SMU

SINGAPORE MANAGEMENT
UNIVERSITY

Masters of IT in Business

ISSS603 - Customer Focused Analytics and IT

Santander

PRODUCT RECOMMENDATION FOR DIGITAL MARKETING

PROJECT REPORT

Group 4

CHIN Chee Yung

Kaushik JAGANATHAN

Priyadarsan SHANKAR

Vaishnavi AGARWAL

Vishali Reddy NALLA

WU Jinglong

Table of Contents

1 Abstract

Santander is a multinational bank in Spain offering an array of financial services and products. 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. Models were built to aid the bank in selling products through mass marketing as well as personalized recommendations.

Customers were segmented based on the number of relations they held with the bank. All customers having more than two accounts were considered and equal width segmentation was applied. Market basket analysis was performed in all three segments and the association rules derived from each segment were used to select the potential products for mass marketing strategies. Those products in rules which had a high lift and low expected confidence were identified as high potential low selling products which could be mass marketed.

Santander's core business involves a client relationship manager guiding a client's investment decisions. Recommender systems can aid relationship managers with making personalized and automated selection of next best products for private banking clients. Different types of collaborative filtering recommender systems were built to help select the next best offer. Since the data had information only about the presence or absence of a product with a customer, Implicit Feedback CF with Alternating Least Squares was implemented where products were modelled as a function of Preference and Confidence. This system was improved by a user-user similarity-based system and further enhanced by incorporating demographic correlations. For the enhanced model, pairwise user-user similarity using cosine similarity using user-demography vectors was calculated. For each user, based on the enhanced similarity scores, 1000 users were identified using the k-Nearest Neighbors technique. The items to be recommended were predicted as a weighted average of the preferences from each user's neighborhood. Precision and Recall @ K were used to evaluate the recommender systems.

Index terms – Equal width segmentation, market basket analysis, association rules, recommender systems, implicit feedback collaborative filtering, alternating least squares, demographic correlations.

2 Introduction

Banco Santander, S.A., doing business as Santander Group, is a Spanish multinational commercial bank and financial services company founded and based in Santander, Spain. It offers an array of financial services and products including retail banking, mortgages, corporate banking, cash management, credit card, capital markets, trust and wealth management, and insurance. The bank is interested in improving its personal product recommendation system for its customers. It collects marketing information about its customers and associates it with bank account products. 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 an effective recommendation system in place, Santander can better meet the individual needs of all customers and ensure their satisfaction.

3 Research article review

3.1 Collaborative Filtering for Implicit Feedback datasets

As the Santander bank's product ownership data contains only unary indicators of product ownership, recommending products to customers is not possible by implementations of traditional recommender systems, which work on explicit feedback data of preference like review ratings provided by users. As the data contained in the dataset is implicit feedback, there is a need to model this data to be suitable for use with Collaborative filtering algorithms. This research paper by "Collaborative Filtering for Implicit Feedback Datasets" *Yifan Hu, Yehuda Koren, Chris Volinsky* provides a method to model the implicit feedback as a function of preference(p) and confidence(c) by assigning user owned accounts a positive preference with a high confidence and the accounts the user never interacted with a negative preference with a low confidence. Their approach suggests the usage of this modelled data in a traditional collaborative filtering by matrix factorization algorithm where the input user-item matrix is factorized into user and item factors whose product produces the predicted rating. Also, their implementation of the algorithm involves performing the matrix factorization using Alternating Least Squares method, as the user-item factor product term in the objective function makes it non-convex, so traditional optimization methods like gradient descent consume a lot of memory and run for more iterations. Since the Santander dataset contains data of around 0.68M customers, there is a requirement for an efficient optimization method, and ALS improves on the matrix factorization by setting the one of the user and item factors to a constant in alternative iterations and the other one as the one to be optimized. So, we've chosen to implement this algorithm to build the recommender system for the project, using it's implementation as a Python library called IMPLICIT.

3.2 Collaborative filtering enhanced by demographic correlations

The implicit feedback data of product ownership in the Santander is not very rich due to it's sparsity in terms of ownership as around 50% customers own 3 products or less and the product portfolio only comprises of 24 products so the recommender system might not perform well. Also, in the banking industry, the predominant features used to profile customers by similarity are the demographics, and based on our EDA and segmentation exercises we have seen that these features offer separability in purchase preferences, so we looked to add the demographic information to the recommender system model in motive to enhance its predictive power. To do this, we tended to a research paper "Collaborative Filtering Enhanced By Demographic Correlations" by *Manolis Vozalis and Konstantinos G. Margaritis* which provides a methodology to enhance a base USER-USER similarity based Collaborative filtering algorithm by incorporating demographic correlations. The research paper provides a method to factor in similarities between users in terms of their demographics alongside their product preference and produce an enhanced similarity measure, which can be used to weight products owned to product recommendations for each user. So, we've chosen to implement this algorithm to build an enhanced recommender system over the baseline user-user similarity based collaborative filtering recommender system.

