Santander Product Recommendation Case Study

Kyu Cho

Agenda

01 – Introduction

02 – Preprocessing

03 – Exploratory Data Analysis

04 – Feature Engineering

05 – Modeling and Results

06 – Results and Finding

07 – Difficulties and Future Steps

---

01 – Introduction

Project Description

Santander Bank offers their customers personalized product recommendations time to time, in order to meet the individual's needs and satisfaction. This case study seeks to improve the recommendation system by predicting which products their existing customers will use in the next month based on their past behaviors.

Data

Source : https://www.kaggle.com/c/santander-product-recommendation/data

- Training set: 13,647,409 rows
- Test set: 929,615 rows
- Categorical: 21
- Continuous: 3
- Customer information index: 1 to 24
- Date: 2015.1 – 2016.5
- Product index: 25 to 48

- Evaluation: Multi-Classifier Recommended Products of 7

---

02 – Preprocessing

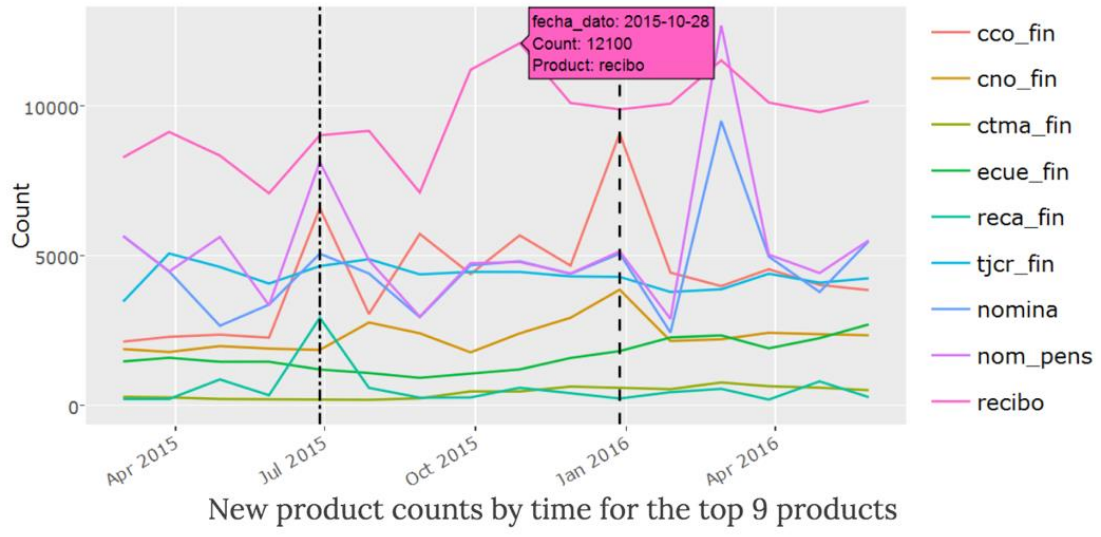
Missing Value Imputation

- Contain Missing Values: 24 features
- ‘indrel_1mes’ : NAs were fill with 0 and non-integer value such as “P” was replaced with 5
- Numeric features such as ‘antiguedad’, ‘ind_nom_pens_ult1’, ‘ind_nomina_ult1’ , NAs were fill with 0
- String categorical features such as ‘segmento’, ‘'pais_residencia', 'canal_entrada', 'nomprov', 'cod_prov', NAs were fill with undefined
    - I found out that instead of imputing the missing value with Mean, Median or Mod, make them as separate column increases the performance of the model
- Features with more than 95 percent of missing values or too high cardinalities, had dropped totally such as 'renta', 'conyuemp', 'tipodom', 'cod_prov', 'fecha_alta'
- Product that has been discontinued or significantly unpopular such as 'ahor', 'aval', 'deco', 'deme’, 'hip', 'pres', 'cder’ were removed from the recommendation list

