## **Santander product Recommendation**

**1. Data**

* **Problem Statement**: ***“****Built a model to predict which products existing customers will use in the following month based on past behavior”*
* **Steps:**

1. Data visualization and cleaning are done in first step. Missing values are imputed as and when required. Imputations to be done by seeing distribution of the features or clustering the dataset and impute the average of the missing feature for each cluster.

2. To come up with a method of creating artificial features to improve the predictive capabilities.

3. To use XgBoost as a go-to model for initial modeling. If it does not perform well then try out different methods/ ensembles.

**Data Source:**

**Kaggle Data Set Source:**  <https://www.kaggle.com/c/santander-product-recommendation/data>

Below is a summary of the dataset size & my compute.

|  |  |
| --- | --- |
| **Dataset** | **Size** |
| Training | 2.3 GB |
| Testing | 107 MB |

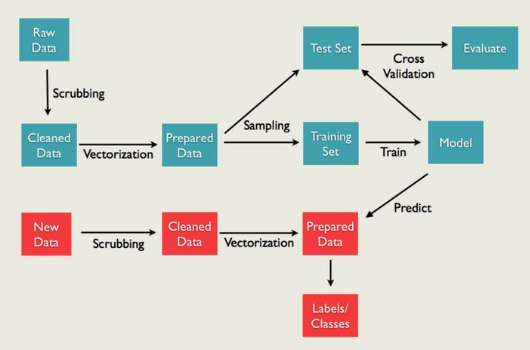
|  |  |
| --- | --- |
| **OS** | **RAM** |
| windows | 8 GB |

**Data Format:**

| **Column Name** | **Description** |
| --- | --- |
| fecha\_dato | The table is partitioned for this column |
| ncodpers | Customer code |
| ind\_empleado | Employee index: A active, B ex employed, F filial, N not employee, P pasive |
| pais\_residencia | Customer's Country residence |
| sexo | Customer's sex |
| age | Age |
| fecha\_alta | The date in which the customer became as the first holder of a contract in the bank |
| ind\_nuevo | New customer Index. 1 if the customer registered in the last 6 months. |
| antiguedad | Customer seniority (in months) |
| indrel | 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month) |
| ult\_fec\_cli\_1t | Last date as primary customer (if he isn't at the end of the month) |
| indrel\_1mes | Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner ),P (Potential),3 (former primary), 4(former co-owner) |
| tiprel\_1mes | Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential) |
| indresi | Residence index (S (Yes) or N (No) if the residence country is the same than the bank country) |
| indext | Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country) |
| conyuemp | Spouse index. 1 if the customer is spouse of an employee |
| canal\_entrada | channel used by the customer to join |
| indfall | Deceased index. N/S |
| tipodom | Addres type. 1, primary address |
| cod\_prov | Province code (customer's address) |
| nomprov | Province name |
| ind\_actividad\_cliente | Activity index (1, active customer; 0, inactive customer) |
| renta | Gross income of the household |
| segmento | segmentation: 01 - VIP, 02 - Individuals 03 - college graduated |
| ind\_ahor\_fin\_ult1 | Saving Account |
| ind\_aval\_fin\_ult1 | Guarantees |
| ind\_cco\_fin\_ult1 | Current Accounts |
| ind\_cder\_fin\_ult1 | Derivada Account |
| ind\_cno\_fin\_ult1 | Payroll Account |
| ind\_ctju\_fin\_ult1 | Junior Account |
| ind\_ctma\_fin\_ult1 | Más particular Account |
| ind\_ctop\_fin\_ult1 | particular Account |
| ind\_ctpp\_fin\_ult1 | particular Plus Account |
| ind\_deco\_fin\_ult1 | Short-term deposits |
| ind\_deme\_fin\_ult1 | Medium-term deposits |
| ind\_dela\_fin\_ult1 | Long-term deposits |
| ind\_ecue\_fin\_ult1 | e-account |
| ind\_fond\_fin\_ult1 | Funds |
| ind\_hip\_fin\_ult1 | Mortgage |
| ind\_plan\_fin\_ult1 | Pensions |
| ind\_pres\_fin\_ult1 | Loans |
| ind\_reca\_fin\_ult1 | Taxes |
| ind\_tjcr\_fin\_ult1 | Credit Card |
| ind\_valo\_fin\_ult1 | Securities |
| ind\_viv\_fin\_ult1 | Home Account |
| ind\_nomina\_ult1 | Payroll |
| ind\_nom\_pens\_ult1 | Pensions |
| ind\_recibo\_ult1 | Direct Debit |

The variables in the format ind\_\*\_fin\_ult1 are the response variables nothing but the products that are offering by the bank. As the target variable, has more than two output values we need to use the multinomial distribution. I was successful in loading 7 million data using python and it is taking less than 5 minutes for loading data.

**Programming Language, packages:**

We would be using the concept of additive modeling for building the predictive model. We would be using Python as the programing language and below is the summary of the package list we are planning for this project. We are using the XGBoost algorithm for modeling. So, I have downloaded the XGBoost package as well.

|  |  |
| --- | --- |
| **Package** | **Purpose** |
| Pandas | Data Frames |
| Numpy | Data manipulation |
| Bokeh | Interactive visualization |
| Sklearn | Machine Learning |
| XgBoost | Machine Learning |
| Keras | Deep Learning |
| Tensorflow | Deep Learning |

## 2. **Preliminary Experiments**

**Data Preprocess Steps:**

1. Converting the attributes to appropriate type. The ‘fecha\_dato’ is the date type variable. So, converting this variable to date type using ‘DatetimeIndex’ function in ‘pandas’ package. The age is not numeric, so converted that variable to numeric using the ‘to\_numeric’ function in pandas
2. We have NA’s for most of the attributes. So, replacing all the NA’s values by observing the distribution of the attribute. For continuous variables like ‘age’ and ‘ind\_actividad\_cliente’ We have replaced the NA’s with median value.
3. For discrete variables, We have replaced the NA values with highest distributed class in the data. For example, npcoders is the customer code. Grouping the customers using the code and replacing the values with highest distributed class for attributes like ‘ind\_nuevo’, ‘indresi’, and ‘indext’.
4. Some of the values for ‘indrel\_1mes’ attribute are integers and some are strings. So, converted all the values into string as this is a discrete variable. Using the dictionary object feature in python We have mapped all the values correctly
5. 'ind\_empleado','pais\_residencia','sexo','canal\_entrada','conyuemp' replaced the NA values with ‘UnKnown’ because values are equally distributed for all the classes

There are still some methods we would try. For example, removing the outliers using 2 sigma deviation i.e giving an Upper and lower control limit for each feature. We would try out these methods if the initial results need some convincing. For the date values, for now we would be using the variable for calculating lags but would try to come up with a different method.