4 Objective

In this project, we will analyze the Santander's bank customer profiles, data, products and transactions. Our aim is to predict which products their existing customers will use in the next month based on their past behavior and that of similar customers. The analysis will include the following:

a. Identify different approaches of segmentation based on customer's needs or values to market products and select the best approach
b. Perform Market Basket Analysis for the segments in the best segmentation approach and market the rules through different techniques.
c. Build recommendation engine to engage in personalized marketing.

5 Dataset Description

The dataset was sourced from https://www.kaggle.com/c/santander-product-recommendation/data. The dataset contains 1.5 years of customers behavior data from Santander bank to predict what new products customers will purchase. The data starts at 2015-01-28 and has monthly records of products a customer has, such as "credit card", "savings account", etc. The training data consists of nearly 1 million users with monthly historical user and product data between January 2015 and May 2016. User data consists of 24 predictors including the age and income of the users. Product data consists of Boolean flags for all 24 products and indicates whether the user owned the product in the respective months. This sample does not include any real Santander Spain customers, and thus it is not representative of Spain's customer base.

Santander CRM data:

Feature	Description
date	Date
customer_code	Customer code
employee_index	Employee index: A active, B ex employed, F filial, N not employee, P passive
country_of_residence	Customer's Country residence
sex	Customer's sex
age	Age
date_of_first_contract	The date in which the customer became as the first holder of a contract in the bank
new_customer_index	New customer Index. 1 if the customer registered in the last 6 months.
seniority	Customer seniority (in months)
customer_type	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
last_date_as_primary	Last date as primary customer (if he isn't at the end of the month)
customer_typ_at_begin	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner), P (Potential),3 (former primary), 4(former co-owner)

customer_relation_at_begin	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer), R (Potential)
residency_index	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
foreign_index	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
spouse_index	Spouse index. 1 if the customer is spouse of an employee
channel_used	channel used by the customer to join
deceased_index	Deceased index. N/S
address_type	Address type. 1, primary address
province_code	Province code (customer's address)
province_name	Province name
activity_index	Activity index (1, active customer; 0, inactive customer)
gross_income	Gross income of the household
segmentation	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated

Product portfolio:

Product	Description	Product	Description
ind_savings_account	Saving Account	ind_eaccount	e-account
ind_guarantees	Guarantees	ind_funds	Funds
ind_current_account	Current Accounts	ind_mortgage	Mortgage
ind_derivada_account	Derivada Account	ind_pensions	Pensions
ind_payroll_account	Payroll Account	ind_loans	Loans
ind_junior_account	Junior Account	ind_taxes	Taxes
ind_mas_particular_account	Más particular Account	ind_credit_cards	Credit Card
ind_particular_account	particular Account	ind_securities	Securities
ind_particular_plus_account	particular Plus Account	ind_home_accounts	Home Account
ind_short_term_deposit	Short-term deposits	ind_payrolls	Payroll
ind_medium_term_deposit	Medium-term deposits	ind_pensions_2	Pensions
ind_long_term_deposit	Long-term deposits		

6 Data Preparation and Integration

Data preparation consist of removing irrelevant data, imputing missing values, excluding skewed variables, recoding variables and performing feature engineering. Since the last month data was to be analyzed, it was extracted at the beginning.

As we inspect the missing value data pattern, the columns, 'last_date_as_primary' and 'spouse_index', had more than 99% of missing values. These two columns were removed. Next, the rows with missing value in other columns such as ‘province_name’ that was less than 0.5% were removed as they were observed to be non-contributory to the analysis. The calculation for the median ‘gross_income’ of each province were performed to prepare for value imputation. The missing values were imputed with the median values of the gross income based on province. Since the dataset was obtained from a Spanish company, the variables were recoded from Spanish notation to English Notation to achieve readability and clarity. For example, gender acronym was recoded from ‘H’ and ‘V’ to ‘F’ and ‘M’ respectively.

The dataset was further examined by checking the distribution of continuous and nominal variables. For continuous variables, the 'seniority' variable had outliers which were removed as they had invalid values of ‘-99999’. The figure below shows the distribution of the continuous variables before processing.

