Week 8 assignment

Batch code: LISUM05
Submission date: 02/12/2022
Submitted to: Data Glacier
Group Name: Data warriors
(Data Analyst :: Cross selling recommendation - Group Project)

Team member's details:

Name: Christopher Irvin Ballon Peralta
Email: cballon@uoc.edu
Country: Peru
College/Company: Universitat Oberta de Catalunya
Specialization: Data Science

Problem description

XYZ credit union in Latin America is performing very well in selling the Banking products (eg: Credit card, deposit account, retirement account, safe deposit box etc) but their existing customer is not buying more than 1 product which means bank is not performing good in cross selling (Bank is not able to sell their other offerings to existing customer). XYZ Credit Union decided to approach ABC analytics to solve their problem.

Data understanding

The features of the data set is described following:

Type of data

The dataset has categorical and numerical features:

fecha_dato	date
ncodpers	int
ind_empleado	int
pais_residencia	string
sexo	string
age	int
fecha_alta	date
ind_nuevo	int
antiguedad	int
indrel	int
ult_fec_cli_1t	date
indrel_1mes	int
tiprel_1mes	int
indresi	int

indext	int
conyuemp	int
canal_entrada	int
indfall	int
tipodom	int
cod_prov	int
nomprov	string
ind_actividad_cliente	float
segmento	string
renta	float
ind_ahor_fin_ult1	int64
ind_aval_fin_ult1	int64
ind_cco_fin_ult1	int64
ind_cder_fin_ult1	int64
ind_cno_fin_ult1	int64
ind_ctju_fin_ult1	int64
ind_ctma_fin_ult1	int64
ind_ctop_fin_ult1	int64
ind_ctpp_fin_ult1	int64
ind_deco_fin_ult1	int64
ind_deme_fin_ult1	int64
ind_dela_fin_ult1	int64
ind_ecue_fin_ult1	int64
ind_fond_fin_ult1	int64
ind_hip_fin_ult1	int64
ind_plan_fin_ult1	int64
ind_pres_fin_ult1	int64
ind_reca_fin_ult1	int64
ind_tjcr_fin_ult1	int64
ind_valo_fin_ult1	int64
ind_viv_fin_ult1	int64
ind_nomina_ult1	float64
ind_nom_pens_ult1	float64
ind_recibo_ult1	int64

The dataset corresponds to registers from clients from a bank, the strategy is to analyze the data to get which are the better combinations of products to offer.

Approach

We're planning to detect NA values using a python function and impugn the missing fields depending on two cases:
First: the number of entries is in order of 80% from the total rows, fill the NA with "empty" or "unknown", if the column is numerical, we can utilize mean filling from the rest of column values.
Second: the number of entries is minor than 50%, we'll delete the column.

Next, we need to graph the distribution values using box plots to detect outliers and utilize quartile margin to remove outliers (python function) and do EDA.