An Internship Report On

PREDICTING TERM DEPOSIT SUBSCRIPTIONS

Submitted in partial fulfillment of requirements.

For the Internship (IT-451)

Submitted By

S. NARENDRA KUMAR (Y20IT112)

NOVEMBER-2023

R.V.R. & J.C. COLLEGE OF ENGINEERING (Autonomous)
NAAC A+ Grade, NBA Accredited
(Approved by A.I.C.T.E)

(AFFILIATED TO ACHARYA NAGARJUNA UNIVERSITY)

DEPARTMENT OF INFORMATION TECHNOLOGY
R.V.R. & J.C. COLLEGE OF ENGINEERING (Autonomous)
GUNTUR-522019

BONAFIDE CERTIFICATE

This is to certify that this internship report "**DATA SCIENCE**" is the bonafide work of **"SIVANGULA NARENDRA KUMAR (Y20IT112)"** who have carried out the work under my supervision, and submitted in partial fulfillment for the award of **INTERNSHIP (IT-451)** during the year 2023-2024.

Smt. B. Manasa
Assistant Professor, IT

Dr. A. SriKrishna
Prof.& HOD, Department of IT

INTERNSHIP CERTIFICATION

INTERNSHALA TRAININGS

Certificate of Training

Narendra Kumar Sivangula

from R.V.R. & J.C.College Of Engineering has successfully completed a 6-week online training on **Data Science**.
The training consisted of Introduction to Data Science, Python for Data Science, Understanding the Statistics for Data Science, Predictive Modeling and Basics of Machine Learning, and The Final Project modules.
In the final assessment, Narendra Kumar scored 82% marks.
We wish Narendra Kumar all the best for future endeavours.

Sarvesh Agarwal
FOUNDER & CEO, INTERNSHALA

Date of certification: 2023-06-04 Certificate no. : 8vkka06di7d

For certificate authentication, please visit https://trainings.internshala.com/verify_certificate

ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without a proper suggestions, guidance and environment. The combination of these three factors acts like backbone to our internship "**DATA SCIENCE**".

I would like to express our gratitude to the Management of **R.V.R&J.C. COLLEGE OF ENGINEERING** for providing us with a pleasant environment and excellent lab facility.

I regard my sincere thanks to our Principal, **Dr. Kolla Srinivas,** for providingsupport and stimulating environment.

I am greatly indebted to **Dr. A. SriKrishna**, Professor, and Head of the Department Information Technology, for her valuable suggestions during the internship.

I would like to express our special thanks to our guide **Smt. B. Manasa**, Assistant Professor who helped us in doing the internship successfully.

SIVANGULA NARENDRA KUMAR
(Y20IT112)

TABLE OF CONTENTS

ABSTRACT

TERM DEPOSIT PREDICTION

In recent years, more and more data are being collected from a variety of sources for scientific researches. At the same time, data mining (DM) and machine learning (ML) are being utilized to analyze special features from the data. Meanwhile in business, this combination is able to produce comprehensive overviews to support human decision-making by showing profitable recommendations. In marketing, it is essential for organizations to know when they can provide their services by analyzing customer's data to make crucial strategies for their businesses, especially in banks. In this paper, with the problem of bank marketing, we will take a look at what types of experimental data are typically used, do a preliminary analysis on them, and generate a Term Deposit Subscription prediction model by using machine learning frameworks. Using a database with thousands of data points gathered in a marketing campaign, the accuracy rates of detection and classification are about 71% and 86% respectively. Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. There are certain variables that we need to take into consideration when making a marketing campaign. A Term deposit is a deposit that a bank or a financial institution offers with a fixed rate (often better than just opening a deposit account) in which your money will be returned back at a specific maturity time.

LIST OF FIGURES

LIST OF ABBRIEVATIONS

ABBRIEVATIONS	EXPANSION
IDE	INTEGRATED DEVELOPMENT ENVIRONMENT
UI	USER INTERFACE
KNN	K NEAREST NEIGHBOUR
SVM	SUPPORT VECTOR MACHINE

CHAPTER 1

INTRODUCTION

1.1 GENERAL

Marketing to potential clients has always been a crucial challenge in attaining success for banking institutions. It's not a surprise that banks usually deploy mediums such as social media, customer service, digital media and strategic partnerships to reach out to customers. The investment and portfolio department of the Bankwould want to be able to identify their customers who potentially would subscribe to their term deposits. As there has been heightened interest of marketing managers to carefully tune their directed campaigns to the rigorous selection of contacts, the task at hand is to find a model that can predict which future clients who would subscribe to the Bank's term deposit. In this post, am going to talk about data exploration, data cleaning, feature extraction, and dealing with class imbalance and developing a robust machine learning algorithm for predicting which potential customers would subscribe to the term deposit using different Machine learning library.

1.3OUTLINE OF THE PROJECT

This project aims to predict if a customer subscribes to a term deposit or not, when contacted by a marketing agent, by understanding the different features and performing predictive analytics.

CHAPTER 2

AIM AND SCOPE

2.1 AIM:

The main objective of this project is to build a model that predicts the customers that would or would not subscribe to bank term deposits using Machine Learning Techniques.

2.2 PROBLEM STATEMENT:

To Predict if a customer subscribes to a term deposit or not, when contacted by a marketing agent, by understanding the different features and performing predictive analytics

2.3 SCOPE:

This project will investigate how machine learning techniques can be used to predict the customer's subscriptions to as term deposit as to whether they will subscribe or not . This aims to develop a model in which there will be a dataset storing allthe dataset consists of several predictor variables and one target variable, Outcome. Predictor variables includes the age, job, marital status, and so on. The main purpose of the prediction is to reduce uncertainty associated to investment decision making and to avoid discrepancy in the deposition of money.The machine learning algorithm that was considered in this project includes Logistic Regression and several other methods which will predict data in an accurate manner

CHAPTER 3

SYSTEM ANALYSIS

3.1 GENERAL:

3.1.1 MACHINE LEARNING:

Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own.

