Creating a Product Recommendation Engine for Large Scale Commercial Bank

Predictive Modeling to Create a Personalized Rank-ordered List of the Seven Products Customers are Most Likely to Purchase

Background: Business Problem

- Santander Bank wants to support customers with a range of financial needs through personalized product recommendations
- Under their current system, a small number of Santander’s customers receive many recommendations while many others rarely see any resulting in an uneven customer experience.
- With a more effective recommendation system in place, Santander can better meet the individual needs of all customers and ensure their satisfaction no matter where they are in life.

Client

- Santander Group is a multinational banking conglomerate
- Its chief holding is Banco Santander the largest bank in Spain
- They are actively looking to improve their business and customer experience with data driven approaches

Data

	fecha_dato	ncodpers	ind_empleado	pais_residencia	sexo	age	fecha_alta	ind_nuevo	antiguedad
0	2015-01-28	1375586	N	ES	H	35	2015-01-12	0.0	6
1	2015-01-28	1050611	N	ES	V	23	2012-08-10	0.0	35
2	2015-01-28	1050612	N	ES	V	23	2012-08-10	0.0	35
3	2015-01-28	1050613	N	ES	H	22	2012-08-10	0.0	35
4	2015-01-28	1050614	N	ES	V	23	2012-08-10	0.0	35

- Anonymized user data was given in the format of one record per user per month
- User product information was 24 binary values for whether they owned each of the 24 products in that given month
- The test set was records for customers in the month of June 2016 with product columns excluded
- The train set included 17 months from January 2015 to May 2016
- There were are also 22 columns offering varied demographic data

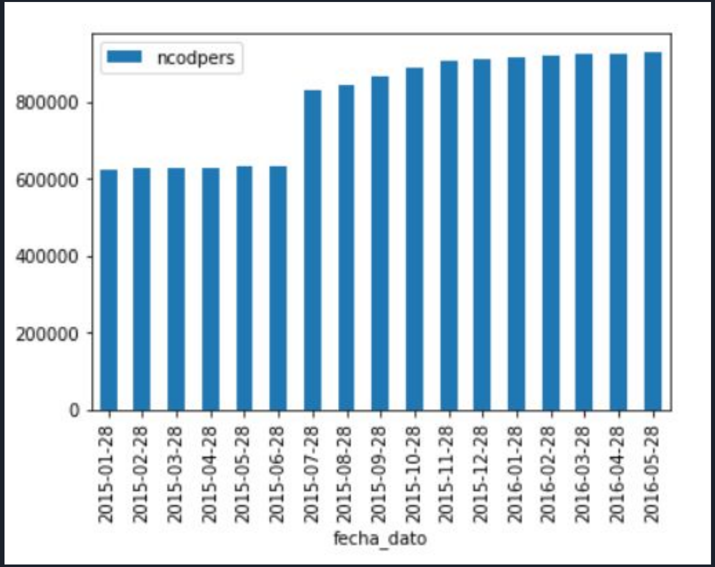
Demographic Columns:

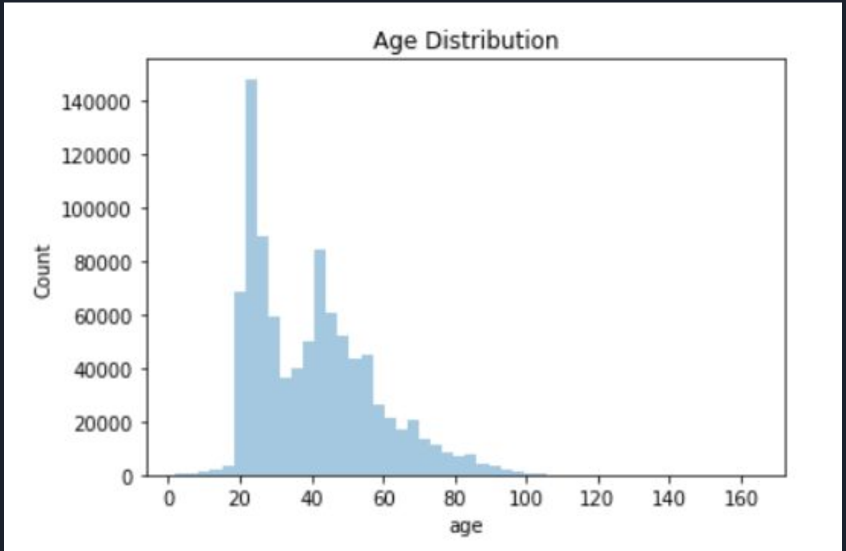
Column Name	Description
fecha_dato	The table is partitioned for this column
ncodpers	Customer code
ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive
pais_residencia	Customer's Country residence
sexo	Customer's sex
age	Age
fecha_alta	The date in which the customer 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)
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_c	Activity index (1, active customer; 0, inactive customer)
renta	Gross income of the household

