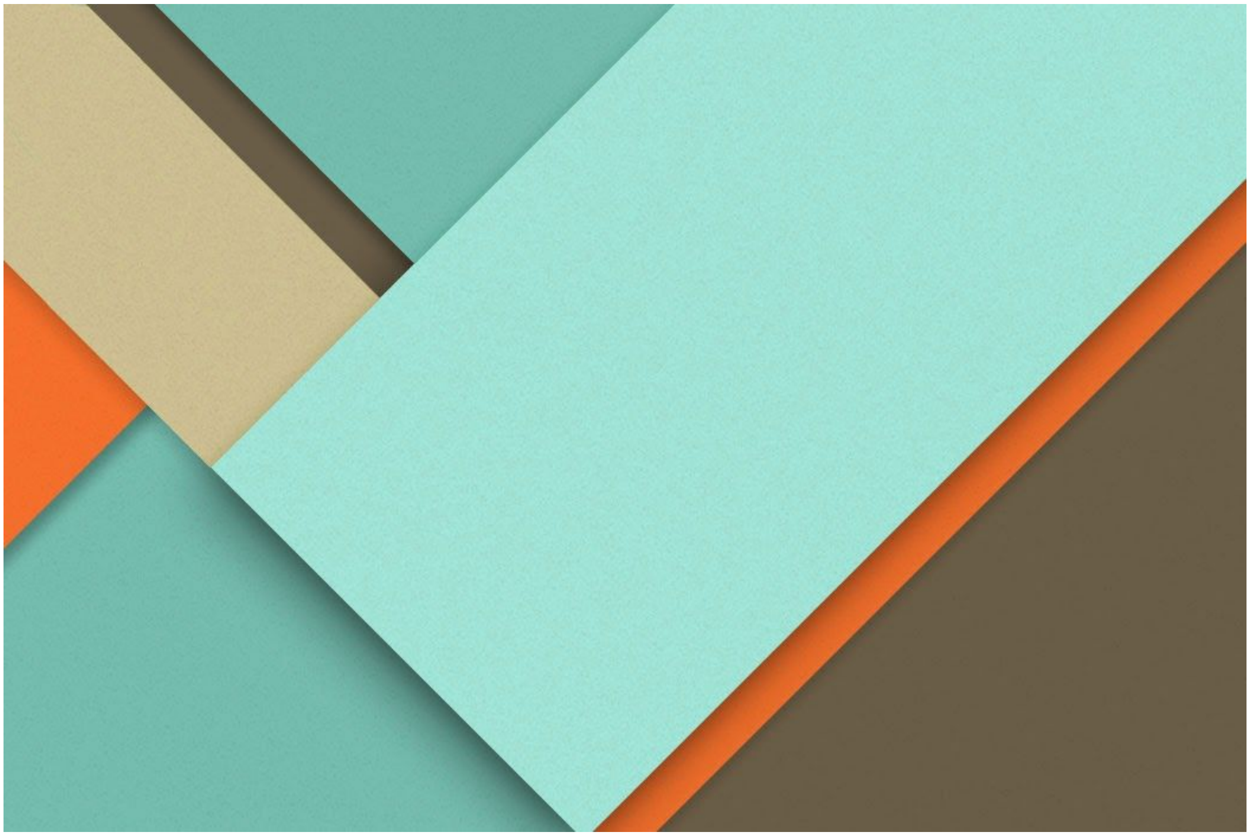

Bank_subscription

ML CODE (for end-sem exam)

Akash Kamerkar
CS19M1001
CSE Department,
National Institute of Technology Puducherry

Overview:

Problem statement of this assignment is as follows: The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Data Set Information:

The data is related with 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.

Attribute Information:

Input variables:

# bank client data:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')

# related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

# social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed to a term deposit? (binary: 'yes','no')

Github Link:

https://github.com/akash9579/work/tree/master/ml_final

Algorithm used:

Exploratory Data Analysis

GaussianNB

RandomForestClassifier

Decision tree classifier

logistic regression

using xgboost [ensemble technique]

Implementation:

We are performing Exploratory Data Analysis on a given dataset. After that we applying various classification algorithm on that dataset

Exploratory Data Analysis does two main things:

1. It helps clean up a dataset.

2. It gives you a better understanding of the variables and the relationships between them.

Components of EDA

To me, there are main components of exploring data:

1. Understanding your variables
2. Cleaning your dataset
3. Analyzing relationships between variables

1. Understanding Your Variables

.shape returns the number of rows by the number of columns for my dataset

.head() returns the first 5 rows of my dataset.

.columns returns the name of all of your columns in the dataset.

.describe() summarizes the count, mean, standard deviation, min, and max for numeric variables

2. Cleaning your dataset

1. Removing Redundant variables
2. Variable Selection
3. Removing Outliers
4. Removing Rows with Null Values

3. Cleaning your dataset

1. Correlation Matrix
2. Scatterplot

After this we gave data to different classification algorithm

Result:

Gaussian Naive Bayes accuracy score -----> 86.19

RandomForestClassifier accuracy score ----->88.62

Decision tree classifier accuracy score -----> 88.69

logistic regression accuracy score ----> 91.03

xgboost accuracy score ----> 91.52