3.1.2 PYTHON:

Python is an interpreted, high-level and general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

3.1.3 IDE (INTEGRATED DEVELOPMENT ENVIRONMENT)

IDE stands for Integrated Development Environment. It’s a coding tool which allows you to write, test, and debug your code in an easier way, as they typically offer code completion or code insight by highlighting, resource management, debugging tools. And even though the IDE is a strictly defined concept, it’s starting to be redefined as other tools such as notebooks start gaining more and more features that traditionally belong to IDEs. The IDE used in this project isColab.

3.1.4 USER INTERFACE

The user interface (UI) is the point of human-computer interaction and communication in a device. This can include display screens, keyboards, a mouse

and the appearance of a desktop. It is also the way through which a user interacts with an application or a website.

3.2 SOFTWARES USED

3.2.1 CODE EDITORS USED

COLAB

Colaboratory, or “Colab” for short, is a product from Google Research. Colab is a free Jupyter notebook environment that runs entirely in the cloud.It allows you to write and execute Python in your browser, with Zero configuration required,It provides free access to GPUs and it is also useful for easily sharing the codes. Colaboratory, or Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

3.2.2 PROGRAMMING LANGUAGES USED

PYTHON

Python is a general purpose, dynamic, high-level, and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is simple and easy to learn and provides lots of high-level data structures.

Python is easy to learn yet powerful and versatile scripting language, which makes it attractive for Application Development.

3.2.3 LIBRARIES AND FRAMEWORKS USED

PANDAS

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

MATPLOTLIB

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. It is better to represent the data through the graph where we can analyse the data more efficiently and make the specific decision according to data analysis.

NUMPY

NumPy stands for numeric python which is a python package for the computation and processing of the multidimensional and single dimensional array elements.

SEABORN

Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with dataframes and the Pandas library. The graphs created can also be customized easily.

SKLEARN

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

3.3 ALGORITHM

LOGISTIC REGRESSION ALGORITHM:

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

LINEAR DISCRIMINANT ANALYSIS ALGORITHM:

Linear Discriminant Analysis or LDA is a dimensionality reduction technique. It is used as a pre-processing step in Machine Learning and applications of pattern classification. ... LDA is a supervised classification technique that is considered a part of crafting competitive machine learning models.

K NEAREST NEIGHBOUR ALGORITHM:

K Nearest Neighbor(KNN) is a very simple, easy to understand machine learning algorithms. KNN used in the variety of applications such as finance, healthcare. In Credit ratings, financial institutes will predict the credit rating of customers. In loan disbursement, banking institutes will predict whether the loan is safe or risky. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.

CART ALGORITHM:

The CART algorithm is a type of classification algorithm that is required to build a decision tree on the basis of Gini'simpurityindex. It is a basic machine learning algorithm and provides a wide variety of use cases.It is used to describe Decision Tree algorithms that may be used for classification or regression predictive modeling issues.

GAUSSIAN NB ALGORITHM:

A Gaussian Naive Bayes algorithm is a special type of NB(Naïve Bayes) algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e, normal distribution.

SVM-SUPPORT VECTOR MACHINE ALGORITHM:

SVM is aone of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

RANDOM FOREST CLASSIFIER ALGORITHM:

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. This algorithm is applied in various industries such as banking and e-commerce to predict behavior and outcomes

3.4 PROJECT DESCRIPTION

The project aims to build a model that predicts customers that would subscribe to a bank term deposit, to which we will be able to achieve by considering various models and using the best one for the prediction.The project focuses on the use of LogisticRegression and other Machine learning techniques to predict categorical values of whether customers or clients will subscribe to the term deposit or not.

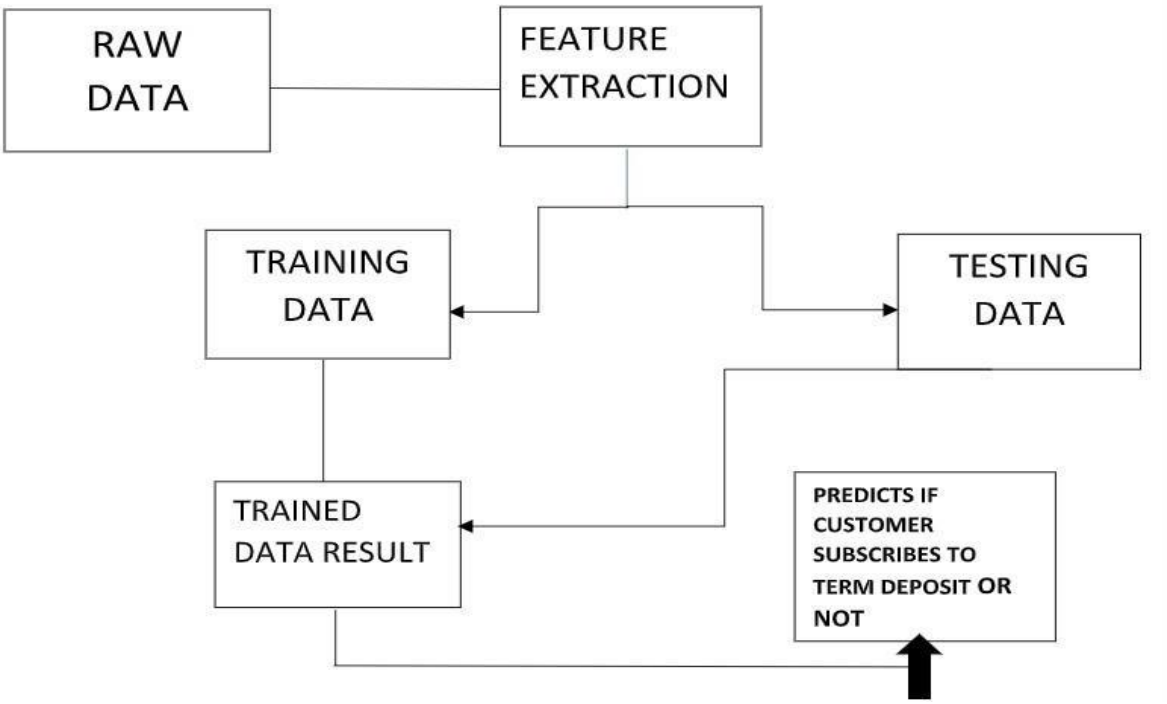

FIG 3.1 DATA FLOW

CHAPTER 4

RESULTS AND DISCUSSION

4.1 SYSTEM REQUIREMENTS

- PROCESSOR:Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz 2.90 GHz
- RAM: 8GB
- SYSTEM TYPE: 64 bit Operating System
- STORAGE:500 GB

