

GROUP NAME: DATA ALCHEMIST

TEAM MEMBERS:

Name: Yuhong Chen
Email: ryechenn@gmail.com
Country: Canada
College/Company: McGill University

Name: Haoyue Chang
Email: kerrychy1215@gmail.com
Country: USA
College/Company: Northeastern University

Name: Gifty Osei
Email: gifty18osei@gmail.com
Country: USA
College/Company: Montana State University

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.

What type of data you have got for analysis

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

Floats, integers, and objects

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

Missing values in training dataset:

```
#missing values checking
```

```
df1.isnull().sum()
```

fecha_dato	0	ind_ahor_fin_ult1	0
ncodpers	0	ind_aval_fin_ult1	0
ind_empleado	27734	ind_cco_fin_ult1	0
pais_residencia	27734	ind_cder_fin_ult1	0
sexo	27804	ind_cno_fin_ult1	0
age	0	ind_ctju_fin_ult1	0
fecha_alta	27734	ind_ctma_fin_ult1	0
ind_nuevo	27734	ind_ctop_fin_ult1	0
antiguedad	0	ind_ctpp_fin_ult1	0
indrel	27734	ind_deco_fin_ult1	0
ult_fec_cli_1t	13622516	ind_deme_fin_ult1	0
indrel_lmes	149781	ind_dela_fin_ult1	0
tiprel_lmes	149781	ind_ecue_fin_ult1	0
indresi	27734	ind_fond_fin_ult1	0
indext	27734	ind_hip_fin_ult1	0
conyuemp	13645501	ind_plan_fin_ult1	0
canal_entrada	186126	ind_pres_fin_ult1	0
indfall	27734	ind_reca_fin_ult1	0
tipodom	27735	ind_tjcr_fin_ult1	0
cod_prov	93591	ind_valo_fin_ult1	0
nomprov	93591	ind_viv_fin_ult1	0
ind_actividad_cliente	27734	ind_nomina_ult1	16063
renta	2794375	ind_nom_pens_ult1	16063
segmento	189368	ind_recibo_ult1	0
		dtype: int64	

missing values in testing dataset:

```
#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_lmes	23
tiprel_lmes	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 antigüedad: There are 11374-11370 outliers, which is 4.

antigüedad	outliers	q1	q3	upper	lower
6	FALSE	24	154	349	-171
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
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
1	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				

For renta: There are 431706 outliers.

	T	U	V	W
renta	outliers	q1	q3	upper
35548.74	TRUE			
22220.04	TRUE			
295590.36	TRUE			
61605.09	TRUE			
68318.46	TRUE			
65608.35	TRUE			
64620.57	TRUE			
64194.99	TRUE			
58728.39	TRUE			
68421.36	TRUE			
70777.59	TRUE			
64398.06	TRUE			
171398.85	TRUE			
64031.25	TRUE			
37075.26	TRUE			
245052.27	TRUE			
57155.34	TRUE			
53631.36	TRUE			
289211.4	TRUE			
34406.58	TRUE			
45345.18	TRUE			
28359.36	TRUE			
60972.81	TRUE			

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 those 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. And for 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.