Santander Bank Product Recommendation

Yisha He(yh2uq), Yanan Gong(yg3ca)
Yinyu Cheng(yc5cb), Shulei Yang(sy8uu)

Introduction of Santander

- Santander Bank is a Spanish financial institution that offers financial services and products to customers worldwide.
- Their main products and services are: savings, mortgages, corporate banking, cash management, credit card etc.
- This project is trying to build a product recommendation system for Santander Bank to predict what products a customer will buy on 2016-06-28 in addition to what they already had on 2016-05-28.

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

Data Summary

No	Variable	Description
1	fecha_dato	The table is partitioned for this column
2	ncodpers	Customer code
3	ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive
4	pais_residencia	Customer's Country residence
5	sexo	Customer's sex
6	age	Age
7	fecha_alta	The date in which the customer became as the first holder of a contract in the bank
8	ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.
9	antiguedad	Customer seniority (in months)
10	indrel	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
11	ult_fec_cli_1t	Last date as primary customer (if he isn't at the end of the month)
12	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)
13	tiprel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
14	indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
15	indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
16	conyuemp	Spouse index. 1 if the customer is spouse of an employee
17	canal_entrada	channel used by the customer to join
18	indfall	Deceased index. N/S
19	tipodom	Addres type. 1, primary address
20	cod_prov	Province code (customer's address)
21	nomprov	Province name
22	ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
23	renta	Gross income of the household
24	segmento	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated

Training set: Monthly observations for 956,000 customers from January 2015 to May 2016.

Testing set: 929,000 customers on June 2016.

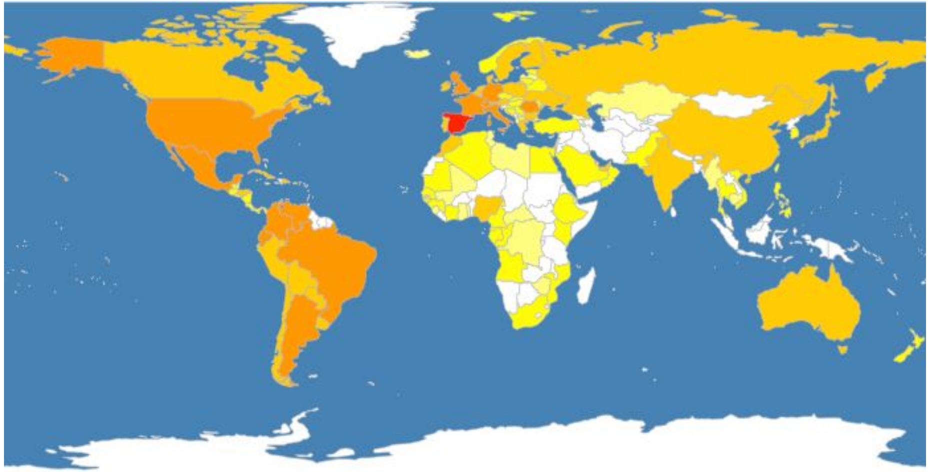
Variables related to customer characterics.

Variables related to Santander’s existing products.

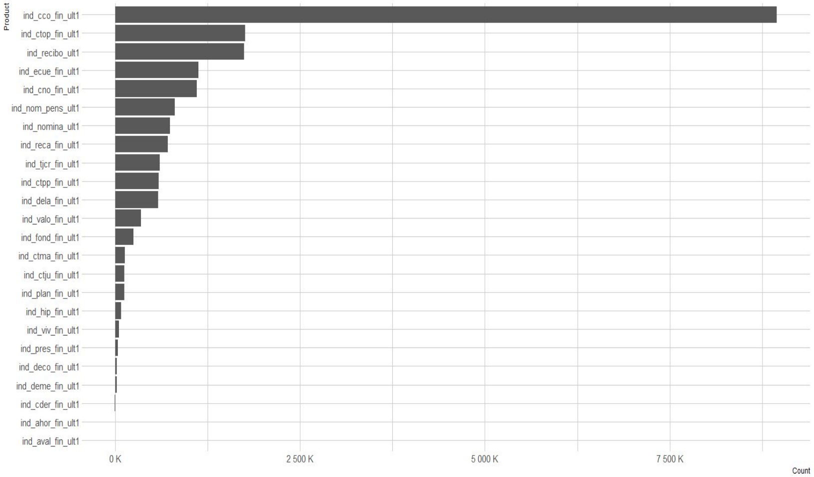
No	Variable	Description
25	ind_ahor_fin_ult1	Saving Account
26	ind_aval_fin_ult1	Guarantees
27	ind_cco_fin_ult1	Current Accounts
28	ind_cder_fin_ult1	Derivada Account
29	ind_cno_fin_ult1	Payroll Account
30	ind_ctju_fin_ult1	Junior Account
31	ind_ctma_fin_ult1	Más particular Account
32	ind_ctop_fin_ult1	particular Account
33	ind_ctpp_fin_ult1	particular Plus Account
34	ind_deco_fin_ult1	Short-term deposits
35	ind_deme_fin_ult1	Medium-term deposits
36	ind_dela_fin_ult1	Long-term deposits
37	ind_ecue_fin_ult1	e-account
38	ind_fond_fin_ult1	Funds
39	ind_hip_fin_ult1	Mortgage
40	ind_plan_fin_ult1	Pensions
41	ind_pres_fin_ult1	Loans
42	ind_reca_fin_ult1	Taxes
43	ind_tjcr_fin_ult1	Credit Card
44	ind_valo_fin_ult1	Securities
45	ind_viv_fin_ult1	Home Account
46	ind_nomina_ult1	Payroll
47	ind_nom_pens_ult1	Pensions
48	ind_recibo_ult1	Direct Debit