---

03 – Exploratory Data Analysis

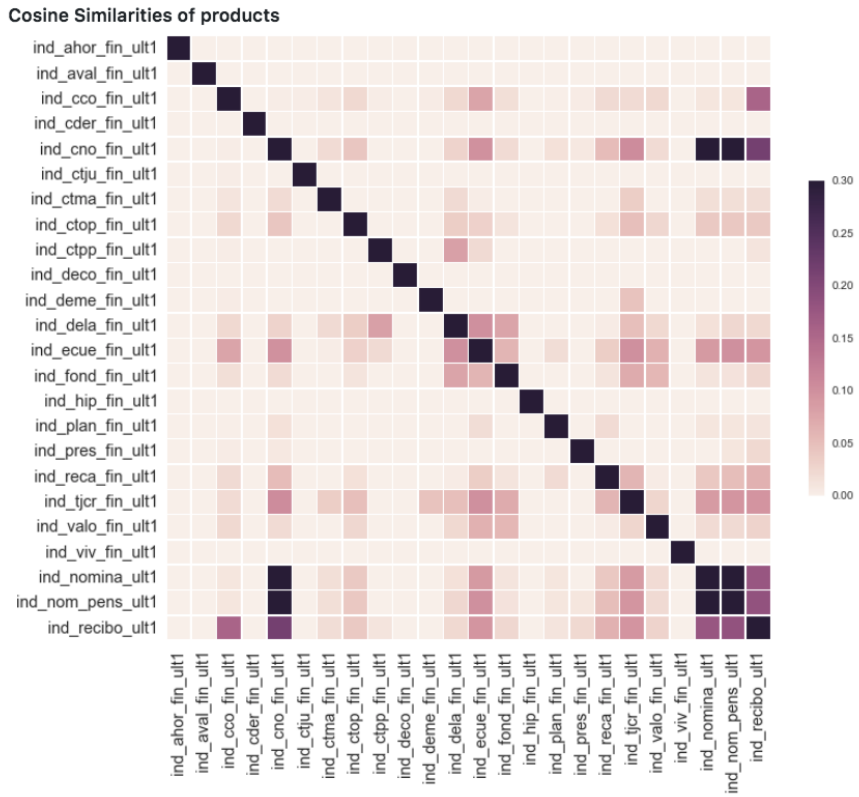
3-1 Seasonal plot of the top products

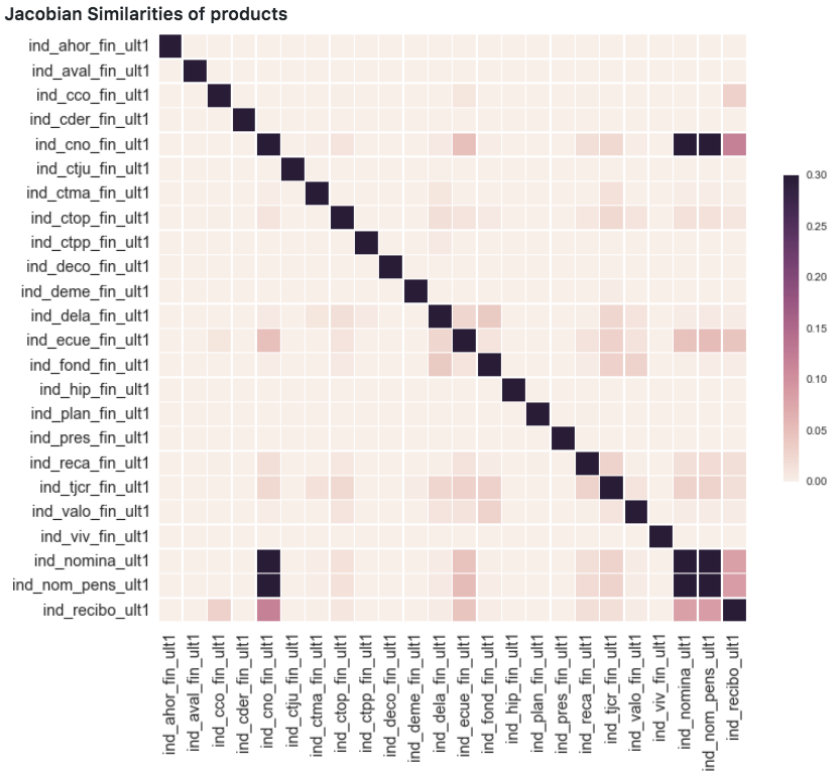


New product counts by time for the top 9 products

- The popularity of the specific product exists along with the seasonal trend
- Since product counts in July 2015 shows the multiple correlation of the products, June 2015 will be extremely helpful feature/indicator to predict new product in June 2016
- Correlation with multiple products exists, let's look at it from the different perspective