Product Columns

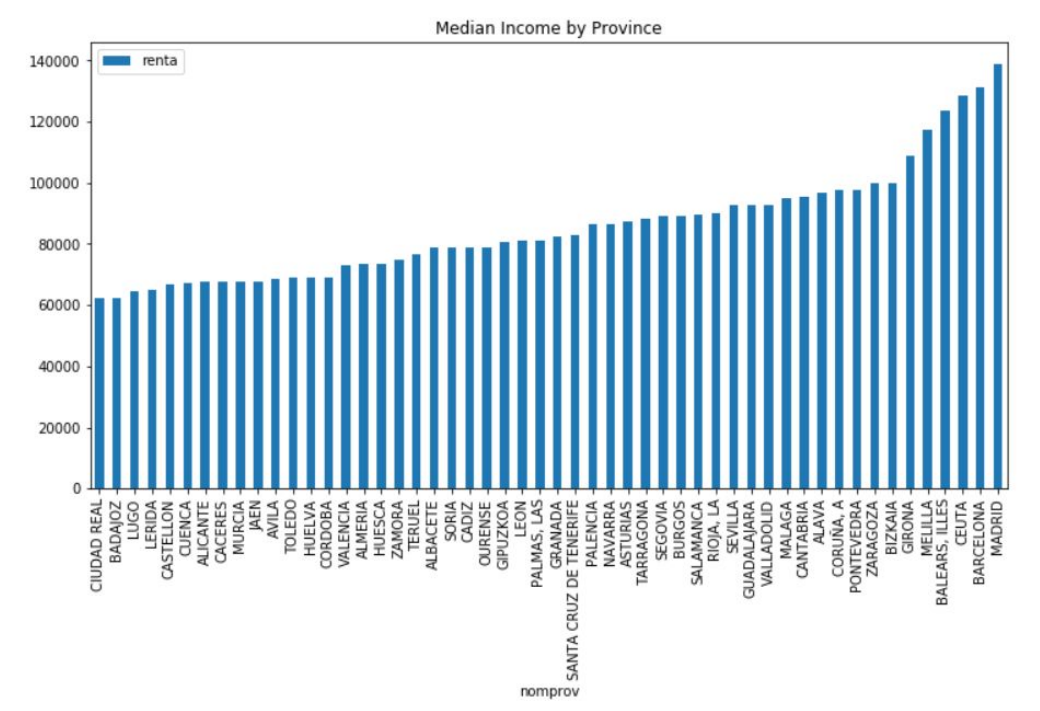
ind_cco_fin_ult1	Current Accounts
ind_cder_fin_ult1	Derivada Account
ind_cno_fin_ult1	Payroll Account
ind_ctju_fin_ult1	Junior Account
ind_ctma_fin_ult1	Más particular Account
ind_ctop_fin_ult1	particular Account
ind_ctpp_fin_ult1	particular Plus Account
ind_deco_fin_ult1	Short-term deposits
ind_deme_fin_ult1	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

