Group Name : Solo

Name: Vishnu Priya Malchetti

Email: vishnupriyam.sapbasis@gmail.com

Country: Ireland

Specialization: Data Analyst

Problem description:

We are determined to help XYZ Bank improve its cross-selling strategies and enhance customer engagement. The bank offers a wide array of financial products and services, including savings accounts, credit cards, mortgages, loans, and investment options. However, we've observed that many of our customers have limited product adoption and aren't fully utilizing the range of services available to them. To tackle this challenge head-on, we plan to implement customer segmentation techniques to gain deeper insights into our customer base. By dividing our customers into distinct groups based on their demographics, financial behavior, and product usage patterns, we hope to identify specific customer segments that are more likely to use products and services. Armed with this valuable information, we aim to create personalized marketing strategies and tailored cross-selling initiatives to boost customer satisfaction and encourage higher product adoption. As part of our data analysis team, the objective is to thoroughly analyze the extensive customer dataset provided by XYZ Bank and conduct a comprehensive customer segmentation analysis. The dataset includes detailed information about each customer, such as age, gender, income, transaction history, product holdings, and tenure with our bank.

Data Understanding:

- Customer demographics: Age, gender, location, and purchase history.

- Website interactions: Clickstream data, session duration, and product views.

- Purchase behavior: Cart abandonment, order history, and customer feedback.

- Customer support interactions: Queries, response times, and issue resolution.

- Different products for sale: Credit Card, particular Account, loans and deposits.

Will address the below questions based on our understanding of the data.

What type of data you have got for analysis?

Floats, Integers and objects

```
dtypes: float64(8), int64(23), object(17)
```

What are the problems in the data (number of NA values, outliers , skewed etc)?

Missing values in the training and test datasets

```
#missing values checking
df1.isnull().sum()
```

fecha_dato	0
ncodpers	0
ind_empleado	27734
pais_residencia	27734
sexo	27804
age	0
fecha_alta	27734
ind_nuevo	27734
antiguedad	0
indrel	27734
ult_fec_cli_1t	13622516
indrel_1mes	149781
tiprel_1mes	149781
indresi	27734
indext	27734
conyuemp	13645501
canal_entrada	186126
indfall	27734
tipodom	27735
cod_prov	93591
nomprov	93591
ind_actividad_cliente	27734
renta	2794375
segmento	189368

```
#missing values checking
df2.isnull().sum()
```

```
fecha_dato                    0
ncodpers                      0
ind_empleado                  0
pais_residencia               0
sexo                          5
age                           0
fecha_alta                    0
ind_nuevo                     0
antiguedad                    0
indrel                        0
ult_fec_cli_1t           927932
indrel_1mes                  23
tiprel_1mes                  23
indresi                       0
indext                        0
conyuemp                 929511
canal_entrada              2081
indfall                       0
tipodom                       0
cod_prov                   3996
nomprov                    3996
ind_actividad_cliente         0
renta                         0
segmento                   2248
dtype: int64
```

Outliers in the training dataset:

For age: There are 15891-11370 outliers, which is 4521.

	A	B
1	age	outliers
263	NA	TRUE
1031	NA	TRUE
1065	NA	TRUE
1156	NA	TRUE
1781	NA	TRUE
1852	NA	TRUE
1869	NA	TRUE
1888	NA	TRUE
1919	95	TRUE
1924	NA	TRUE
1926	96	TRUE
2144	NA	TRUE
2420	NA	TRUE
2489	NA	TRUE
2991	NA	TRUE
3345	NA	TRUE

For antigucdad: There are 11374-11370 outliers, which is 4

antiguedad	outliers	q1	q3	upper	lower
6	FALSE	24	154	349	-171
35	FALSE				
35	FALSE				
35	FALSE				

For cod_prov: There are 18332 outliers outcome but they are all from NA values so no outliers.

cod_prov	outliers	q1	q3	upper	lower
29	FALSE	18	33	55.5	-4.5
13	FALSE				
13	FALSE				
50	FALSE				
50	FALSE				
45	FALSE				
24	FALSE				
50	FALSE				
20	FALSE				
10	FALSE				
50	FALSE				
17	FALSE				
49	FALSE				
50	FALSE				
49	FALSE				
8	FALSE				
37	FALSE				
13	FALSE				
13	FALSE				
45	FALSE				
13	FALSE				

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

For missing object values in the dataset, we will delete them by removing rows directly. This approach is straightforward but maybe lead to a loss of valuable data.

For missing values in the columns that contain floats and integers, we will fill in the missing values with estimated or substituted values.

Common methods include using mean, median, or mode for numerical variables, or using the most frequent category for categorical variables.

Outliers in the dataset, first we would handle with missing values and then we deal with outliers. There are some columns have outliers because of NA values, and there are less

outliers that are real outliers in the dataset, so we would deal the real outliers in same way as missing values, whatever substitute them with median, mean or else.