4.2 RESULT

As a result,The main aim of the research was to analyze the Customer behavior for making a Termdeposit in banks, by considering some basic attributes that can be collectedandanalyzedeasily. The research was evaluated using seven different machinelearningtechniques(LogisticRegression,LDA,KNN,SVM,DecisionTrees(CART),Gaussian NB,SVM,Random Forest Classifier,). Also, all these techniqueswere tested using Python. The best results were obtained for performing Random Forest Classifierwith the highest predictive accuracyof91 percent.The implemented system performs a predictiveanalysis for the customers who are more responsive to apply for the term-deposits in thebank. The model can help banks in focusing and providing services tothecustomerswho have more probability to apply for term-deposits in the bank. In Future, the modelcan be improved by considering a large dataset as compared to ourmodel.Also,someensemble models can be used to get more accuracy.

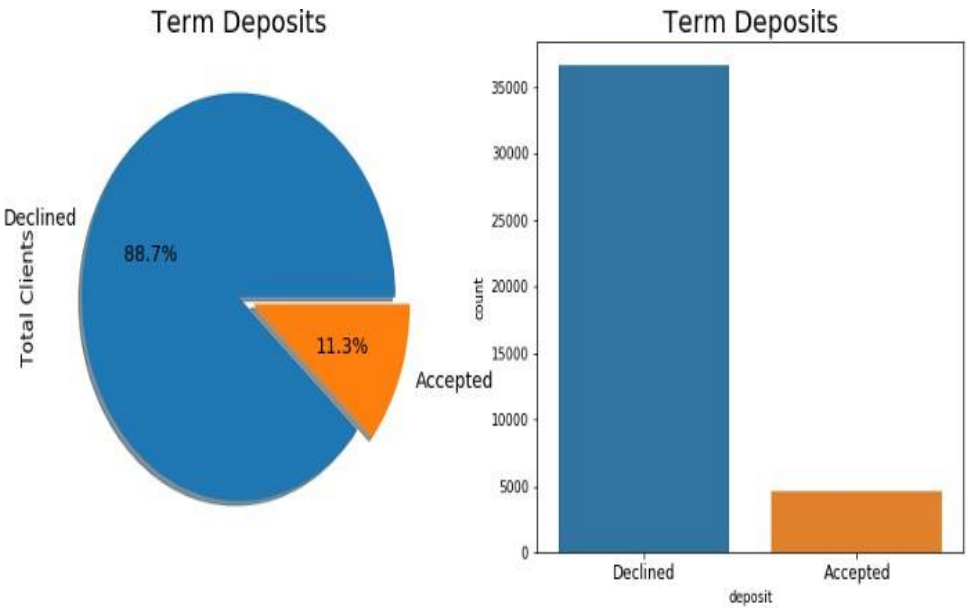

FIG 4.1 PIE CHART AND BAR CHART

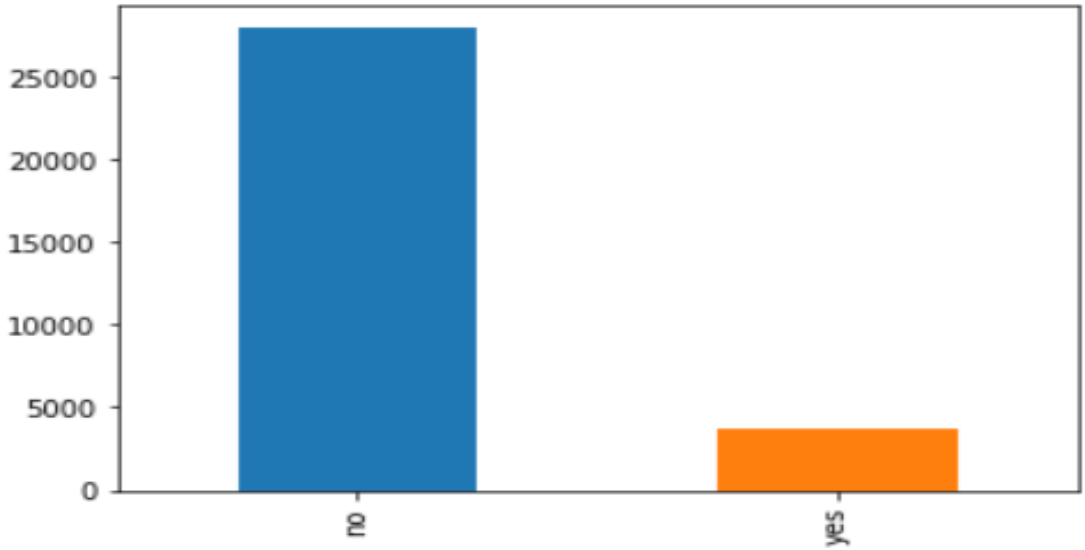

FIG 4.2 UNIVARIATE ANALYSIS

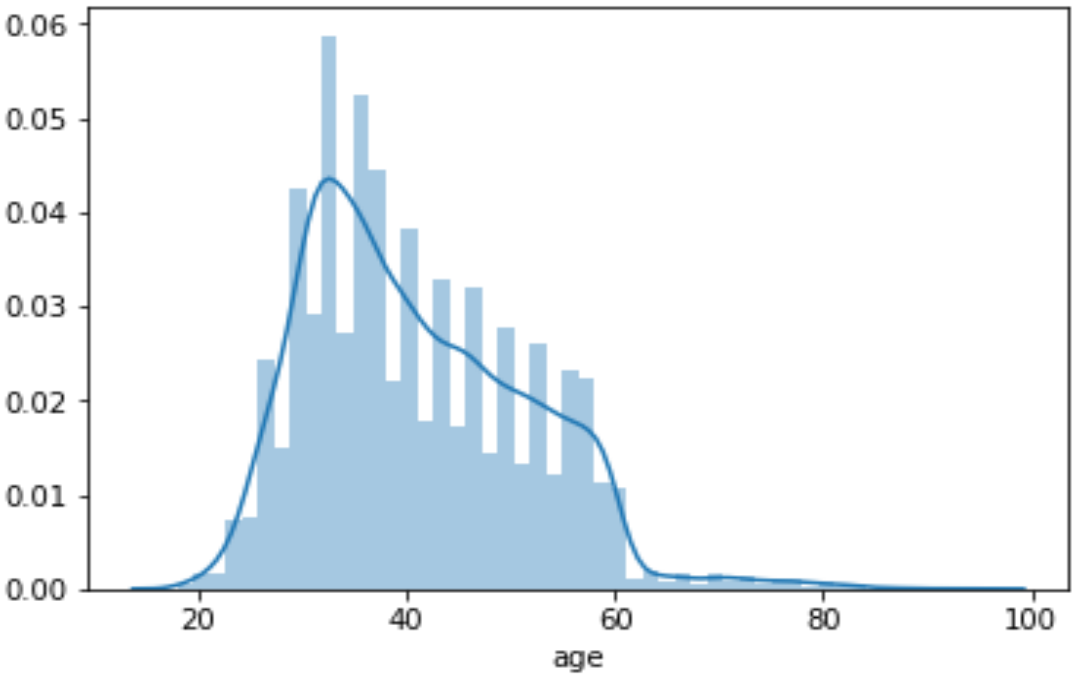

FIG 4.3 DISTRIBUTION OF AGE VARIABLE

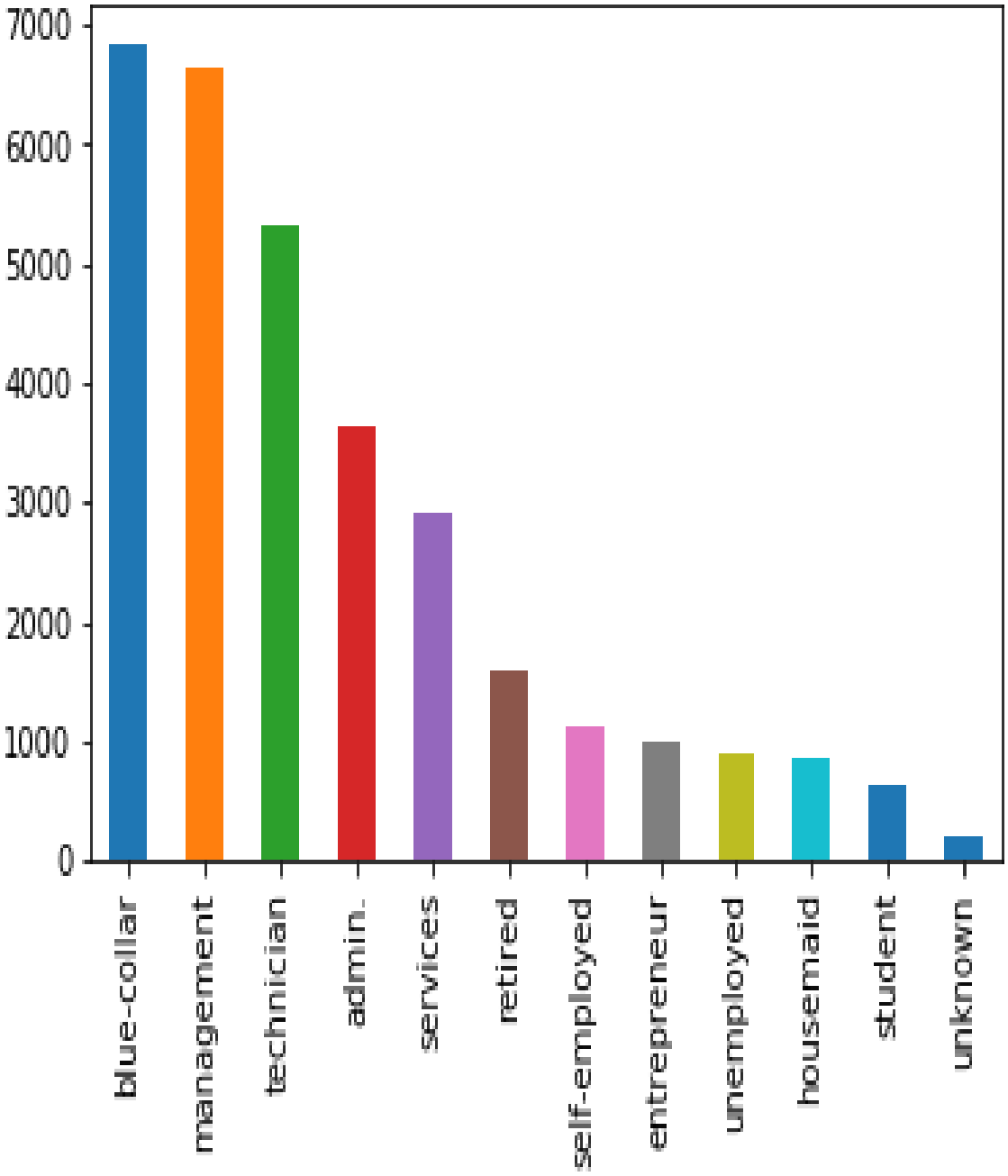

FIG 4.4 DIFFERENT TYPES OF JOBS OF THE CLIENTS

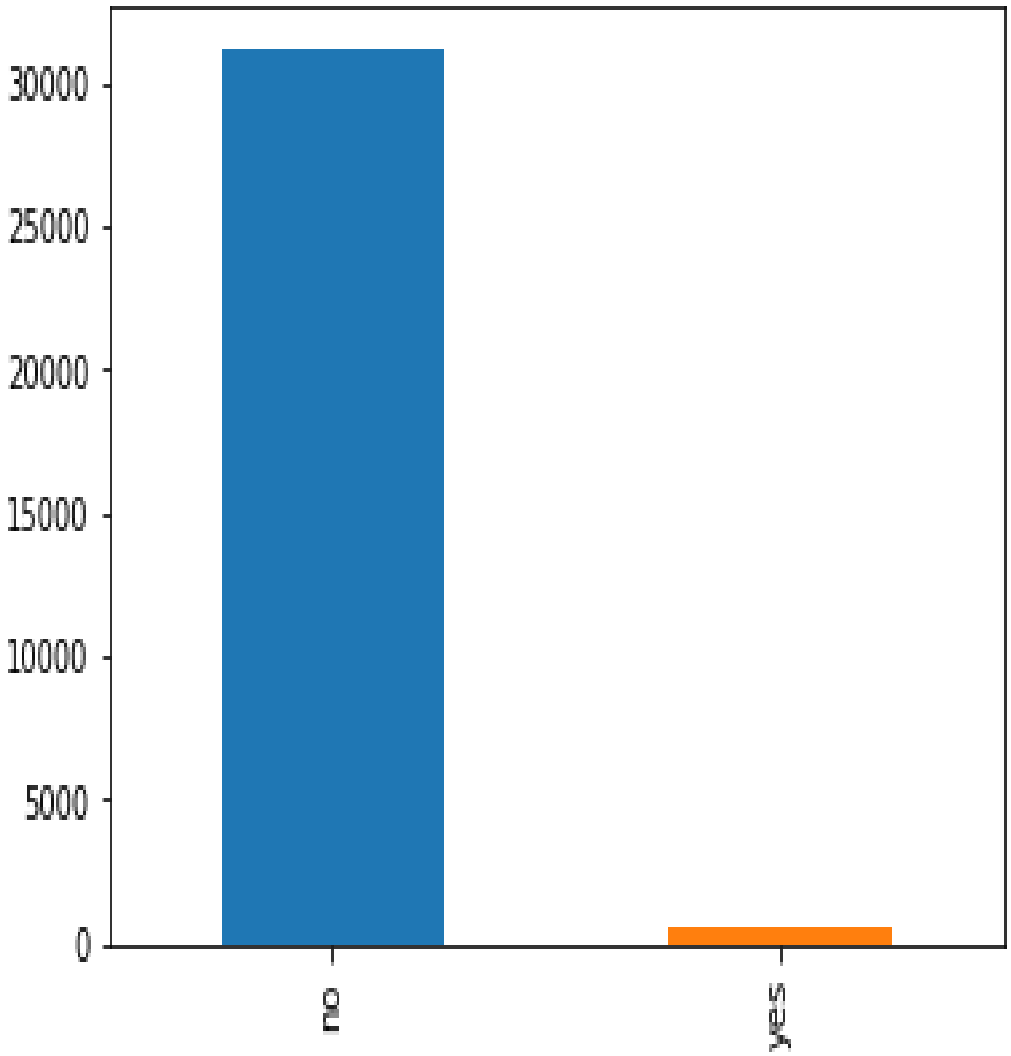

FIG 4.5 CLIENTS DEFAULT HISTORY

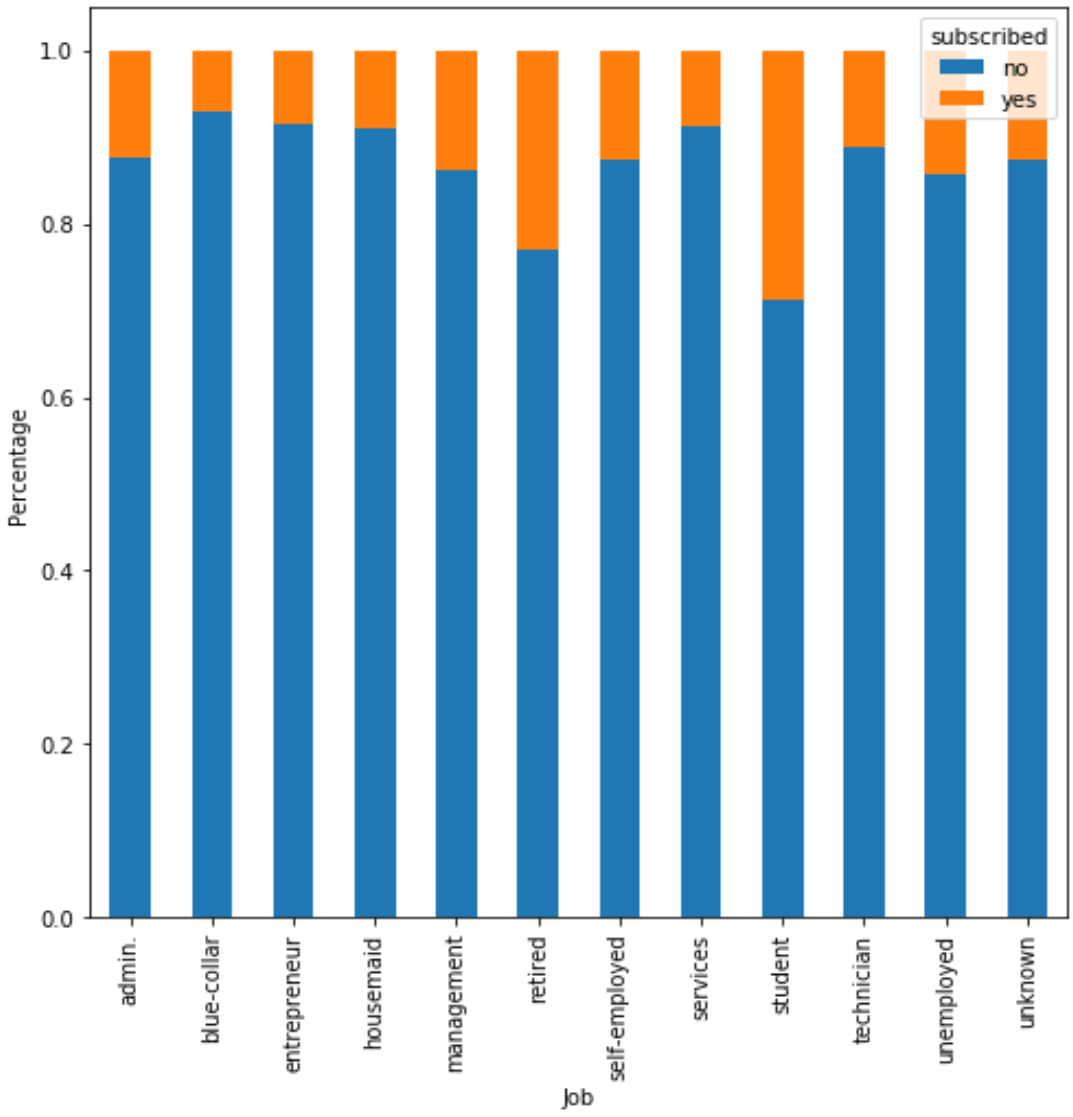

FIG 4.6 BIVARIATE ANALYSIS

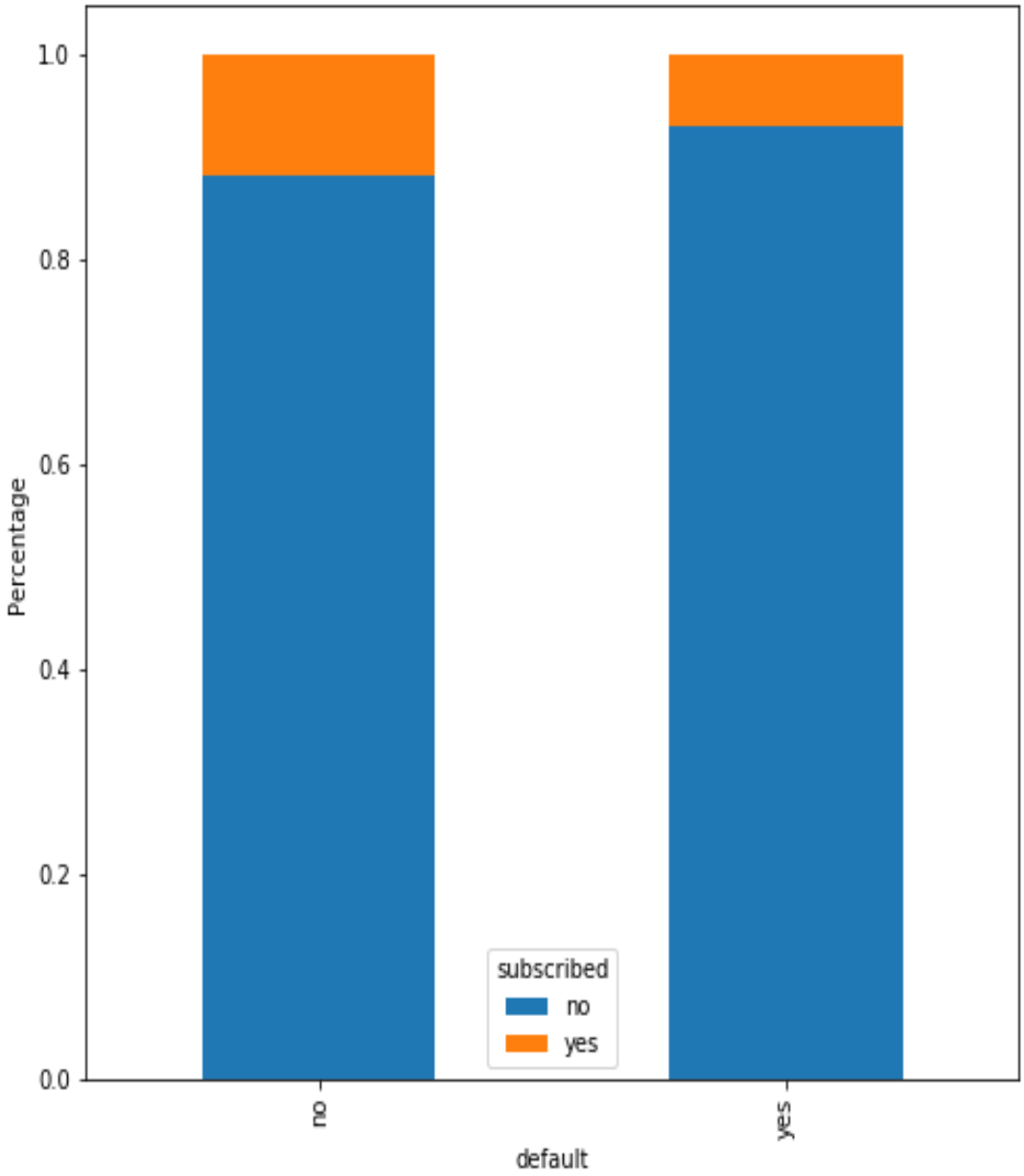

FIG 4.7 PREVIOUS DEFAULT HISTORY

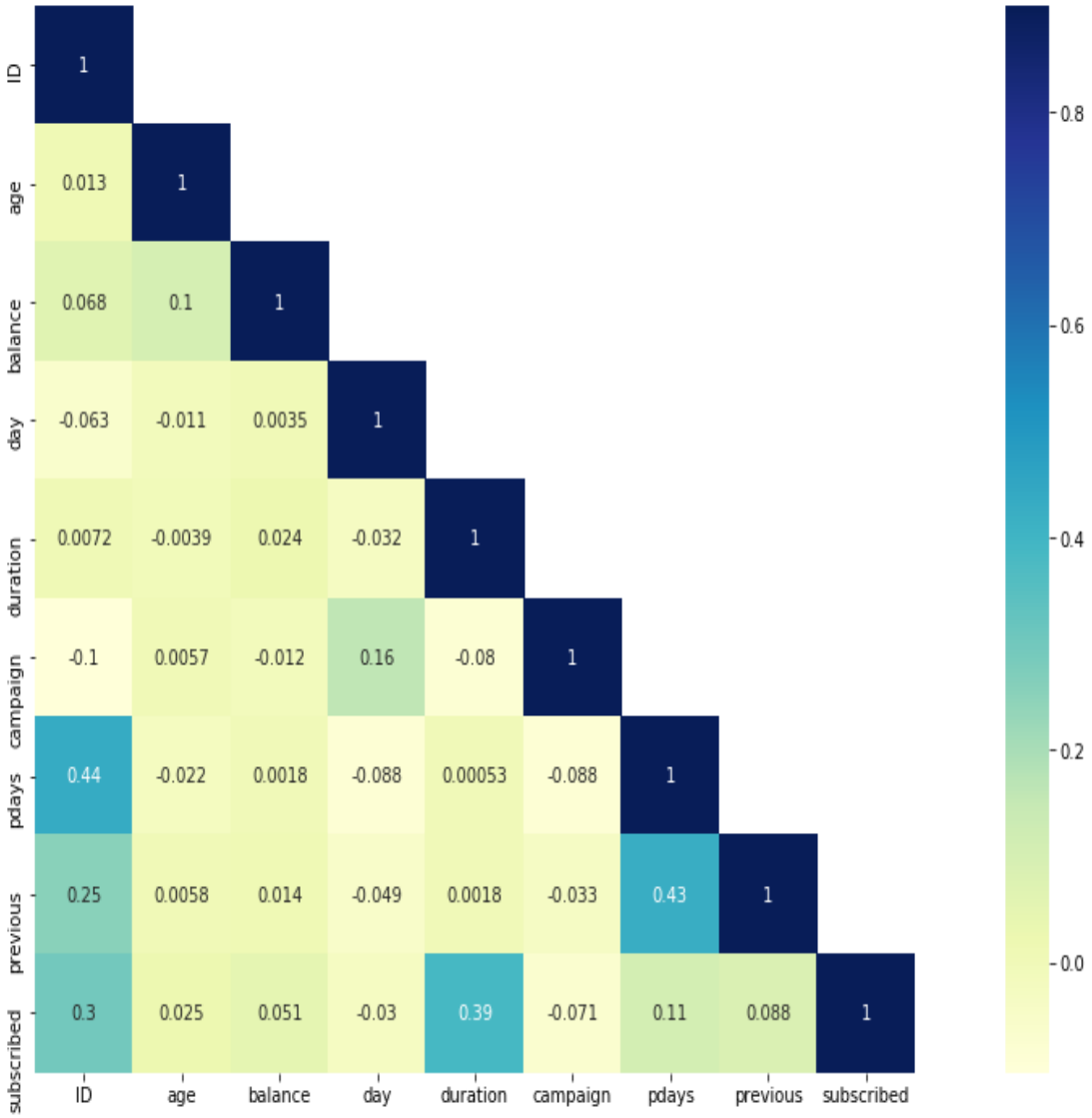

