Name: Deepthika Shiwani Muralikrishnan, Payal Upadhyay, Abida S Bhatti

Group Name: Team Data Analysts

Email: deepthikamk@gmail.com , payalupadhyay456@gmail.com , sarwarabida123@gmail.com

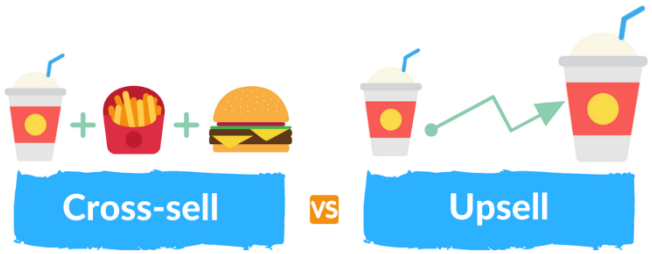
Country: India, South Africa, USA

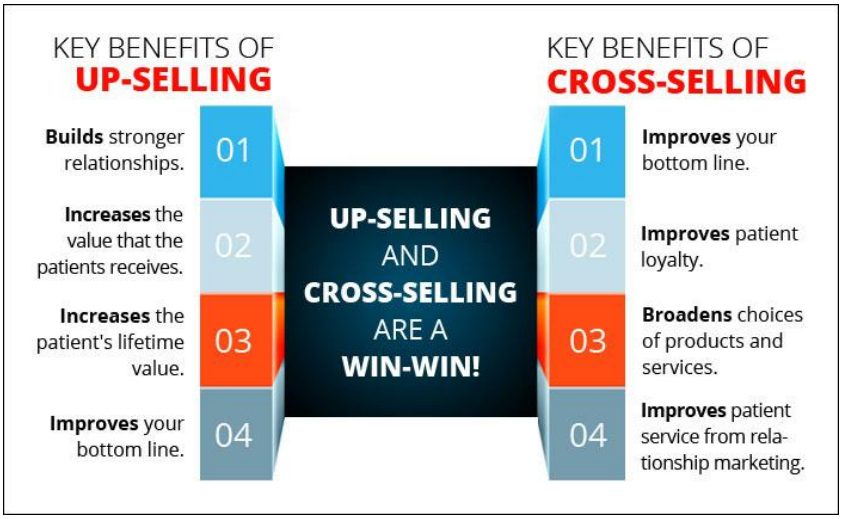
College/Company: -

Specialization: Data Analyst

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. ABC company came up with a framework which will be utilizing machine learning algorithm in the core to increase cross selling. But as a data analyst you need to inspect the data and suggest what action bank can take to increase cross selling (without using ML)





DATA UNDERSTANDING

Column Name	Description
fecha_dato	The table is partitioned for this column
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
ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.
antiguedad	Customer seniority (in months)
indrel	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
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)
tiprel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
conyuemp	Spouse index. 1 if the customer is spouse of an employee
canal_entrada	channel used by the customer to join
indfall	Deceased index. N/S
tipodom	Addres type. 1, primary address
cod_prov	Province code (customer's address)
nomprov	Province name
ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
renta	Gross income of the household
segmento	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated
ind_ahor_fin_ult1	Saving Account

ind_aval_fin_ult1	Guarantees
ind_cco_fin_ult1	Current Accounts
ind_cder_fin_ult1	Derivada Account
ind_cno_fin_ult1	Payroll Account
ind_ctju_fin_ult1	Junior Account
ind_ctma_fin_ult1	Más particular Account
ind_ctop_fin_ult1	particular Account
ind_ctpp_fin_ult1	particular Plus Account
ind_deco_fin_ult1	Short-term deposits
ind_deme_fin_ult1	Medium-term deposits
ind_dela_fin_ult1	Long-term deposits
ind_ecue_fin_ult1	e-account
ind_fond_fin_ult1	Funds
ind_hip_fin_ult1	Mortgage
ind_plan_fin_ult1	Pensions
ind_pres_fin_ult1	Loans
ind_reca_fin_ult1	Taxes
ind_tjcr_fin_ult1	Credit Card
ind_valo_fin_ult1	Securities
ind_viv_fin_ult1	Home Account
ind_nomina_ult1	Payroll
ind_nom_pens_ult1	Pensions
ind_recibo_ult1	Direct Debit

There are 24 different types of products offered.

PROBLEMS IN DATA

Train.csv

Columns with NA values:

ind_empleado	27734
pais_residencia	27734
sexo	27804
fecha_alta	27734
ind_nuevo	27734
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
ind_nomina_ult1	16063
ind_nom_pens_ult1	16063

20.47% of income values are NA

Test.csv

Columns with NA values:

ult_fec_cli_1t	927932
indrel_1mes	23
tiprel_1mes	23
conyuemp	929511
canal_entrada	2081
cod_prov	3996
nomprov	3996
segmento	2248

Outliers:

- In train data set, the age group between 1-18 is only 0.8 % of the total data and have minimum impact in the overall analysis in regression.
- In train data set, age group between 80 -164 is only 2.7 % of the total data and have minimum impact in the overall analysis in regression.

APPROACHES TO OVERCOME PROBLEMS

In our data analysis we used following approaches to resolve NA values and outlier issues:

1. To resolve NA issue in age column, create data set for average income of each group and apply it to each age row where income is 'NA'.

Age Groups:

Group 1 (19-25 yr old customers):

E-account, saving account, short-term deposit, medium-term deposit, long-term deposit, loans, and credit cards.

Group 2 (26-45 yr old customers):

Mortgage, e-account, savings, credit cards, pension plan, loans, and securities.

Group 3 (46-65+ yr old customers):

E-account, loans, short-term deposit, savings, credit cards, securities, and pension plan.

2. To resolve outlier, we will ignore age groups between 1-18 (.8% of dataset) and 85-164(2.7% of data set) as their impact is minimal on overall analysis.

3. Non-numeric missing values can be replaced with the mode (most occurring value) in those categories.

4. Delete products that nobody buys (only a few customers).

5. Delete rows where a lot/all values are missing.

Github repository link : https://github.com/deepthikashiwani/Cross-selling-Recommendation-How-to-increase-cross-selling-of-Banking-Products