

Group Name : Dream Crushers

Name : Nafeu Hassan

Email: nafeuassanhelal@gmail.com

Country :Bangladesh

College: Jiangxi Normal University

Specialization (Data Science, NLP, Data Analyst): Data Sciences

Name : Aiman Lameesa

Email: st122876@ait.asia

Country :Bangladesh

College: Asian Institute of Technology

Specialization (Data Science, NLP, Data Analyst): Data Science

We have selected “**Customer Segmentation**” for our group project.

Problem description:

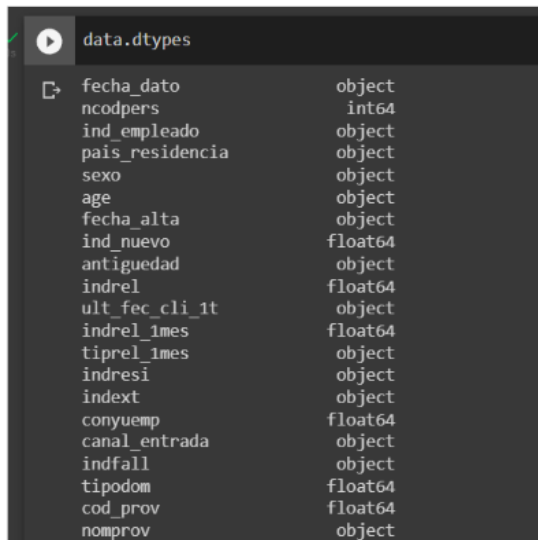
XYZ bank wants to introduce their Christmas offers to their customers. But they do not want to roll out the same offer to all their customers since it will not be profitable to introduce the same offer for different types of customers. Instead, they decide to initiate personalized offers to different sets of customers. Moreover, it is not efficient and beneficial for them to manually understand the hidden patterns in their customer data. That is why, they approached an analytics company, ABC, to help them to understand their customer behaviors in order to introduce Christmas offers. They mentioned to the company that they prefer to have at most 5 groups of customers to maintain the efficiency of their campaign.

Data understanding:

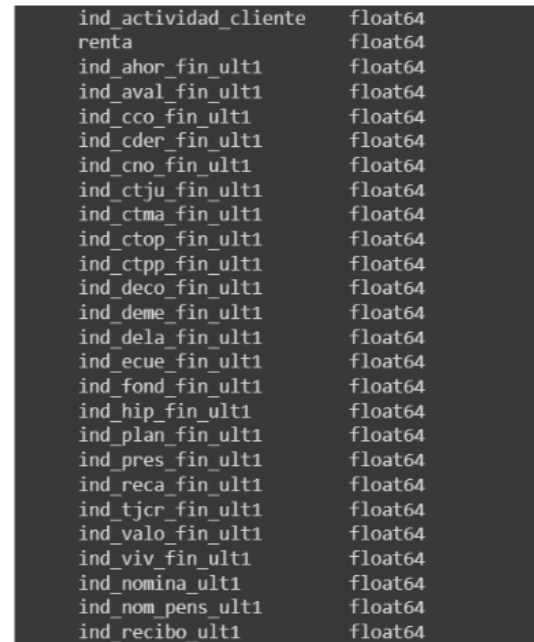
From the problem description, this requires a customer segmentation approach. Since the bank wants to roll out personalized offers to particular sets of customers, it requires Customer Segmentation, which involves analyzing customer behavior based on certain features given in the dataset. From the given data, we will categorize customers into groups based on their behavior, where the customers with the same behavior will form a single category. Thus, we will get several categories for different groups of customers. This approach is handy since not all customers have the same needs and patterns, they however, have similar actions to a particular customer group. To achieve this, we will build an unsupervised classification model based on the data features collected by the bank.

Type of Data

This dataset is a collection of customer data which was collected from the bank . This dataset holds various features about customers' information. The features are mainly numerical and categorical. We performed basic exploratory data analysis on our selected dataset, where we found that most of the columns are numerical and some of them are categorical.



data.dtypes	
fecha_dato	object
ncodpers	int64
ind_empleado	object
pais_residencia	object
sexo	object
age	object
fecha_alta	object
ind_nuevo	float64
antiguedad	object
indrel	float64
ult_fec_cli_1t	object
indrel_1mes	float64
tiprel_1mes	object
indresi	object
indext	object
conyuemp	float64
canal_entrada	object
indfall	object
tipodom	float64
cod_prov	float64
nomprov	object



ind_actividad_cliente	float64
renta	float64
ind_ahor_fin_ult1	float64
ind_aval_fin_ult1	float64
ind_cco_fin_ult1	float64
ind_cder_fin_ult1	float64
ind_cno_fin_ult1	float64
ind_ctju_fin_ult1	float64
ind_ctma_fin_ult1	float64
ind_ctop_fin_ult1	float64
ind_ctpp_fin_ult1	float64
ind_deco_fin_ult1	float64
ind_deme_fin_ult1	float64
ind_dela_fin_ult1	float64
ind_ecue_fin_ult1	float64
ind_fond_fin_ult1	float64
ind_hip_fin_ult1	float64
ind_plan_fin_ult1	float64
ind_pres_fin_ult1	float64
ind_reca_fin_ult1	float64
ind_tjcr_fin_ult1	float64
ind_valo_fin_ult1	float64
ind_viv_fin_ult1	float64
ind_nomina_ult1	float64
ind_nom_pens_ult1	float64
ind_recibo_ult1	float64

Problems in Data

In this dataset, there are columns with null values and outliers as obtained from our exploratory data analysis. If we need to get best output from the unsupervised model, then these problems have to be addressed.