FIG 4.8 CORRELATION BETWEEN VARIABLES

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

This project is undertaken using machine learning and evaluates the performance of prediction by usinglogistic regression algorithm, Linear Discriminant Analysis(LDA)algorithm,KNN,CART,SVM,GaussianNBand Random Forest Classifier Algorithm. In our proposed model, all the methods are involved. The implementation of above system would help in better prediction of whether the customer is interested to subscribe to a term deposit or not, than by assumptions or manual methods.

5.2 FUTURE WORK

This project describes prediction of subscriptions to term deposits in Banks. In future the Bank tellerscan predict thesubscription of customers to the term deposit using the computerised method unlike the normal method using the machine Learning techniques discussed proposed in this project.

REFERENCES

1) Y Dasril, A Alamsyah and T Mustaqim," Bank predictions for prospective long-term deposit investors using machine learning LightGBM and SMOTE", Published under licence by IOP Publishing Ltd(2021)

2) Phan Dun Yung,Tran Duc Hanh and Ta Duc Hanh, "Term Deposit Subscription Prediction Using Spark MLlib and ML Packages", ICEBA 2019: Proceedings of the 2019 5th International Conference on E-Business and Applications (2019)

3) SipuHou,Zongzhen Cai, Jiming Wu, Hongwei Du and Peng Xie, " Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription",published under International Journal of Business Analytics(2016)

4) AshalataPanigrahi and Manik Chand Patnaik," Customer Deposit Prediction Using Neural Network Techniques",published under International Journal of Applied Engineering Research ISSN 0973-4562 Volume 15, Number 3 (2020)

APPENDIX

A.SCREENSHOTS

Splitting Of Data

```
[ ] #splitting of data
    from sklearn.model_selection import train_test_split

    Xtrain,Xtest,ytrain,ytest = train_test_split(X,y,test_size=0.3,random_state=10)
```

```
Xtrain.shape,Xtest.shape,ytrain.shape,ytest.shape
```

```
((28831, 63), (12357, 63), (28831,), (12357,))
```

```
[ ] #Logistic Regression
    from sklearn.linear_model import LogisticRegression
```

```
[ ] model = LogisticRegression()
```

```
[ ] model
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

Building and Evaluation Of Models

