Santander Product Recommendation | Week 2 — Ingest and Explore the Dataset

Maria Alice Vieira and Aritra Ray

- **What is the target variable and why?**

  The target variable in the Santander competition is which product a client will buy in the last month, precisely on **2016-06-28**, that they did not already have the month before **2016-05-28**. These target variables are the last 24 columns, initially named ind_(xyz)_ult1, now reflecting the product in a clearer way. These columns reflect specific product categories such as savings accounts, credit cards, and mortgages. In other words, the purpose is to estimate which of these products will be added by a client in June 2016.

  The problem statement is based on the idea of developing a recommendation system for Santander to forecast which things their clients are likely to buy in the coming month. The goal behind this is to ensure that the bank can provide more personalized suggestions, hence increasing client happiness by offering suitable items.

- **What are the predictors and why?**

  This dataset's predictors include a variety of consumer behavior, demographic, and engagement factors that provide insight into past actions and characteristics, assisting in forecasting future product adoptions. Key predictors are the product columns from prior months, which show which products a consumer already has, as this past behavior might forecast future purchases. Furthermore, customer demographics such as age, income, and segment (VIP, individual, or college graduate) provide context for financial needs, while engagement features such as seniority with the bank, activity level, and customer relationship status reflect the level of interaction with the bank's products. These predictors work together to construct a detailed profile of each consumer, allowing the model to estimate which goods they are most likely to adopt next.

- **Exploration of the dataset: definition of variables, data types, general dataset stats: count of rows, count of columns, etc.**

  The raw dataset, provided by Santander, initially contained 1.5 years of data and was over 2.2 GB in size. To be able to work in Jupyter, we reduced the file size to under 2 GB by removing the first 6 months of data. We are going to focus on the most recent customer-product interactions, since they are crucial for our goal of matching products with customers in the last month of the dataset.

Our initial exploration of the data was done using chunks in pandas to visualize the data and determine the range of dates in the dataset. To handle the large volume of data efficiently, we used Dask, a computing library that provides parallel processing capabilities. It allows us to work with datasets larger than the available memory by dividing the data and processing it in parallel across multiple processors or machines. Its functionality is similar to a pandas dataframe but operates in parallel. After exploring the data and determining the appropriate date cutoff and data types for each column using chunks, we used Dask to ingest the dataset. We then proceeded to make the necessary changes, initially clean the dataset, and create the final training version that was uploaded to Jupyter.

The raw data provided by Santander had column names in Spanish. To facilitate the analysis, we changed the column names to English based on the descriptions provided by the bank. The final training dataset now consists of 1.5GB, 45 columns and 10,501,007 rows, with dates ranging from June 1st, 2015 to May 28th, 2016.The dataset is structured into two main categories of columns. The first 21 columns contain client demographics and information about their relationship with the bank, including the date of the record. These columns vary in type, including date, string, object, float, and integer. Some key variables in this category are sex, age, first contract date, seniority in months, whether a new customer or not, province name, and income. The remaining 24 columns represent the products offered by the bank. These are all dummy variables, indicating whether a customer has a particular product (1) or not (0) on the recorded date. Examples of these products include savings account, current account, mortgage, loans, and credit card. After processing, the final data types in the database are as follows: 2 columns of datetime64, 1 column of float64, 30 columns of int32, and 12 columns of object type.

During the data preparation process, we encountered some NA values. We had two approaches to handle these: rows with missing data across multiple columns were dropped, while other NA values were temporarily filled with 0. This approach will be refined in the upcoming EDA phase. Two variables that require particular attention are income and age. The income variable initially had 2,240,788 null values, which are currently set to 0. Given the likely importance of this variable, we plan to develop a more sophisticated approach to impute these values, possibly using the average income by province. The age variable also needs adjustment, as it currently ranges from 2 to 164 years, which includes some unrealistic values.

Appendix

Columns and datatypes:

```
Data columns (total 45 columns):
 #   Column                Dtype
---  ------                -----
 0   date                  datetime64[ns]
 1   customer_code         int32
 2   employee_index        object
 3   country               object
 4   sex_H                 object
 5   age                   int32
 6   first_contract_date   datetime64[ns]
 7   new_cust              int32
 8   seniority_in_months   int32
 9   primary_cust          int32
 10  last_date_primary     object
 11  cust_type             object
 12  cust_relationship     object
 13  residency_spain       object
 14  birth_spain           object
 15  join_channel          object
 16  deceased              object
 17  province_name         object
 18  active_cust           int32
 19  income                float64
 20  segment               object
 21  savings_acct          int32
 22  guarantees            int32
 23  current_acct          int32
 24  derivada_acct         int32
```

```
 25  payroll_acct          int32
 26  junior_acct           int32
 27  mas_particular_acct   int32
 28  particular_acct       int32
 29  particular_plus_acct  int32
 30  short_term_depo       int32
 31  medium_term_depo      int32
 32  long_term_depo        int32
 33  e_acct                int32
 34  funds                 int32
 35  mortgage              int32
 36  pension               int32
 37  loans                 int32
 38  taxes                 int32
 39  credit_card           int32
 40  securities            int32
 41  home_acct             int32
 42  payroll_acct          int32
 43  pensions_2            int32
 44  direct_debt           int32
dtypes: datetime64[ns](2), float64(1), int32(30), object(12)
```

Data description:

col_name	**description**
date	The table is partitioned for this column
customer_code	Customer code
employee_index	Employee index: A active, B ex employed, F filial, N not employee, P passive
country	Customer's Country residence
sex_H	Customer's sex. 1 for "H", 0 for "V"

age	Age
first_contract_date	The date in which the customer became as the first holder of a contract in the bank
new_cust	New customer Index. 1 if the customer registered in the last 6 months.
seniority_in_months	Customer seniority (in months)
primary_cust	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
last_date_primary	Last date as primary customer (if he isn't at the end of the month)
cust_type	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
cust_relationship	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
residency_spain	Residence index (1 (Yes) or 0 (No) if the residence country is the same than the bank country)
birth_spain	Foreigner index (1 (Yes) or 0 (No) if the customer's birth country is different than the bank country)
employee_spouse	Spouse index. 1 if the customer is spouse of an employee

join_channel	Channel used by the customer to join
deceased	Deceased index. 1 if YES, 0 if NO
address_type	Addres type. 1, primary address
province_code	Province code (customer's address)
province_name	Province name
active_cust	Activity index (1, active customer; 0, inactive customer)
income	Gross income of the household
segment	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated
savings_acct	Saving Account
guarantees	Guarantees
current_acct	Current Accounts
derivada_acct	Derivada Account
payroll_acct	Payroll Account
junior_acct	Junior Account
mas_particular_acct	Más particular Account

particular_acct	particular Account
particular_plus_acct	particular Plus Account
short_term_depo	Short-term deposits
medium_term_depo	Medium-term deposits
long_term_depo	Long-term deposits
e_acct	e-account
funds	Funds
mortgage	Mortgage
pension	Pensions
loans	Loans
taxes	Taxes
credit_card	Credit Card
securities	Securities
home_acct	Home Account
payroll_acct	Payroll

pensions_2	Pensions
direct_debt	Direct Debit