For example, we can observe that columns “conyuemp” and “ult_fec_cli_1t” contain a high number of null values.

data.isnull().sum()	
fecha_dato	0
ncodpers	0
ind_empleado	101
pais_residencia	101
sexo	101
age	0
fecha_alta	102
ind_nuevo	102
antiguedad	1
indrel	102
ult_fec_cli_1t	19457
indrel_1mes	102
tiprel_1mes	102
indresi	102
indext	102
conyuemp	19480
canal_entrada	102
indfall	102
tipodom	102
cod_prov	108
nomprov	108
ind_actividad_cliente	102
renta	3833

ind_ahor_fin_ult1	1
ind_aval_fin_ult1	1
ind_cco_fin_ult1	1
ind_cder_fin_ult1	1
ind_cno_fin_ult1	1
ind_ctju_fin_ult1	1
ind_ctma_fin_ult1	1
ind_ctop_fin_ult1	1
ind_ctpp_fin_ult1	1
ind_deco_fin_ult1	1
ind_deme_fin_ult1	1
ind_dela_fin_ult1	1
ind_ecue_fin_ult1	1
ind_fond_fin_ult1	1
ind_hip_fin_ult1	1
ind_plan_fin_ult1	1
ind_pres_fin_ult1	1
ind_reca_fin_ult1	1
ind_tjcr_fin_ult1	1
ind_valo_fin_ult1	1
ind_viv_fin_ult1	1
ind_nomina_ult1	27
ind_nom_pens_ult1	27
ind_recibo_ult1	1

Moreover, we can also observe some outliers in a few columns such as “ncodpers” and “renta”, which requires further investigation and data analysis.

Q1 = data.quantile(0.25) Q3 = data.quantile(0.75) IQR = Q3 - Q1 print(IQR)	
ncodpers	15088.750
ind_nuevo	0.000
indrel	0.000
indrel_1mes	0.000
conyuemp	NaN
tipodom	0.000
cod_prov	24.000
ind_actividad_cliente	1.000
renta	66182.415
ind_ahor_fin_ult1	0.000
ind_aval_fin_ult1	0.000
ind_cco_fin_ult1	0.000
ind_cder_fin_ult1	0.000
ind_cno_fin_ult1	0.000
ind_ctju_fin_ult1	0.000
ind_ctma_fin_ult1	0.000
ind_ctop_fin_ult1	0.000
ind_ctpp_fin_ult1	0.000
ind_deco_fin_ult1	0.000

There is also skewness among the given data, whereas most of the columns are positively skewed.

`print(data.skew())`

ncodpers	2.555614
ind_nuevo	69.586276
indrel	28.976765
indrel_1mes	0.000000
conyuemp	NaN
tipodom	0.000000
cod_prov	0.126253
ind_actividad_cliente	0.505486
renta	19.435175
ind_ahor_fin_ult1	0.000000
ind_aval_fin_ult1	0.000000
ind_cco_fin_ult1	-3.623115
ind_cder_fin_ult1	0.000000
ind_cno_fin_ult1	5.131387
ind_ctju_fin_ult1	13.576886
ind_ctma_fin_ult1	31.162487
ind_ctop_fin_ult1	0.000000
ind_ctpp_fin_ult1	0.000000

ind_deco_fin_ult1	69.767470
ind_deme_fin_ult1	98.681304
ind_dela_fin_ult1	11.226185
ind_ecue_fin_ult1	4.112409
ind_fond_fin_ult1	22.001173
ind_hip_fin_ult1	98.681304
ind_plan_fin_ult1	37.263544
ind_pres_fin_ult1	98.681304
ind_reca_fin_ult1	9.206944
ind_tjcr_fin_ult1	10.087176
ind_valo_fin_ult1	19.278383
ind_viv_fin_ult1	0.000000
ind_nomina_ult1	6.830283
ind_nom_pens_ult1	6.517667
ind_recibo_ult1	3.252783

Approaches to Overcome the Problems

To deal with null values: If there is any column with a high number of null values, then we will check if that column is important or not. If we find out the feature importance of that column is low or very low, then we will drop that column. Otherwise, we will implement methods like forward and backward fill, mean and median fill or categorical imputation techniques. Moreover, we can implement iterative imputation with a machine learning model to solve this problem.

To deal with null values: There are several ways to deal with outliers in a dataset. We can remove the outliers from the dataset if this will not significantly change the data. Otherwise, we can assign new values instead of these outliers so that a favorable outcome can be generated by the unsupervised model at the end. Moreover, we can transform the outliers by scaling, log transformation, cube root normalization and many other techniques. We can also apply imputation for these outliers and then work on the data.

To deal with skewed data: There are some useful methods to handle skewed data in the dataset like log transformation or square root transformation. We can also try to apply box-cox transformation to check which method performs the best for our selected dataset.

Github Repo Link:

<https://github.com/NafeuHassan/customerSegment-W7>
[aimanlameesa/Week-7 \(github.com\)](https://github.com/aimanlameesa/Week-7)

