Data Intake Report

Name: Cross-Selling Opportunity Analysis for XYZ Credit Union
Report date: 19 May 2025
Internship Batch: LISUM44
Version: 1.0
Data intake by: Raj Pawar, Naga Pavithra Jajala
Data intake reviewer:
Data storage location: Local (Original file source – GitHub – Data Glacier)

Tabular data details:

1. **Train.csv**

Total number of observations	~13 million
Total number of files	1
Total number of features	48
Base format of the file	.csv
Size of the data	2.13 GB

2. **Test.csv**

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

1. Deduplication Validation

- Initial review indicates ***ncodpers*** is the unique customer ID
- Will validate uniqueness and drop duplicates if required after join/merge operations
- ***.duplicated()*** method to be used after feature engineering to ensure no row-level duplication

2. Assumptions for Data Quality

- Columns like ***age, antiguedad,*** and ***renta*** were stored as strings and required conversion.
- **age** and ***antiguedad*** contain values like **"NA"** and **"NA"** - will be stripped and converted to numeric.
- Income ***(renta)*** is extremely sparse in the train set (~21.4% missing).
- Columns ***ult_fec_cli_1t*** and ***conyuemp*** are almost entirely null — to be dropped or imputed based on domain knowledge.

- Data spans monthly customer records and includes numerous binary flags for product ownership.

Null Values Summary

Train.csv

- **High Missing Columns:**
  - *ult_fec_cli_1t*: 13.6M missing
  - *conyuemp*: 13.6M missing
  - *renta*: 2.79M missing
- **Moderate Missing:**
  - *segmento*: 189k
  - *canal_entrada*: 186k
  - *indrel_1mes / tiprel_1mes*: ~149k

Test.csv

- **High Missing:**
  - *ult_fec_cli_1t*: 927k missing
  - *conyuemp*: 929k missing
- **Moderate Missing:**
  - *segmento*: 2.2k
  - *canal_entrada*: 2k

Observations

- Majority of product ownership columns are binary flags (0/1)
- Many categorical columns are ***string[pyarrow]*** (e.g., ***segmento, sexo, pais_residencia***)
- Several identifiers (like ***indrel_1mes***) contain mixed types (***1, 2, 'P'***) and are treated as **object**
- Data integrity is generally strong outside of a few sparse and stringified columns
- Use of **Dask** is essential to scale processing and transformation on the 2GB+ train file