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]

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

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

• 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

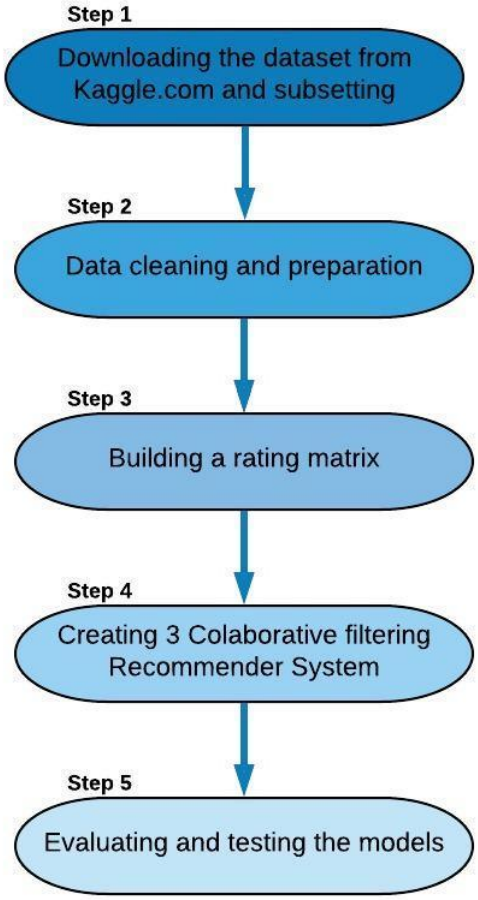

Figure 1. Steps of the analysis

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.