# **Santander Dataset**

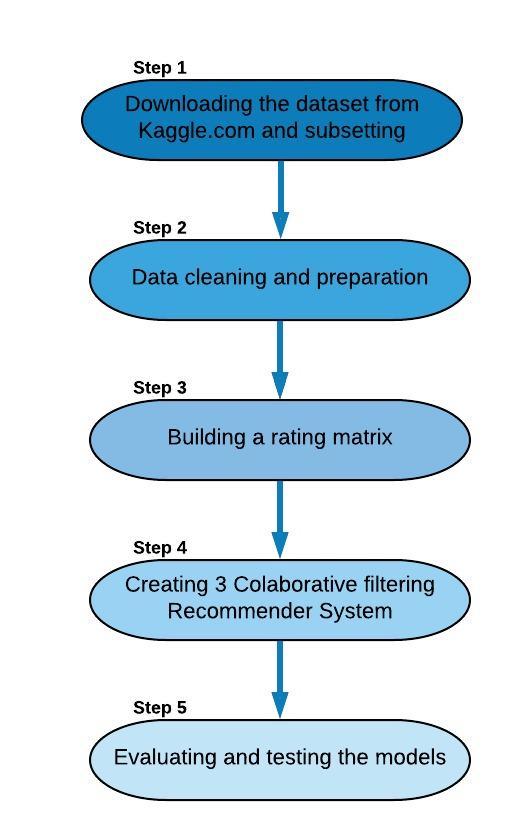
The dataset provided by a Spanish bank contains the financial information of approximatively 1 million customers. It contains 13.647.309 observations from 28 of January 2015 until 28 of May 2016, which include the personal information of the customers and the bank products that they bought.

There are 48 attributes, from which the first 24 are client information, and the rest are the bank products.

Table 1. “Description of the attributes”[[1]](#footnote-1)

|  |  |  |
| --- | --- | --- |
| Attributes | Attributes in  English | * Description |
| * Fecha\_dato | Date | * The date of the transaction |
| * ncodpers | Id | * Unique code that identifies the customer |
| * ind\_empleado | Emp\_status | * Employee status: A(active), B(past employee), F(filial), N(Not employee), P(Pasive) |
| * pais\_residencia | Country | * Country of residence |
| * sexo | Sex | * sex |
| * age | Age | * Age |
| * fecha\_alta | Date2 | * The customer is first holder of a contract in the bank at this date |
| * ind\_nuevo | New\_Customer | * New customer:1(less than6 months) |
| * antiguedad | Seniority | * For how long the customer has been with the bank |
| * indrel | Primary\_customer | * 1 (Primary customer), * 99 (Primary customer during the month but not at end of the month) |
| * “ult\_fec\_cli\_1t | Pr\_customer\_ld | * Last date as primary customer (if he isn't at the end of the month) |
| * indrel\_1mes | Customer\_type | * Type of customer (at the start of the month), * 1(Primary customer) * 2 (co-owner ) * P (Potential), * 3 (former primary), * 4(former co-owner) |
| * tiprel\_1mes | Customer\_st\_end\_m | * Customer status (beginning of the month):   A (active), I (inactive), P (ex customer),R (Potential) |
| * indresi | Residency | * Resident (S (Yes) or N (No) in case the client’s residence and the bank are in the same country |
| * indext | Foreigner | * Foreigner (S (Yes) or N (No) (the client was born in a different country) |
| * conyuemp | emp\_spouse | * Spouse of employee: 1 (the client is an employee’s spouse) |
| * canal\_entrada | Entry\_chanal | * Modes of entry |
| * indfall | Deceased | * Deceased N(No)/S(Yes) |
| * tipodom | Address\_type | * Type of address: 1(primary address) |
| * cod\_prov | Province\_code | * Code of Customer’s Province |
| * nomprov | province\_name | * Nameof the province |
| * ind\_actividad\_cliente | Activity\_st | * Customer status:1( active); 0(inactive) |
| * renta | Household | * Household income |
| * segmento | Segment | * segment : 01(VIP), 02(Individuals), 03(college graduated) |
| * ind\_ahor\_fin\_ult1 | Savings | * Saving Account |
| * ind\_aval\_fin\_ult1 | Guarantees | * Guarantees |
| * ind\_cco\_fin\_ult1 | Current\_Acc | * Current Accounts |
| * ind\_cder\_fin\_ult1 | Derivada | * “Derivada” Account |
| * ind\_cno\_fin\_ult1 | Payroll\_Acc | * Payroll Account |
| * ind\_ctju\_fin\_ult1 | Junior\_Acc | * Junior Account |
| * ind\_ctma\_fin\_ult1 | M\_particular\_Acc | * More particular Account |
| * ind\_ctop\_fin\_ult1 | Particular\_Acc | * Particular Account |
| * ind\_ctpp\_fin\_ult1 | Particular\_Plus\_Acc | * Particular Plus Account |
| * ind\_deco\_fin\_ult1 | Short\_term\_dep | * Short-term deposits |
| * ind\_deme\_fin\_ult1 | Medium\_term\_dep | * Medium-term deposits |
| * ind\_dela\_fin\_ult1 | Long\_term\_dep | * Long-term deposits |
| * ind\_ecue\_fin\_ult1 | e-account | * e-account |
| * ind\_fond\_fin\_ult1 | Funds | * Funds |
| * ind\_hip\_fin\_ult1 | Mortgage | * Mortgage |
| * ind\_plan\_fin\_ult1 | Pensions\_Acc | * Pensions |
| * ind\_pres\_fin\_ult1 | Loans | * Loans |
| * ind\_reca\_fin\_ult1 | Taxes | * Taxes |
| * ind\_tjcr\_fin\_ult1 | Credit\_Card | * Credit Card |
| * ind\_valo\_fin\_ult1 | Securities | * Securities |
| * ind\_viv\_fin\_ult1 | Home\_Acc | * Home Account |
| * ind\_nomina\_ult1 | Payroll | * Payroll |
| * ind\_nom\_pens\_ult1 | Pensions | * Pensions |
| * ind\_recibo\_ult1 | Direct\_Debit | * Direct Debit |

# **Approach**

****

*Figure 1.* Steps of the analysis

## **Downloading the dataset and subsetting**

The provided database available on Kaggle.com (<https://www.kaggle.com/c/santander-product-recommendation>) is quite large, which makes it difficult and time consuming to process.

For this research only a portion of the dataset will be analyzed. Therefore, the information from October 2015- December 2015 interval will be used as are the last three months of the financial year.

In the initial phase the chosen months are extracted from the initial database, obtaining a dataset of 2.710.381 observation of 48 variables regarding 915.898 unique customers and their monthly products.

1. (2017).*Santander product recommendation. Data description*. Retrieved from <https://www.kaggle.com/c/santander-product-recommendation/data> [↑](#footnote-ref-1)