Emre Yesilyurt

yesilyurttemre@gmail.com

Turkey

Dokuz Eylul University

Problem Description:

ABC bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

The Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers whose chances of buying the product is more. This will save resource and time (which is directly involved in the cost (resource billing)).

GitHub Link:

https://github.com/emreyesilyurt/customer_segmentation_in_marketing_campaign

Data Cleansing and Transformation:

1. 2. 3.

4.

I choose the bank_additional_full dataset

Checking for null values
Checking for duplications
does not show that there exist clients with the same details, but it shows that duplicates occur while entering the data. Thus they must be dropped.
Handling Outliers
except for the age, campaign, and duration features. The outliers in the previous features are handled using the IQR method such as:

numerical_features=['age','duration','campaign'] for col in numerical_features:

```
Q1=data[col]. quantile (0.25)
Q3=data[col]. quantile (0.75)
IQR= Q3-Q1
lower_limit=Q1-1.5*IQR
upper_limit=Q3+1.5*IQR
filter=(data[col] >= lower_limit) &(data[col] <= upper_limit) data=data.loc[filter]
```

Two other approaches were also adopted for the dataset: Multivariate approach using the Chi-squared test and the z-score approach using standard deviation. But the best results came from the IQR approach and that is what has been used in the process.

: There is no NA values in our dataset
: There is duplications and these duplicates

: We can notice that all the features have no outliers

Education Feature – Category Clubbing

5.
categories in education into one which means that

'basic.9y','basic.6y','basic.4y' are combined for one educational level which is middle school.

6.
month feature by its real numerical values and also the day_of_week feature by its real numerical values.
Encoding the 999 value of pdays feature by 0:
999 value in pdays feature into 0 numerical value.
Encoding loan, housing, and default features: We have encoded these features using a specific dictionary such as unknown:-1, yes:1, no:0.

: We have clubbed all the

Encoding the month and day_of_week features:

7.

8.

We have encoded the We have converted the

9.
performed One Hot Encoding to the above features and dropped the original features. In addition to that we have dropped the dummy_failure since its result on the target variable is known which means that if the outcome of the previous campaign is failure so consequently the client will make no deposit.

Dummy encoding of contact and poutcome features:

We have

10.
encoded the job feature based on its frequency which means that

Frequency Encoding of Education and Job features: We have

{'admin.':9104, 'blue-collar':8074,'technician':5884, 'services':3450, 'management':2545, 'entrepreneur':1270, 'self-employed':1221, 'retired':1135, 'housemaid':896, 'unemployed':887, 'student':779, 'unknown':279}

Also we have encoded the education feature based on its frequency in such a way {‘middle school’:10678, ‘university.degree’:10548, ‘high.school’:8278, ‘professional.course’:4548, ‘unknown’:1458, ‘illiterate’:14}

11.
using the LabelEncoder() function such a way that we have the following results {‘married’:1, ‘single’:2, ‘divorced’:0, ‘unknown’:3}

Encoding the marital feature:

We have encoded the marital feature