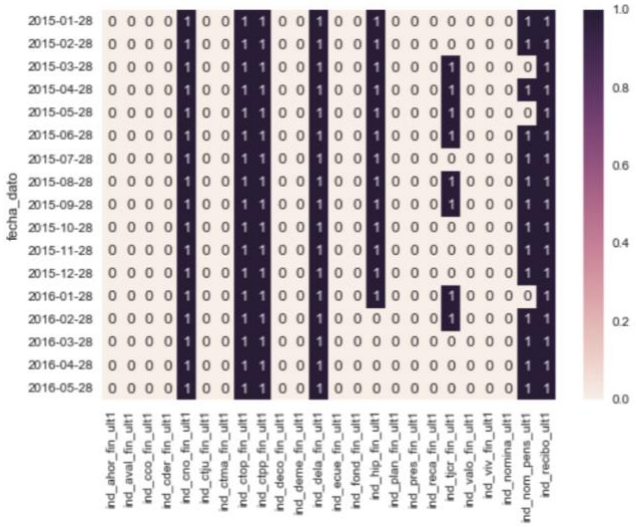
3-2 Correlation plot of all products

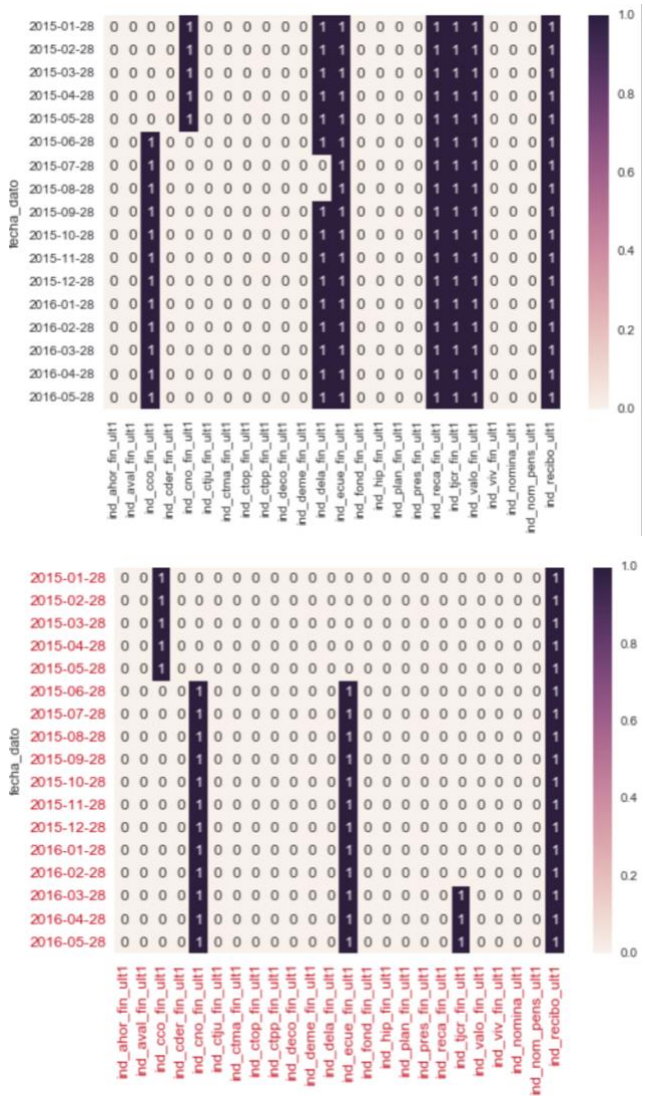




- ‘cno’ with ‘nom_pens’ and ‘nomina’ are extreamly correlated
- ‘nomina’ with ‘nom_pends’ and ‘recibo’ are highly correlated
- ‘nom_pends’ with ‘recibo’ are highly correlated

3-3 Historical plots for the ownerships of the product over time





- Customers are dynamically adding new products and dropping old products

[Take away]

- Build new features that can capture the pattern, trend throughout the season and the characteristics of the months.
- Build new features that can capture the dynamic movement of the client's behavior such as dropping or adding new products

04 – Feature Engineering

Based on the EDA above and the intuitions, I've created following features.