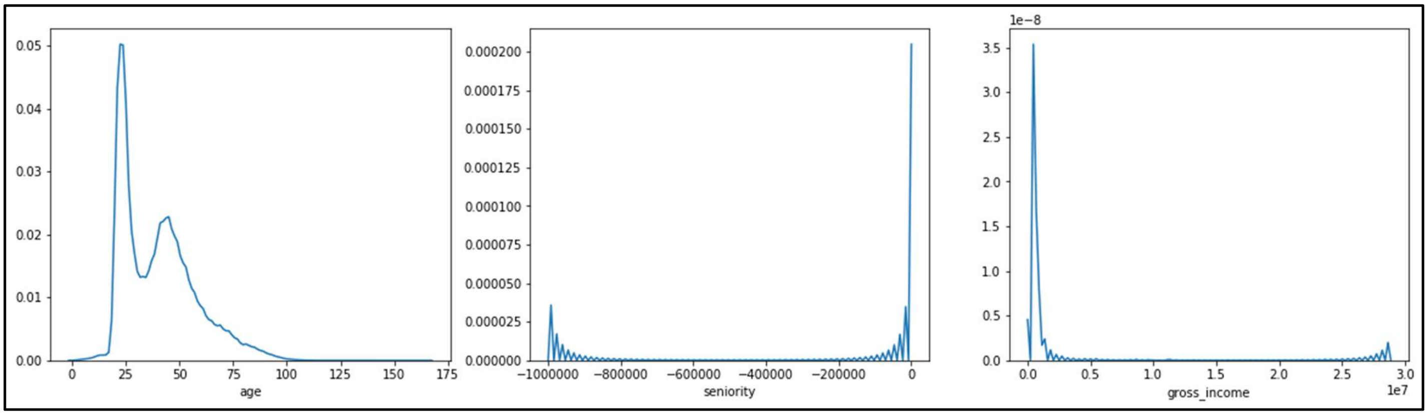


Figure 1: Continuous feature distribution

For nominal variables, 'country_of_residence', 'employee_index', 'deceased_index', 'customer_type' variables had a notably minimum of 99.8% of records that had the same value. The distributions can be seen in the figures below.

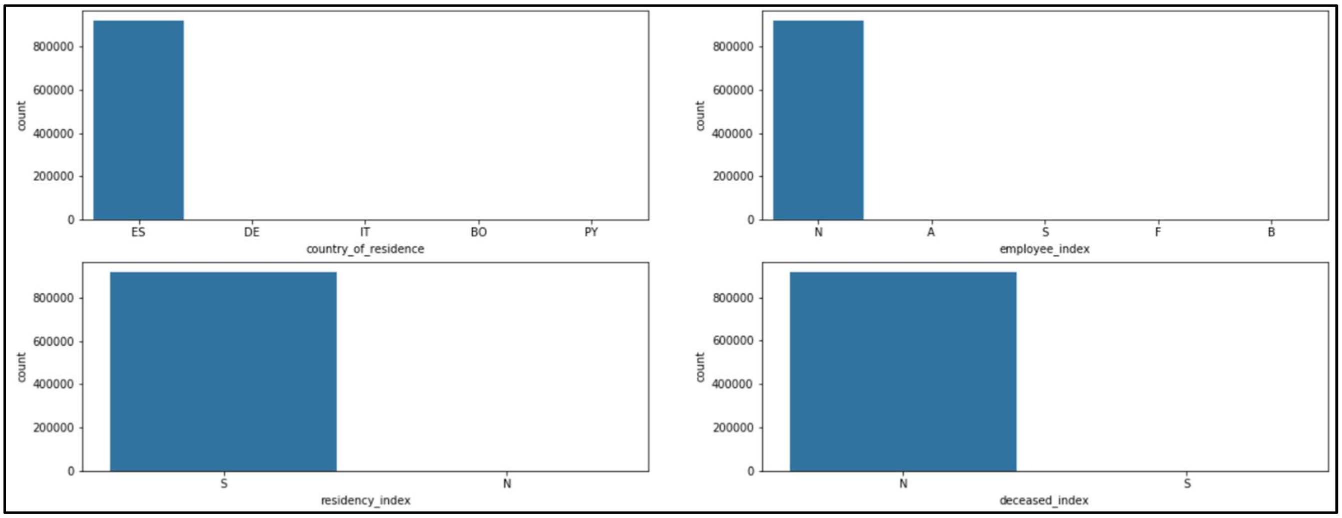


Figure 2: Categorical feature distribution

For 'country_of_residence', customers from other values except ‘ES’ were removed since the small records could not be analysed objectively. Thus, the records of variables 'customer_type', 'country_of_residence', 'employee_index', 'customer_typ_at_begin', 'residency_index' and 'deceased_index' were removed using this similar method. Furthermore, these columns were removed which were observed to have only one level of data. The variables, 'date_of_first_contract', 'address_type', 'province_code' and 'channel_used' were observed to be correlated or could be inferred to another variable. In particular, ‘date_of_first_contract’ was removed because it was correlated with ‘seniority’. ‘address_type’ and ‘province_code’ were removed because they could be inferred from ‘province_name’ and ‘channel_used’ was removed because the data dictionary did not provide further elaboration on the meaning of the values displayed.

