Week 9 - Data Analyst: Cross-selling recommendation

Team Member Details: Individual project (no team)

Name: Asha K. C.

Email: aasha.kc@outlook.com

Country: India

Company: DataGlacier

Specialization: Data Analyst

Problem Description:

The XYZ bank is having difficulty cross-selling its products to existing customers. Customers are not buying additional products sold by their bank. Hence, as data analysts, we must provide the information required to enhance their cross-selling methods.

Github Repo link:

https://github.com/Asha-KC-07/Data-Glacier-Internship-2025---LISUM43/blob/main/Week%209%20-%20Data%20Analyst_Cross-selling%20recomendation/CleanData.ipynb

Data cleansing and transformation that was done on the data:

1. Identified the range of customer records based on min & max of 'fecha_dato'

   Date Range in 'fecha_dato':

   Start Date: 2015-01-28 00:00:00

   End Date: 2016-05-28 00:00:00

2. Fetched the overall missing values count and percentage in the dataframe. Displayed them in descending order of nan percentage

	Count	Percentage
conyuemp	*13645501*	*99.986752*
ult_fec_cli_1t	*13622516*	*99.818330*
renta	*2794375*	*20.475648*

segmento 189368 1.387585

canal_entrada 186126 1.363829

indrel_1mes 149781 1.097513

tiprel_1mes 149781 1.097513

nomprov 93591 0.685784

cod_prov 93591 0.685784

sexo 27804 0.203732

tipodom 27735 0.203227

indfall 27734 0.203220

ind_actividad_cliente 27734 0.203220

ind_empleado 27734 0.203220

pais_residencia 27734 0.203220

indext 27734 0.203220

indresi 27734 0.203220

indrel 27734 0.203220

ind_nuevo 27734 0.203220

fecha_alta 27734 0.203220

ind_nomina_ult1 16063 0.117701

ind_nom_pens_ult1 16063 0.117701

3. Analysed duplicates in the ncodpers column (This is customer code, hence one record of each is enough for analysis). After validating all features for a single customer, I decided to keep only the last available record per customer.

Number of duplicate customers in dataset: 12690664

Top 10 duplicate values and their counts:

ncodpers

1375586 17

42515 17

42636 17

42684 17

42685 17

42686 17

```
42690    17
42695    17
42697    17
42704    17
Name: count, dtype: int64
```

Percentage of duplicate values: 92.99%

4. Analyse values in the conyuemp column (Spouse index)

```
Percentage of NaN values in 'conyuemp' column: 99.99%
```

Since most of the conyuemp column is 'nan', we remove the column from the dataframe.

5. Analyse values in the ult_fec_cli_1t column (date showing last date in month when customer is primary)

```
Percentage of NaN values in 'ult_fec_cli_1t' column: 97.98%
```

Apply the max date to the nan values in 'ult_fec_cli_1t' where 'indrel' is 1. 'indrel' = 1 means they are primary customers

```
Number of rows with indrel = 1: 930285
Number of rows with indrel = 1 AND ult_fec_cli_1t = NaN: 930285
Number of rows with indrel = 1 AND ult_fec_cli_1t = NaN after filling
with a date: 0
Number of rows updated: 930285

Percentage of NaN values in 'ult_fec_cli_1t' column: 0.73%
```

6. Analysing the next highest nan column - 'rento' (income of the customer). Computed the MOD value for NaN values

Percentage of NaN values in 'renta' column after filling: 0.00%

7. Analysing the next highest nan column - 'segmento'. Dropped values that did not contribute to the analysis. Same way dropped values for ‘canal_entrada’, 'indrel_1mes', 'tiprel_1mes', 'cod_prov'. These showed .5% of NaN values in the dataframe.
8. Converted date columns to a datetime format for 'fecha_dato' & 'fecha_alta'.

   Converted ‘age’, ‘antiguedad’, ‘indrel_1mes’ to int dtype.

9. Replacing ‘sexo’ feature H - Male and V - Female for easy readability.
10. ‘cod_prov’ & ‘nomprov’ had the same information. Country of the customer. Hence, removed cod_prov.