Exploratory Data Analysis



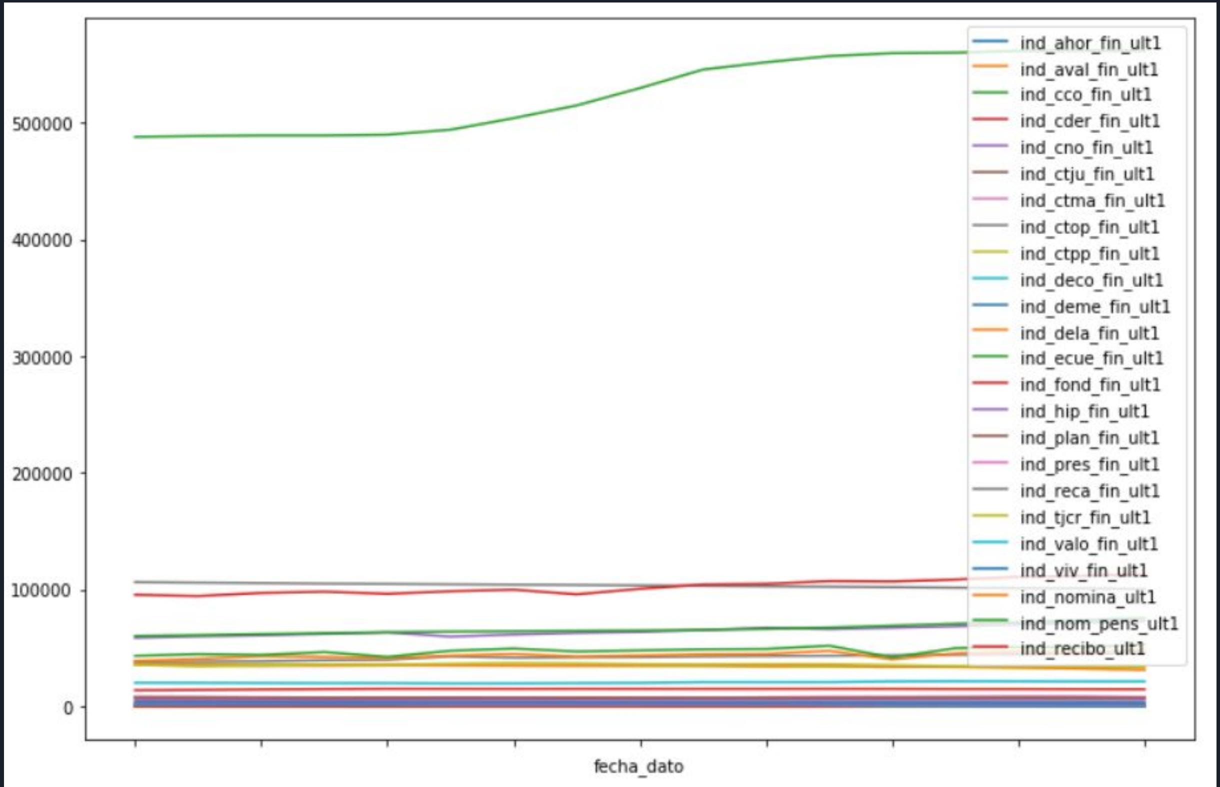


EDA

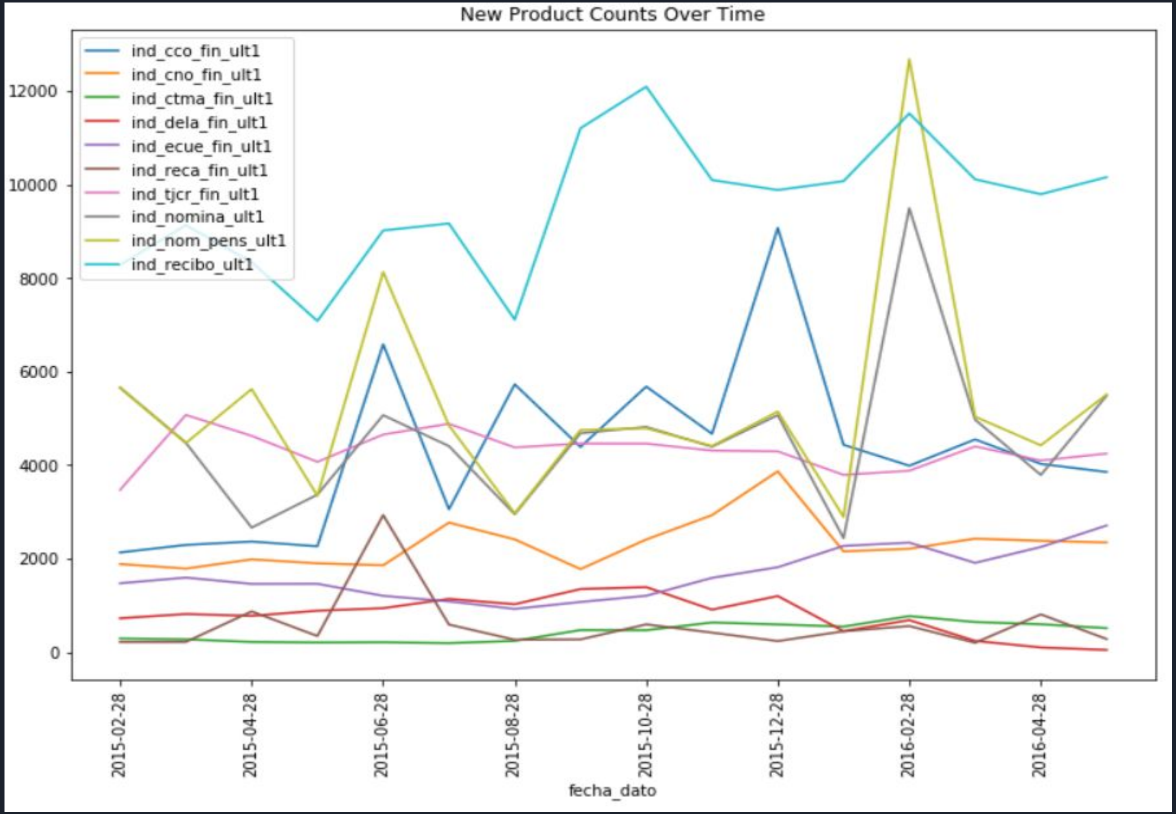


Median Income by Province
renta
140000
120000
100000
80000
60000
40000
20000
0
CIUDAD REAL
BADAJOZ
LUGO
LERIDA
CASTELLON
CUENCA
ALICANTE
CACERES
MURCIA
JAEN
AVILA
TOLEDO
HUELVA
CORDOBA
VALENCIA
ALMERIA
HUESCA
ZAMORA
TERUEL
ALBACETE
SORIA
CADIZ
OURENSE
GIPUZKOA
LEON
PALMAS, LAS
GRANADA
SANTA CRUZ DE TENERIFE
PALENCIA
NAVARRA
ASTURIAS
TARRAGONA
SEGOVIA
