INTERNSHIP–DATA SCIENCE & ANALYTICS-PROJECT 2

Customer Segmentation for Marketing Strategy

Objective: Aim to predict with precision whether bank customers will subscribe to term deposits after the marketing campaigns. Provide actionable insights to help the bank make informed decisions and reinvigorate its revenue streams

Problem: To help the bank to predict accurately whether the customer will subscribe to the focus product for the campaign- Term Deposit after the campaign?

DATA UNDERSTANDING:

The Train dataset has 31647 rows and 18 columns. There are 10 features of object datatype, 3 features of integer and 5 features of float datatype.

The Test dataset has 13564 rows, 17 columns with 10 object datatype features, 2 integer datatype features and 5 features of float datatype.

Statistical Summary:

1. Customer Age ranges from 18 years to 97 years
2. Balance from-8020 to 102128
3. Day of the month from 1st to 31st of the month
4. Last contact duration 0 secs to 4900 secs ie.1.36 hours
5. Number of contacts in campaign ranging from 1 contact to 63 contacts made during the campaign.
6. Days since previous campaign contact ranging from 1 day to 871 days since last contact ie.28 months (2.3years).
7. Number of contacts in previous campaign from 0 to 275 contacts made.
8. Term Deposit subscribed is 0 and 1. (0 for not subscribed and 1 for subscribed)

Features with missing values are Age, Marital, Balance, Personal Loan, Last Contact Duration, Number of contacts in campaign, Days since previous campaign contact.

Numerical Features:

1. Age of customer
2. Balance in the bank account
3. Day of month the customer was contacted

4. Duration of last contact in seconds
5. Number of contacts made during the campaign
6. Number of days since contacted during previous campaign
7. Number of contacts made during previous campaign
8. Term deposit subscribed

Categorical Features:

1. Type of Job of the customer
2. Marital status of customer
3. Education level of customer
4. Default status of any credit
5. Any Housing loan taken
6. Any personal loan taken
7. Mode of communication
8. Month in which customer was contacted
9. Outcome of previous campaign

EXPLORATORY DATA ANALYSIS:

Univariant Analysis shows that

Most of the customers are:

1. in the age group of 30-40
2. Have Balances below 20000
3. Contacted around 20th of the month
4. Call duration less than 1000 secs
5. Contacted upto 10 times
6. Around 200 days ie. 7 months since previous campaign contact
7. Number of contacts in previous campaign is less than 50
8. Majority have not subscribed to term deposits

21.5% of customers have a blue-collar job, 21% are management professionals and 16% are technicians.

60% customers are married, 28% are single and 11% are divorced.

51% customers have had secondary level of education, 29.5% with tertiary level of education

98% customers have not defaulted

55% of customers have housing loan and 84% have not taken personal loan

64.7 % were contacted via cellular mode, 29% unknown

30% of customers were contacted in the month of May and 15% in July and least in December

Only 3.4 % success rate outcome of previous campaign, 11% failed outcomes and 81 % are unknown outcomes

DATA PREPROCESSING:

The missing values greater than 80% was removed. The remaining missing values were imputed using median and mode. The idea is to retain as much data as possible.

Outliers were initially treated using IQR method. But it is to be noticed that removing outliers can lead to missing out on key information and hence for later part of the analysis, outliers were treated differently.

One hot encoding was used initially for all the categorical variables. 2 features – month and job type were reduced to fewer categories and One-Hot Encoding was applied to all the categorical features.

DATA TRANSFORMATION:

Yeo Johnson Transformer was used to treat outliers. It is best used when the data contains negative values. The feature 'Balance' had negative balances and hence it was chosen for treating outliers.

FEATURE REDUCTION:

For Feature reduction, initially number of contacts in campaign and number of contacts in previous campaign was initially clubbed to see if it improved the accuracy. But later it was left as it is.

Principal component Analysis (PCA) was used for feature reduction. It enhanced the accuracy of the algorithm used. Components were reduced to 4 features when only numerical features were used in the algorithm. When all the features were used, PCA reduced it to 12.

FEATURE SCALING:

Robust Scaler was initially used to treat outliers inorder to check the effect on accuracy. But since Yeo Johnson Transformer was used, I didn't use Robust Scaler later on.

Standard Scaler was best for good score. Hence Standard Scaler was preferred over MinMaxScaler.

PREDICTION:

As both Train and Test dataset was provided, Logistic Regression was used to check the prediction by training the Train dataset and using Test Dataset to predict. Test Dataset gave 2 labels (0 and 1). Accuracy of the model was 0.9. Prediction was done only to check if the Test dataset would follow the trained model well or not.

CLUSTERING:

KMeans Clustering, Hierarchical Clustering and DBSCAN Clustering were used for analysis. For week 2 and week 3 submissions Hierarchical Clustering was used. But since it takes high computational time, it was not used for the final code.

DBSCAN turned out to be the best Clustering Algorithm considering its capability to produce better results for data that is dense and having outliers.

All features were initially used for clustering but the scores were low around 0.13 and hence I decided to perform clustering only for the numerical features.

It is considered better to use clustering algorithms only for numerical features and not for categorical. The clustering thus done can then be checked across all the features after clusters are formed.

CUSTOMER SEGMENTATION – BASED ON CLUSTERS FORMED:

DBSCAN Clustering Algorithm has helped in generating 2 clusters.

Cluster 0 : represents those who are not likely to subscribe to Term Deposits

Cluster 1 : represents those who are more likely to subscribe to Term Deposits

Cluster -1 : represents the outliers

MARKETING STRATEGY:

1. Key to a successful marketing campaign is to maintain a regular connect with the customer.
2. For that as the data suggests : maintain regular contact with those clients who maintain higher balances.
3. Approach the blue collar and management professionals who are well educated with secondary or tertiary degrees and who have higher balances. Although their risk appetite is high, for liquidity we can suggest them to invest in Term deposits as well.
4. Those who have less debt, have more disposable income and hence approach those whose liability towards home loan and personal loan is less.
5. Bank should provide attractive interest rates inorder to attract the senior citizens. Retirement corpus for these clients will be on the higher side and they tend to invest in Term deposits and hence attractive interest rate will draw them in.
6. Contact around the 2nd week of the month and not wait till the month end when customers tend to pay any outstanding bills.
7. Approach customers in May, June and November as well, May and June being the mid year and November is before the winter holiday expenses.
8. Married folks are most likely earning double income (couple) and hence for liquidity and to inculcate savings Term Deposits are a good investment option.