```
[ ] # Building  models
    from sklearn.linear_model import LogisticRegression
    from sklearn.tree import DecisionTreeClassifier
    from sklearn.neighbors import KNeighborsClassifier
    from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
    from sklearn.naive_bayes import GaussianNB
    from sklearn.svm import SVC
    from sklearn.ensemble import RandomForestClassifier

    # Spot Checking Algorithms
    models = []
    models.append(('LR', LogisticRegression()))
    models.append(('LDA', LinearDiscriminantAnalysis()))
    models.append(('KNN', KNeighborsClassifier()))
    models.append(('CART', DecisionTreeClassifier()))
    models.append(('NB', GaussianNB()))
    models.append(('SVM', SVC()))
    models.append(('DecisionTree',DecisionTreeClassifier()))
    models.append(('Random Forest',RandomForestClassifier()))
```

```
[ ] # evaluation of each model
    results = []
    names = []
    for name, model in models:
     kfold = model_selection.KFold(n_splits=10, random_state=None)
     cv_results = model_selection.cross_val_score(model, Xtrain, ytrain, cv=kfold, scoring='accuracy')
     results.append(cv_results)
     names.append(name)
     msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
     print(msg)
```

Showing the prediction

```
[ ] # Making predictions on validation dataset
    rf = RandomForestClassifier()
    rf.fit(Xtrain, ytrain)
    predictions = rf.predict(Xtest)
    print(accuracy_score(ytest, predictions))
    print(confusion_matrix(ytest, predictions))
    print(classification_report(ytest, predictions))
```

```
0.9116290361738286
[[10629   306]
 [  786   636]]
              precision    recall  f1-score   support

           0       0.93      0.97      0.95     10935
           1       0.68      0.45      0.54      1422

    accuracy                           0.91     12357
   macro avg       0.80      0.71      0.74     12357
weighted avg       0.90      0.91      0.90     12357
```

```
[ ]
```

Removing Null values

	ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
0	26110	56	admin.	married	unknown	no	1933	no	no	telephone	19	nov	44	2	-1	0	unknown
1	40576	31	unknown	married	secondary	no	3	no	no	cellular	20	jul	91	2	-1	0	unknown
2	15320	27	services	married	secondary	no	891	yes	no	cellular	18	jul	240	1	-1	0	unknown
3	43962	57	management	divorced	tertiary	no	3287	no	no	cellular	22	jun	867	1	84	3	success
4	29842	31	technician	married	secondary	no	119	yes	no	cellular	4	feb	380	1	-1	0	unknown

B. SOURCE CODE

```
# importing libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sn
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
# loading the data
train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
# Print data types for each variable
train.dtypes
#printing first five rows of the dataset
train.head()
UNIVARIATE ANALYSIS
train['subscribed'].value_counts()
# Normalize can be set to True to print proportions instead of number
train['subscribed'].value_counts(normalize=True)
```

 # plotting the bar plot of frequencies

```
train['subscribed'].value_counts().plot.bar()

sn.distplot(train["age"])

train['job'].value_counts().pl

train['default'].value_counts().plot.bar()
```

BIVARIATE ANALYSIS

```
print(pd.crosstab(train['job'],train['subscribed']))

job=pd.crosstab(train['job'],train['subscribed'])

job.div(job.sum(1).astype(float), axis=0).plot(kind="bar", stacked=True,

figsize=(8,8))

plt.xlabel('Job')

plt.ylabel('Percentage')

print(pd.crosstab(train['default'],train['subscribed']))

default=pd.crosstab(train['default'],train['subscribed'])

default.div(default.sum(1).astype(float), axis=0).plot(kind="bar", stacked=True,

figsize=(8,8))

plt.xlabel('default')

plt.ylabel('Percentage')
```

```
train['subscribed'].replace('no', 0,inplace=True)

train['subscribed'].replace('yes', 1,inplace=True)

corr = train.corr()

mask = np.array(corr)

mask[np.tril_indices_from(mask)] = False

fig,ax= plt.subplots()

fig.set_size_inches(20,10)

sn.heatmap(corr, mask=mask,vmax=.9, square=True,annot=True,

cmap="YlGnBu")

train.isnull().sum()
```

MODEL BUILDING

```
target = train['subscribed']
train = train.drop('subscribed
```

 # applying dummies on the train dataset

```
train = pd.get_dummies(train)


from sklearn.model_selection import train_test_split
```

 # splitting into train and validation with 20% data in validation set and 80%

data in train set.

```
X_train, X_val, y_train, y_val = train_test_split(train, target, test_size = 0.2,

random_state=12)
```

LOGISTIC REGRESSION

```
from sklearn.linear_model import LogisticRegression

# defining the logistic regression model

lreg = LogisticRegression()

# fitting the model on  X_train and y_train

lreg.fit(X_train,y_train)

# making prediction on the validation set

prediction = lreg.predict(X_val)

from sklearn.metrics import accuracy_score

# calculating the accuracy score

accuracy_score(y_val, prediction)
```

DECISION TREE

```
from sklearn.tree import DecisionTreeClassifier


# defining the decision tree model with depth of 4, you can tune it further to

improve the accuracy score

clf = DecisionTreeClassifier(max_depth=4, random_state=0)

# fitting the decision tree model

clf.fit(X_train,y_train)

# making prediction on the validation set

predict = clf.predict(X_val)

# calculating the accuracy score
```

```
accuracy_score(y_val, predict)

test = pd.get_dummies(test)

test_prediction = clf.predict(test)

submission = pd.DataFrame()

# creating a Business_Sourced column and saving the predictions in it

submission['ID'] = test['ID']

submission['subscribed'] = test_prediction

submission['subscribed'].replace(0,'no',inplace=True)

submission['subscribed'].replace(1,'yes',inplace=True)

submission.to_csv('submission.csv', header=True, index=False)
```