BURGOS
SALAMANCA
RIOJA, LA
SEVILLA
GUADALAJARA
VALLADOLID
MALAGA
CANTABRIA
ALAVA
CORUÑA, A
PONTEVEDRA
ZARAGOZA
BIZKAIA
GIRONA
MELILLA
BALEARS, ILLES
CEUTA
BARCELONA
MADRID
nomprov

Total Product Counts Over Time



Newly Purchased Product Counts Over Time



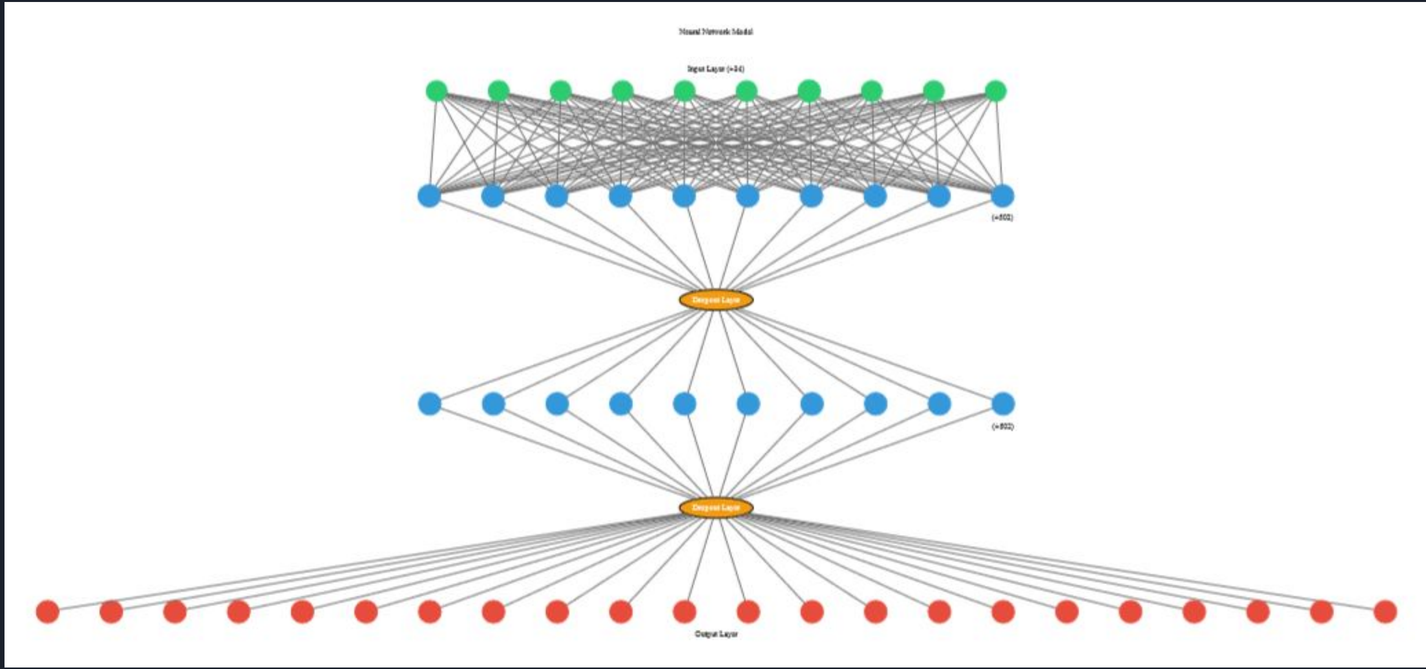
Model Selection and Performance

- Overall Evaluation Metric is Mean Adjusted Precision at 7 (MAP@7)
- Benchmark = 0.0042109
- Maximum Score on this Dataset = 0.031409

