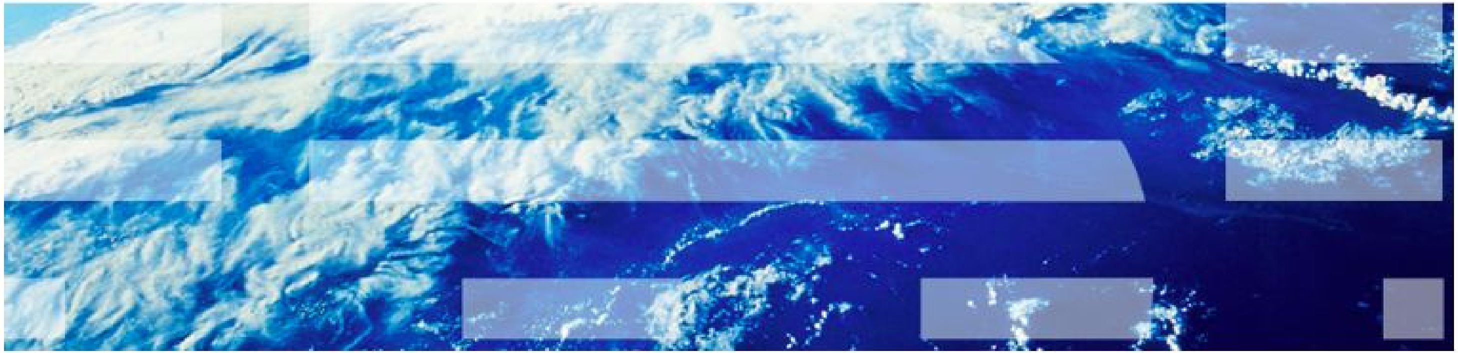
E6893 Big Data Analytics:

Santander Product Recommendation

Team Members:
Huilong An(hz2399)
Wenhang Bao(wb2304)
Yifan Zhang(yz2831)



December 15, 2016

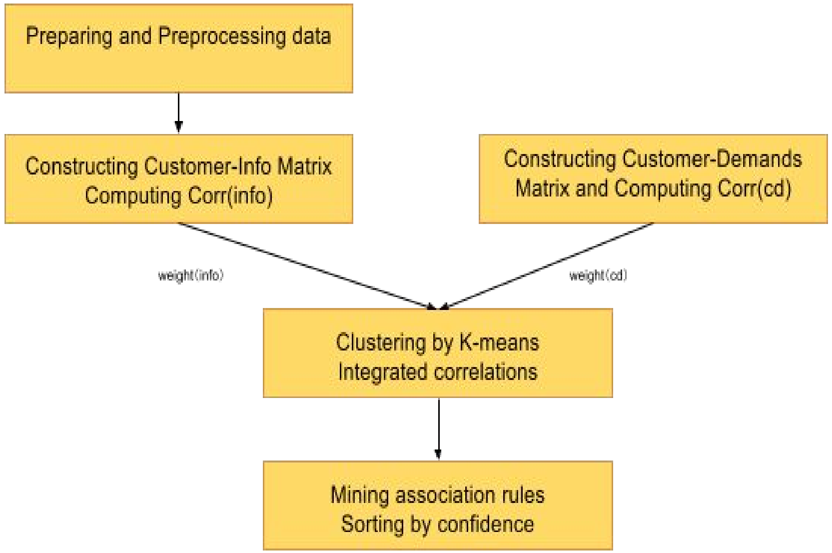
Project Description

- Objective
  - Our project is trying to explore an effective recommendation algorithm to predict which bank product a consumer will be most likely to purchase in the following month based on their past behavior and that of similar customers.
- Data Overview
  - We downloaded our data from the following website and the uncompressed data size is over 2.3GB. https://www.kaggle.com/c/santander-product-recommendation/data
- Technology Used
  - Spark
  - Python
  - R

Outline

Santander

- Data Preparing and Preprocessing
    - Data Cleaning
    - Data Exploration and Visualization
- Customer-Information Matrix
    - Transformation of Data
- Customer-Demands Matrix
    - Applied User-Based Recommendation
- Integrate Correlations
- Clustering by K-means
    - Based on similarity matrix
- Training association rules within each cluster to give recommendations