Visualization

- Santander Customers Geograpgic Distribution: Most customers are from Spain.

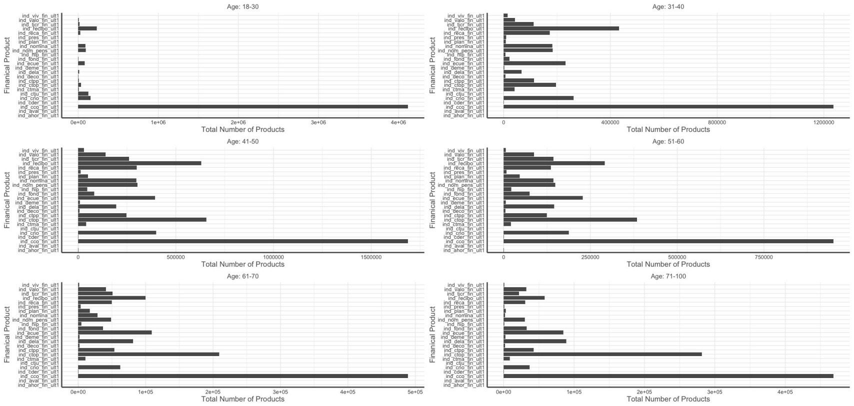


- Santander's Past Product Purchase Distribution: Most bought products are: current account, particular account, direct debit, e-account.



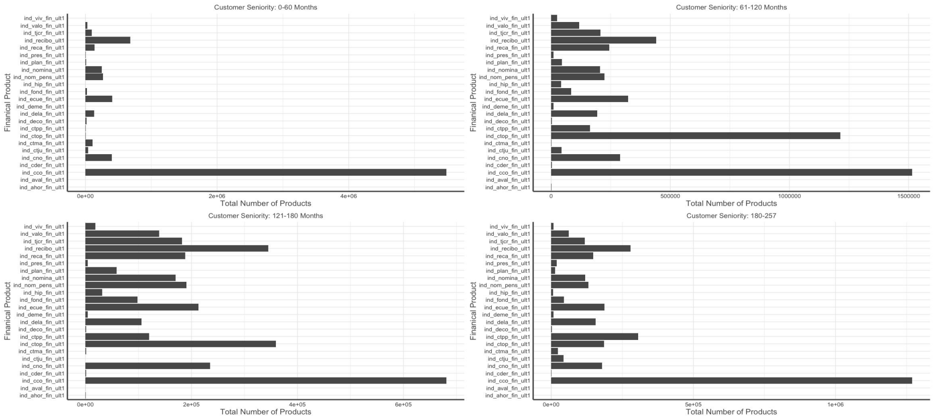
Age’s and Customer Seniority’s Influence