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

Model Selection and Performance

Neural Network Multiclass Classifier: MAP@7 = 0.0205575



Model Selection and Performance

- Collaborative Filtering: latent vector factorization

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	0	3	0	3	0
User 2	4	0	0	2	0
User 3	0	0	3	0	0
User 4	3	0	4	0	3
User 5	4	3	0	4	0

A matrix of user/item ratings

	Feature 1	Feature 2
User 1	?	?
User 2	?	?
User 3	?	?
User 4	?	?
User 5	?	?

X

	Item 1	Item 2	Item 3	Item 4	Item 5
Feature 1	?	?	?	?	?
Feature 2	?	?	?	?	?

=

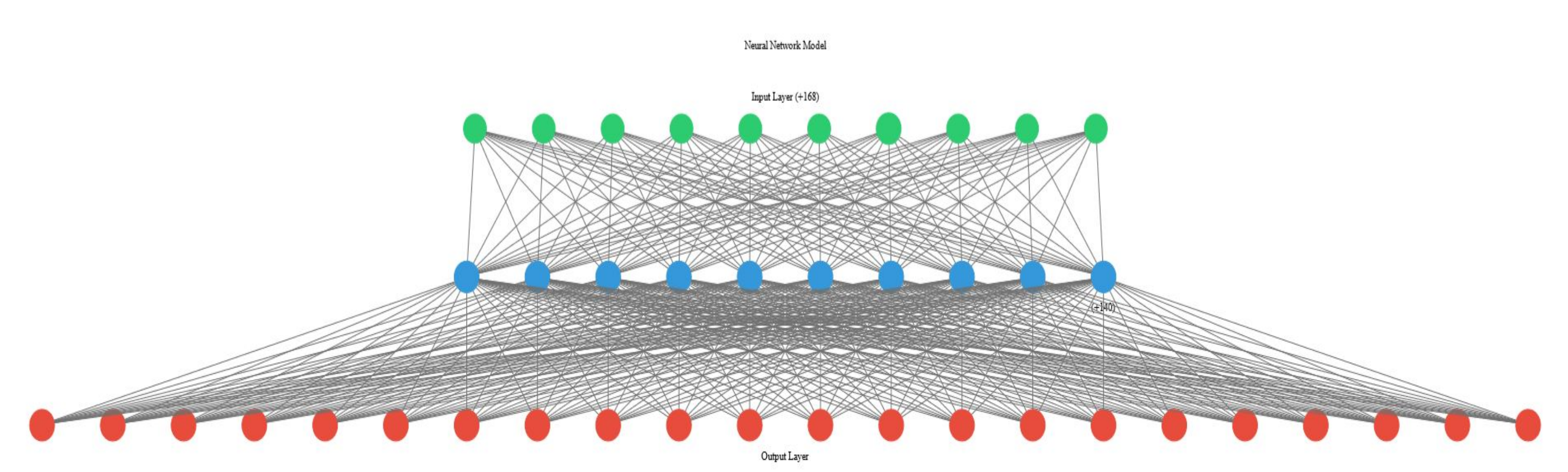
	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	0?	3	0?	3	0?
User 2	4	0?	0?	2	0?
User 3	0?	0?	3	0?	0?
User 4	3	0?	4	0?	3
User 5	4	3	0?	4	0?

Model Selection and Performance

- LightFM using only user-item interactions: MAP@7 = 0.0229795
- Random Forest Classifier with latent features: MAP@7 = 0.0233802
- Surprise! Using SVD algo on just May 2016 user-item interactions: MAP@7 = 0.0233802
- Surprise SVD May 2016 averaged with weights of June 2015 product distribution: MAP@7 = 0.024061
- Simply recommending the 7 most common newly purchased products in June 2015 that the customer does not already have: MAP@7 = 0.024061

Model Selection and Performance

- Neural Network without dropout layers trained on just records of newly purchased products in June 2015 with 5 month lags of products owned: MAP@7 = 0.030072



Recommendation to Client

- The most robust, agile solution is to employ SVD based collaborative filtering
- This can then be tuned to account for seasonal changes in purchasing patterns
- If maximally precise recommendations are required detailed seasonality analysis and lag feature engineering will accomplish that