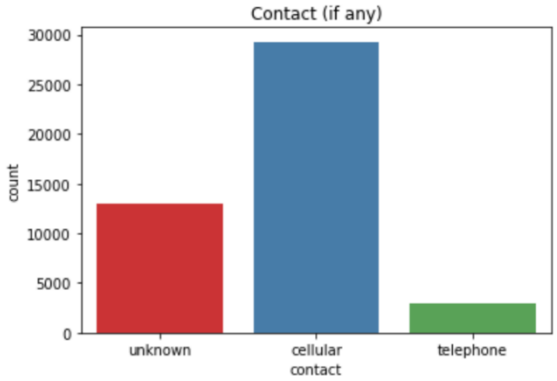

EDUCATION




LOAN

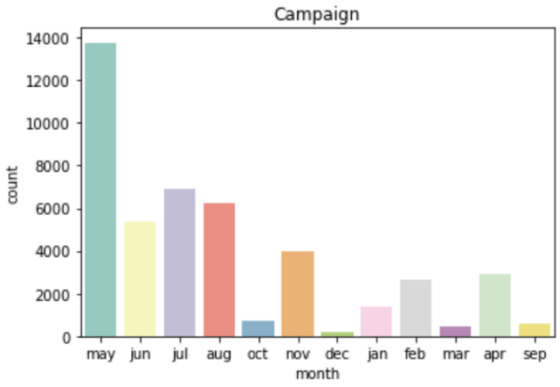



DEFAULT

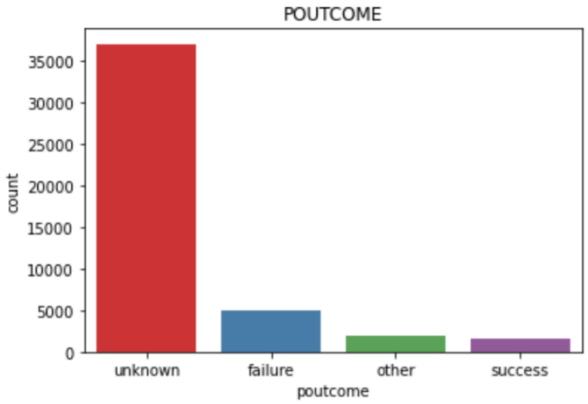



CONTACT

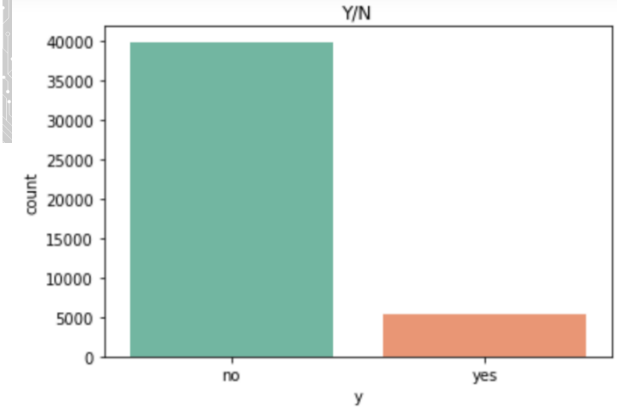
Name- Shashank Shivaji Patil
Internship Program- Data Science with Machine Learning and Python
Batch- Apr 2022-Jun 2022
Certificate ID- TCRIB3R84
Date of submission- 19-07-2022




MONTH of CAMPAIGN




POUTCOME




Y/N

From the above output we are finding the categorical feature distribution, Outcome:

- The clients with job type as management records are high in the given dataset.
- Clients who married are high in records in the given dataset.
- Clients whose education background is secondary are in high numbers.
- Default feature seems to be doesn't play important role as it as value of no at high ratio to value yes which we can drop.

**Statistics from categorical variables (based on univariate analysis)**:

- **Job**: The highest number (around 25%) of applications comes from the position of administrator.
- **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.