Age



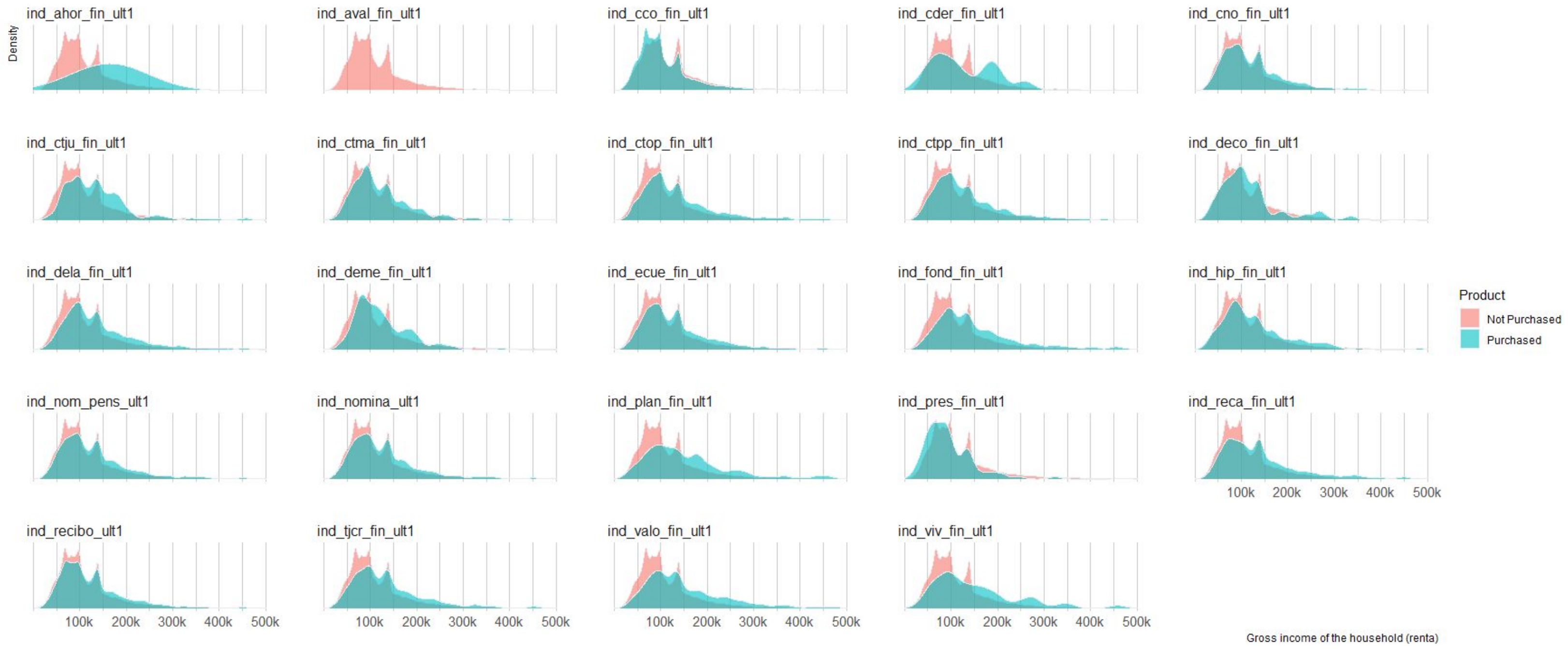
- Shows customers' purchase preferences within different age intervals.

Customer Seniority

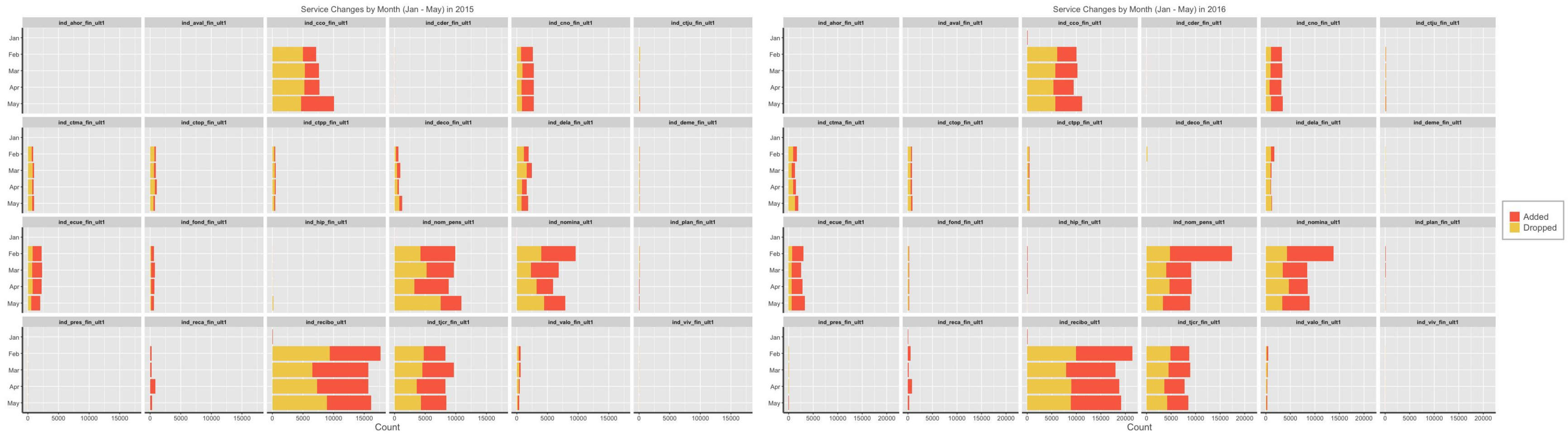


- Shows how customer’s seniority influence their purchase preference.