Product related

- list = lag of all products
- int = flag if the product was added, dropped or unchanged from previous month (switching from 0 to 1, 1 to 0, 0 to 0, 1 to 1)
- int = consecutive months that customer not had product X (length of continuous 0 for each product in last 5 months)
- int = consecutive months that customer had product X (length of continuous 1 for each product in last 5 months)
- float = ratio of months that customer had owned product X
- flaot = ratio of months that customer had changed the product X

Customer related

- int = flag change of customer features from previous month such as 'segmento', 'ind_actividad_cliente', cod_prov, 'canal_entrada', 'indrel_1mes', 'tiprel_1mes'

Date related

- int = lag of past 5 months

Simple Example of Lagged Features

Original

1/28/15
2/28/15
3/28/15
4/28/15
5/28/15
6/28/15
7/28/15
8/28/15
9/28/15

10/28/15

Lagged 5 Features

1/28/15	2/28/15	3/28/15	4/28/15	5/28/15	6/28/15
2/28/15	3/28/15	4/28/15	5/28/15	6/28/15	7/28/15
3/28/15	4/28/15	5/28/15	6/28/15	7/28/15	8/28/15
4/28/15	5/28/15	6/28/15	7/28/15	8/28/15	9/28/15
5/28/15	6/28/15	7/28/15	8/28/15	9/28/15	10/28/15

---

05 – Model and Results

Model	**Data**	**Private Score**	**Public Score**
Randomforest	Raw	0.0226121	0.0224749
XGB	Raw	0.0272445	0.0268538
Randomforest	Feature Engineered	0.0292582	0.0288355

---

06 – Results and Finding

1. Building lagged features helped to capture both to seasonal component and general trend for these products.
2. Different months have different patterns and trends, having multiple windows on 5 lagged helped to capture those.
3. The algorithm uses the lagged months features to figure out which month the data came from – and it would predict probabilities in line with distributions from that month.
4. Specific age distribution of users that buy ‘nom_pens’ in a certain month but don't buy ‘nomina’.
5. Based on the 'renda' feature, multiple customers belonged to the same household tends to have same behavior pattern of adding or dropping products.
6. Customers had had a product before, they were more likely to have it again.

7. During March, May and Jan, the number of the product ownership dropped significantly which introduce high probabilities of adding 'new' product at next month.
8. The new customer in July 2016 which is about 170k accounts, have different behaviors (more active) comparing to the other group.

---

07 – Difficulties and Future Steps

Difficulties

- Data was much dirtier than expected, preprocessing the data without losing information were challenging.
- Due to time limitation, I was focused mainly around feature engineering which has the most impact on the scores rather than advanced feature selection, hyper parameters tuning, and ensemble models.

Future Improvement

- Adding more features such as
  - int = number of products owned in previous months
  - list = exponential weighted average of each product
  - float = average of products group by ‘canal_entrada’, ‘segmento’, ‘cod_prov’, ‘age_group’
  - string = concatenation of 'tiprel_1mes', 'ind_actividad_cliente' and the products
  - int = time since change of customer features such as
    - ‘segmento’, ‘ind_actividad_cliente’, ‘cod_prov, canal_entrada’, ‘indrel_1mes’, ‘tiprel_1mes’
  - categorical = age group by percental
  - float = ratio of income mean group by ‘canal_entrada’, ‘segmento’, ‘cod_prov’, ‘age_group’
- Advanced Feature Selection
- Hyper-parameter tuning
- Ensemble models
- Post-processing to calibrate the mode with the scores on the leaderboard

---