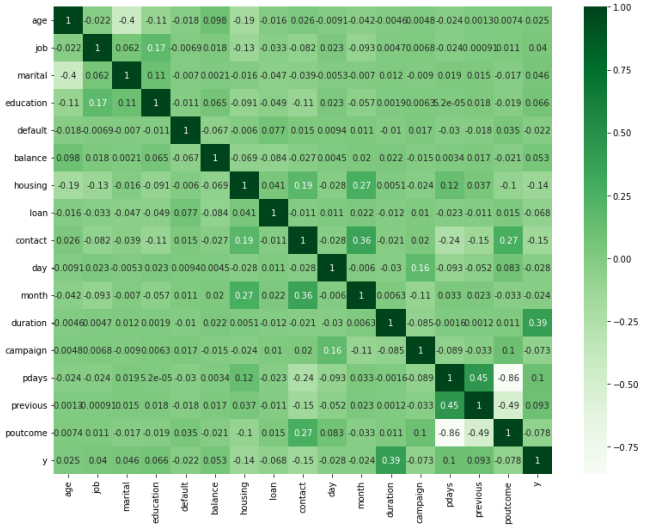
Name- Shashank Shivaji Patil
Internship Program- Data Science with Machine Learning and Python
Batch- Apr 2022-Jun 2022
Certificate ID- TCRIB3R84
Date of submission- 19-07-2022

## IV. CORRELATION

Correlation is an indication about the changes between two variables. In our previous chapters, we have discussed Pearson's Correlation coefficients and the importance of Correlation too. 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:


HEATMAP (CORRELATION MATRIX)

The categorical data is then converted to the numerical data and the new correlation matrix is as follows:




HEATMAP (CATEGORICAL DATA; CORRELATION MATRIX)

## 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.

```
# Training And Testing the dataset

y = df_bank.y.values
x_data = df_bank.drop(["y"], axis = 1)

x = (x_data - np.min(x_data)) / (np.max(x_data) - np.min(x_data)).values

#train test split

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

x_trian = x_train
y_train = y_train
x_test = x_test
y_test = y_test
```

Training and Testing

➔ Models Used for the Evaluation:

1. Random Forest Classifier:

Name- Shashank Shivaji Patil
Internship Program- Data Science with Machine Learning and Python
Batch- Apr 2022-Jun 2022
Certificate ID- TCRIB3R84
Date of submission- 19-07-2022

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
accuracies = {}
# 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
accuracies["Random Forest"] = r_accuracy
print("Random Forest Algorithm Accuracy Score : {:.2f}%".format(r_accura
```

```
Random Forest Algorithm Accuracy Score : 90.09%
```

Random Forest Classifier

2. Logistic Regression:

```
Lr = LogisticRegression()
Lr.fit(x_train, y_train)
accuracy = Lr.score(x_test, y_test)*100

accuracies["Logistic Regression"] = accuracy
print("Test Accuracy {:.2f}%".format(accuracy
```

```
Test Accuracy 88.80%
```

3. K – Nearest Neighbour:

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

accuracies["K-Nearest Neighbour (KNN) Classification"] = accuracy
print("Test Accuracy {:.2f}%".format(accuracy))
```

```
Test Accuracy 88.69%
```

## VI. CONCLUSION

After applying Random Forest classification algorithm, machine learning model is able to predict the results with 90.09% accuracy. Using K means algorithm and Logistic regression, we can predict the same results, but I found that this Random Forest Algorithm provides better results than other classification algorithms.

## VII. REFERENCES

1. S. B. Kotsiantis, P. E. Pintelas and I. D. Zaharakis. “Machine Learning: a review of classification and combination techniques”, Artificial Intelligence Rev, Vol, 26,pp. 159-109,2006

2. K. Wisaeng. “A Comparison of Different classification Techniques for Bank direct Marketing”. International Journal of Soft Computing and Engineering (IJSCE), vol. 3, no. 4, pp. 116-119, 2013.

3. S. B. Kotsiantis, P. E. Pintelas and I. D. Zaharakis. “Machine Learning: a review of classification and combining techniques”, Artificial Intelligence Rev, vol, 26, pp. 159-190, 2006