Customer Household Income and Purchase Preference



Service Changes by Month



- Together demonstrate the service changes from January to May for 2015 and 2016.
- The service change patterns in each month are very similar from 2015 to 2016. Therefore, we suspect that the the services change in June, 2016 might be similar to the services change in June 2015.

Models : XGBoost

- A scalable machine learning system for tree boosting.
- Innovative ideas :
    - Regularized learning objective : loss function + penalty term.

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

    - Second-order approximation of loss : Taylor expansion.
- Splitting finding:
    - Exact greedy algorithm.
    - Approximate algorithm.
    - Sparsity-aware split finding algorithm.

Analysis : XGBoost

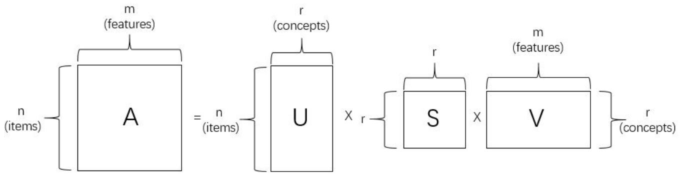
- Dealing with predictors : select 17 of 22 predictors.
- Dealing with responses :
  - For each customer, record the indexes of products newly bought on 2015-06-28 compared to 2015-05-28.
  - Combine each index with the 17 predictors of this customer as a new sample. If a customer buys 7 new products, there will be 7 new samples for him/her.
- New training set : combine all the samples generated above together.
  - A record of customers who buy new products on 2015-06-28.
  - New predictors : 17 predictors.
  - New response : index of product newly bought.

Analysis : Tuning Paremeters

Paremeter	Chosen value	Description
objective	multi:softprob	setting XGBoost to conduct multiclass classification
est	0.05	step size shrinkage used in update to prevents overfitting
max_depth	6	maximum depth of a tree
num_class	24	the number of classes that are going to deal with
eval_metric	mlogloss	evaluation metrics for validation data
min_child_weight	2	minimum sum of instance weight needed in a child
subsample	0.8	subsample ratio of the training instances
colsample_bytree	0.8	subsample ratio of columns when constructing each tree
num_rounds	100	the number of rounds for boosting

Models : SVD

- Singular Value Decomposition is a Matrix Factorization approach.
- Decompose a given matrix A : A = USV'.



- SVD in recommendation system:
  - For a given user u: p_u measures matrix with users and factors.
  - For a given item i: q_i^T measures which the item possesses those factor.
  - $\hat{r}_{ui} = q_i^T p_u$ captures the interaction between user u and item i.

$$Min_{(q,p)} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2.$$

Analysis : SVD

- Create user - item matrix:
  - Keep "ncodpers" and the 24 response, which are users and items in our dataset.
  - Choose 2 timestamp: "2016-04-28" and "2016-05-28" to decrease the computational cost and prevent large sparse matrix.
  - Remove "ncodpers" and make the map between number of rows and "ncodpers".

- Shape of matrix: 931453 rows and 24 columns.
- Find the best $\hat{r}_{ui}$, rank the value of items and get the top 7 products.

Result :

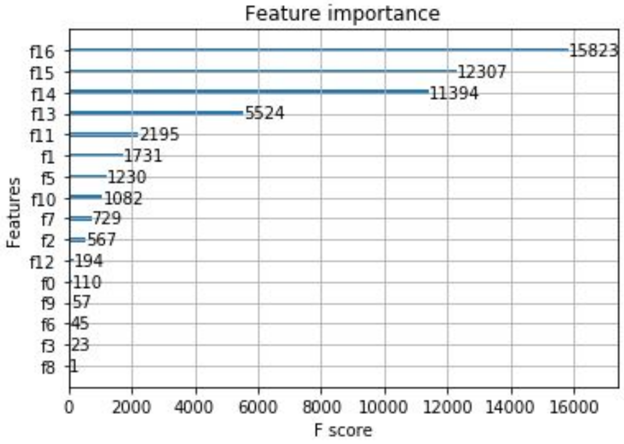
- Evalutation:

$$MAP@7 = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{1}{min(m,7)} \sum_{k=1}^{min(n,7)} P(k)$$

- XGBoost: 0.01821; SVD:0.01915.

- Feature importance selected by XGBoost:

  f16, f15, f14 correspond to "renta", "antiguedad", "age" separately.



Conclusion :

- Both XGBoost and SVD work well in our problem and can successfully predict what products a customer will newly buy on 2016-06-28.

- Strengths:
    - Data visualization: find each variable’s influence on customer’s final purchase preference.
    - XGBoost: prevent overfitting, provide a evaluation for the performance of each split.
    - SVD: simple, uncover latent relations between customers and products.

- Limitations and Future improvement:
    - Do not fully utilize feature information: do more feature exploration.
    - Can not deal with cold start problem: build advanced methods.