Data Preprocessing

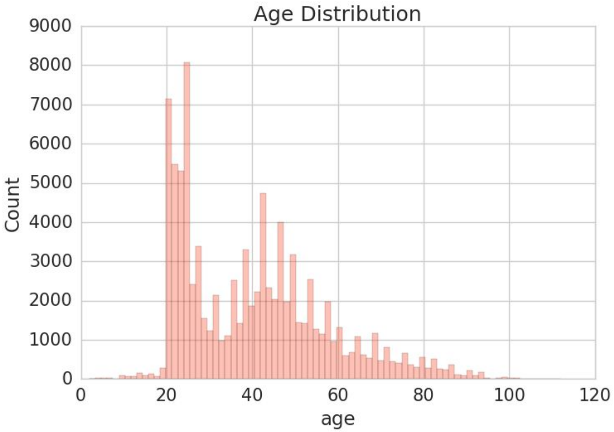
- Data Description
    - Over 10 million users' records with 24 features and 24 labels
    - Feature variables include Age, sex, employment, residence and etc
    - Label variables are all dummy variables to show if the user is currently having the certain product

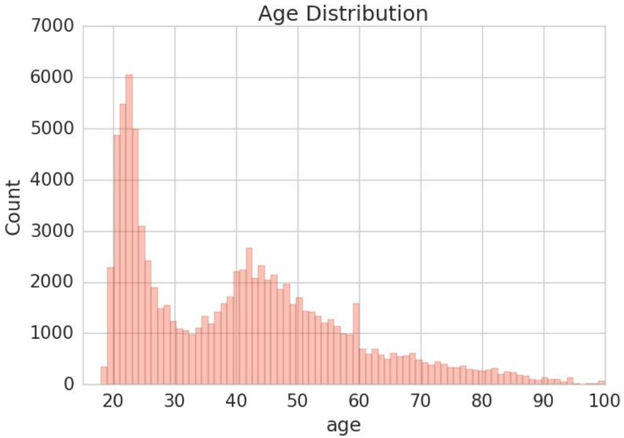
	Features							Labels			
	fecha_dato	ncodpers	ind_empleado	pais_residencia	sexo	age	fecha_alta	ind_nuevo	ind_reca_fin_ult1	ind_tjcr_fin_ult1	ind_valo_fin_ult1
9704	1/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
120754	2/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
186516	3/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
252157	4/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
272539	5/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
338163	6/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
454973	7/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
494807	8/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
636849	9/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
674624	10/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
825030	11/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
863829	12/28/2015	952138	N	ES	H	30	9/30/2011	0	0	0	0
1030664	1/28/2016	952138	N	ES	H	30	9/30/2011	0	0	0	0
1127817	2/28/2016	952138	N	ES	H	30	9/30/2011	0	0	0	0
1191618	3/28/2016	952138	N	ES	H	31	9/30/2011	0	0	0	0
1293000	4/28/2016	952138	N	ES	H	31	9/30/2011	0	0	0	0
1345086	5/28/2016	952138	N	ES	H	31	9/30/2011	0	0	0	0

- Data Cleaning
    - Outliers---Removing and smoothing
    - Missing Data---Filling case by case
    - Empty Strings---Assigning Unknown

Data Exploration and Visualization

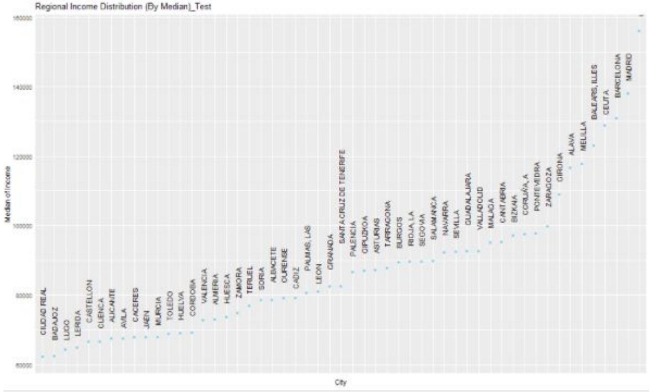
- Age Distribution
  - It's shown that the age distribution is bimodal. There are a large number of university aged students, and then another peak around middle-age.
  - Separate the distribution and move the outliers to the mean of the closest one.





Data Exploration and Visualization

- Filling Missing Values
  - By judging other variables Eg: Ind_nuevo
  - By the more common status
  - By different medians
    - Use Income as an example
    - We can see obvious variations between different provinces’ medians
    - Not reasonable to apply the total median
    - Assigning missing incomes by province instead



