Week 8

Team member's details

Group Name: Ctrl_C&Ctrl_V

Name:

1. Yihsuan Sun
2. Tiantian Xie

Email:

- Yihsuan S.: yihsuan.sun88@gmail.com
- Tiantian X.: tenxie0411@gmail.com

Country: United States

College/Company:

- Yihsuan S.: Washington State University
- Tiantian X.: Case Western Reserve University

Specialization: Data Analyst

Problem description & Business understanding

XYZ Credit Union, located in Latin America, does well in selling banking products such as: credit cards, deposit accounts, retirement accounts, safe deposit boxes, etc. However, after statistics, they found that their existing customers basically only buy one product, which means that the bank does not perform well in cross-selling. So XYZ Credit Union wants analysts to build models such as marketing models through machine learning to solve their problems**.**

Data Understanding

Test.csv details:

Total number of observations	929,615
Total number of files	1
Total number of features	24
Base format of the file	csv
Size of the data	105 MB

Train.csv details:

Total number of observations	13,647,309
Total number of files	1
Total number of features	48
Base format of the file	csv
Size of the data	2.13 GB

Important columns:

ncodpers-Customer code
ind_empleado-Employee index: A active, B ex employed, F filial, N not employee, P pasive
pais_residencia-Customer's Country residence
sexo-Customer's sex
age-Age
fecha_alta-The date in which the customer became as the first holder of a contract in the bank
antiguedad-Customer seniority (in months)
ult_fec_cli_1t-Last date as primary customer (if he isn't at the end of the month)
indrel_1mes-Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
nomprov-Province name
renta-Gross income of the household

What type of data you have got for analysis

The data is stored in a CSV file and it contains different data types. Most of the columns we use are integer and string, one can easily make charts and another one can help us do classification, we will provide the results after finishing EDA.

What are the problems in the data

Contains lots of NAs

```
#Some columns contain NA values
df0.isna().any()
```

```
fecha_dato               False
ncodpers                 False
ind_empleado              True
pais_residencia           True
sexo                      True
age                      False
fecha_alta                True
ind_nuevo                 True
antiguedad               False
indrel                    True
ult_fec_cli_1t            True
indrel_1mes               True
tiprel_1mes               True
indresi                   True
indext                    True
conyuemp                  True
canal_entrada             True
indfall                   True
tipodom                   True
cod_prov                  True
nomprov                   True
ind_actividad_cliente     True
renta                     True
segmento                  True
ind_ahor_fin_ult1        False
ind_aval_fin_ult1        False
ind_cco_fin_ult1         False
ind_cder_fin_ult1        False
ind_cno_fin_ult1         False
ind_ctju_fin_ult1        False
ind_ctma_fin_ult1        False
ind_ctop_fin_ult1        False
ind_ctpp_fin_ult1        False
ind_deco_fin_ult1        False
ind_deme_fin_ult1        False
ind_dela_fin_ult1        False
ind_ecue_fin_ult1        False
ind_fond_fin_ult1        False
ind_hip_fin_ult1         False
ind_plan_fin_ult1        False
ind_pres_fin_ult1        False
ind_reca_fin_ult1        False
ind_tjcr_fin_ult1        False
ind_valo_fin_ult1        False
ind_viv_fin_ult1         False
ind_nomina_ult1           True
ind_nom_pens_ult1         True
ind_recibo_ult1          False
dtype: bool
```

```
pd.set_option('display.max_columns', None)
df0.head()
print("num columns: ", df0.shape[0])
```

```
num columns:  13647309
```

```
#Drop column "ult_fec_cli_1t" and "conyuemp" because all values are NAN
del df0['ult_fec_cli_1t']
del df0['conyuemp']
```

```
#Remove all rows which has NA value. Now there is no NA value in data

df = df0.dropna()
df.isna().any()
```

```
fecha_dato              False
ncodpers                False
ind_empleado            False
pais_residencia         False
sexo                    False
age                     False
fecha_alta              False
ind_nuevo               False
antiguedad              False
indrel                  False
indrel_1mes             False
tiprel_1mes             False
indresi                 False
indext                  False
canal_entrada           False
indfall                 False
tipodom                 False
cod_prov                False
nomprov                 False
ind_actividad_cliente   False
renta                   False
segmento                False
ind_ahor_fin_ult1       False
ind_aval_fin_ult1       False
ind_cco_fin_ult1        False
ind_cder_fin_ult1       False
ind_cno_fin_ult1        False
ind_ctju_fin_ult1       False
ind_ctma_fin_ult1       False
ind_ctop_fin_ult1       False
ind_ctpp_fin_ult1       False
ind_deco_fin_ult1       False
ind_deme_fin_ult1       False
ind_dela_fin_ult1       False
ind_ecue_fin_ult1       False
ind_fond_fin_ult1       False
ind_hip_fin_ult1        False
ind_plan_fin_ult1       False
ind_pres_fin_ult1       False
ind_reca_fin_ult1       False
ind_tjcr_fin_ult1       False
ind_valo_fin_ult1       False
ind_viv_fin_ult1        False
ind_nomina_ult1         False
ind_nom_pens_ult1       False
ind_recibo_ult1         False
dtype: bool
```

All the rows show that there’s no more NAs. (The process are above)