Customer-Information Matrix

ncodpers	age	fecha_alta	New customer Index	Customer relation type	Gross income of the household	segmentation
586885	55	12/22/05	0	A	155478.39	02 - PARTICULARES
1136578	44	2006/7/13	0	A	59450.88	02 - PARTICULARES
434773	35	2009/12/3	0	A	155128.05	02 - PARTICULARES
154668	45	12/30/99	0	A	214848.03	02 - PARTICULARES
1008154	44	2003/9/12	0	I	76315.26	02 - PARTICULARES
236319	47	2004/2/1	0	I	188164.53	02 - PARTICULARES
1375581	37	2001/12/15	0	A	55587.81	02 - PARTICULARES
632181	33	2008/7/6	0	A	54351.6	02 - PARTICULARES
1281835	22	7/28/14	0	A	98466.54	03 - UNIVERSITARIO
141511	57	9/1/99	0	A	139093.44	02 - PARTICULARES
586885	55	12/22/05	0	A	155478.39	02 - PARTICULARES
337151	57	4/15/02	0	A	114594.75	02 - PARTICULARES
1105763	40	11/26/12	0	A	88675.24	02 - PARTICULARES
51852	49	12/13/96	0	A	112996.59	02 - PARTICULARES
809170	55	10/26/08	0	A	301241.48	02 - PARTICULARES

Transform data to construct a Customer-Info Matrix

ncodpers	age	fecha_alta	New customer Index	Customer relation type	Gross income of the household	segmentation
586885	0.4375	0.51235121	0	1	0.00633098	1
1136578	0.3	0.861000896	0	1	0.002281703	1
434773	0.1875	0.405862025	0	1	0.006316207	1
154668	0.3125	0.2328171	0	1	0.008834473	1
1008154	0.3	0.802764623	0	0	0.002992838	1
236319	0.3375	0.29156534	0	0	0.007709286	1
1375581	0.2125	0.935748112	0	1	0.002118805	1
632181	0.1625	0.541533342	0	1	0.002066677	1
1281835	0.025	0.914245488	0	1	0.003926911	0
141511	0.4625	0.217458083	0	1	0.005640061	1
586885	0.4375	0.51235121	0	1	0.00633098	1
337151	0.4625	0.339946243	0	1	0.004607003	1
1105763	0.25	0.836298477	0	1	0.003514032	1
51852	0.3625	0.090490209	0	1	0.004539612	1
809170	0.4375	0.645334699	0	1	0.012477503	1

Customer-Demands Matrix

ncodpers	Current Accounts	Payroll Account	particular Account	Taxes	Payroll
586885	1	0	1	0	0
1136578	0	1	0	0	1
434773	0	1	1	1	1
154668	1	0	0	0	0
1008154	1	0	0	1	0
236319	0	0	0	1	0
1375581	1	0	0	0	0
632181	1	0	0	0	0
1281835	1	0	0	0	0
141511	0	1	0	0	1
586885	1	0	1	0	0
337151	1	0	0	0	0
1105763	0	1	0	1	1
51852	1	0	0	0	0
809170	1	0	0	0	0

Correlation Computing and Integration

Santander

Customer-Information Matrix

$$Corr_{\text{info}}(C_i, C_j) = \frac{\sum_{s\in V}(Winfo_{ci,s} - \overline{Winfo_{ci}})(Winfo_{cj,s} - \overline{Winfo_{cj}})}{\sqrt{\sum_{s\in V}(Winfo_{ci,s} - \overline{Winfo_{ci}})^2 (Winfo_{cj,s} - \overline{Winfo_{cj}})^2}}$$

Customer-Demands Matrix

$$corr_p(c_i, c_j) = \frac{\sum_{s\in I}(r_{c_i,s} - \bar{r}_{c_i})(r_{c_j,s} - \bar{r}_{c_j})}{\sqrt{\sum_{s\in I}(r_{c_i,s} - \bar{r}_{c_i})^2 \sum_{s\in I}(r_{c_j,s} - \bar{r}_{c_j})^2}}$$

Integrated Correlation

$$Corr_{integrated}(C_i, C_j) = W_{info} \times Corr_{info}(C_i, C_j) + W_{cd} \times Corr_{cd}(C_i, C_j)$$

Sample Output

ncodpers	added_products
15889	particular Account
15890	e-account
15892	Credit Card
15893	Credit Card
15894	Securities
15895	particular Account
15896	Mortgage
15897	particular Account
15898	e-account
15899	particular Account
15900	Derivada Account
15901	particular Account
15902	Loans
15903	particular Plus Account
15906	e-account

Conclusion:
1.The distribution of added products is significantly different from the distribution of original products in this specific problem.

2.The process of tuning weights of personal preference cannot be perfect.

3.Substituting distance function in K-means is tricky.

Reference

- [1] Toine Bogers, Collaborative and Content-based Filtering for Item Recommendation on Social Bookmarking Websites
- [2] Ya-Yueh Shih, Hybrid recommendation approaches: collaborative filtering via valuable content information
- [3]https://www.kaggle.com/apryor6/santander-product-recommendation/detailed-cleaning-visualization-python
- And many more

- The End
- Thank you!