In Feature Engineering, the provinces were group based on their GDP resulting in a new column ‘province_segment’, and a new column ‘No. of Accounts’ were created to show the number of accounts that a customer holds. Customers with no accounts were then removed. The final data after cleaning and preparation has 689449 rows with 35 columns.

7 Analysis Process Flow

The step by step process flow is as below:

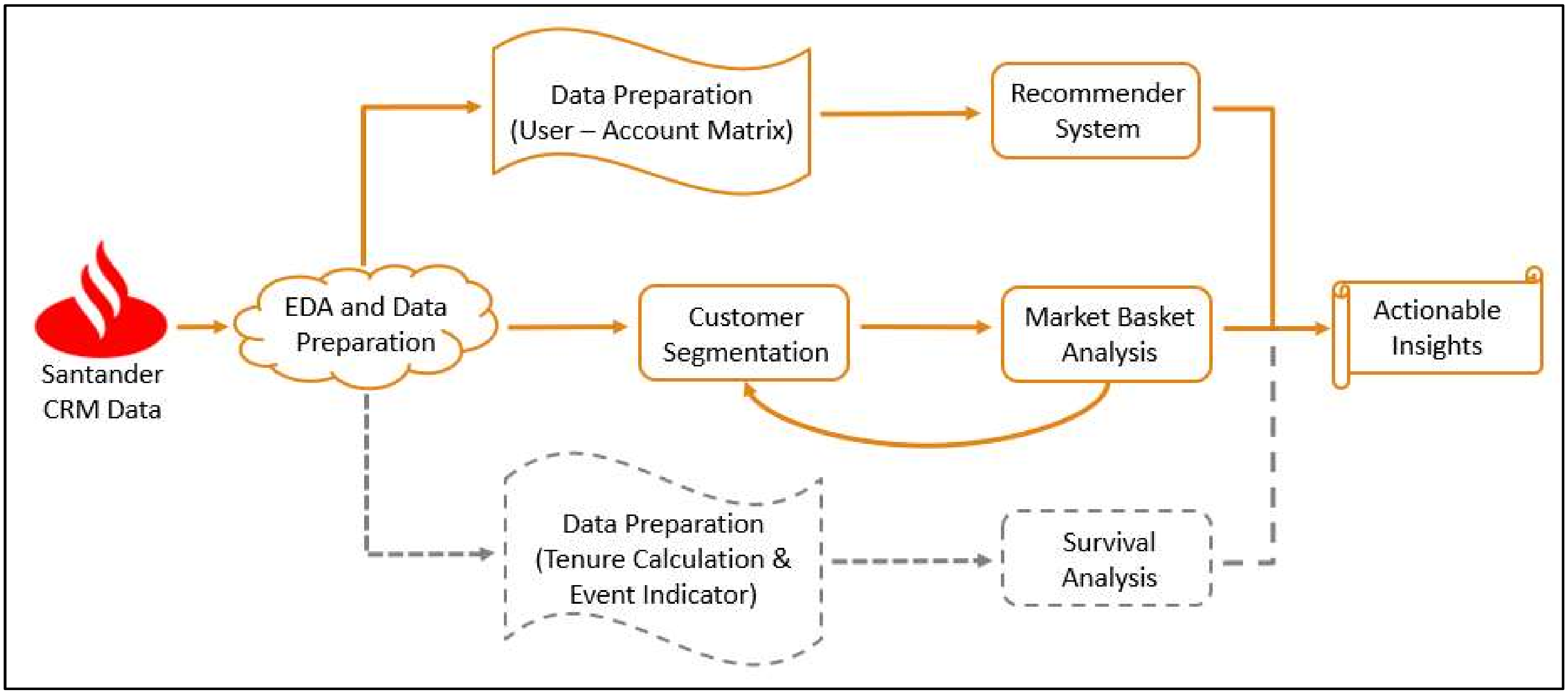


Figure 3: Analysis Process Flow

8 Exploratory Data Analysis

There were numerical and categorical variables. Within the demographics data section, ‘age’, ‘sex’, ‘seniority’, ‘gross_income’, ‘province_segment’ and ‘no_of_accounts’ were deemed meaningful to be analysed. First, the distribution of the numerical variables was plotted as shown in the figures below.

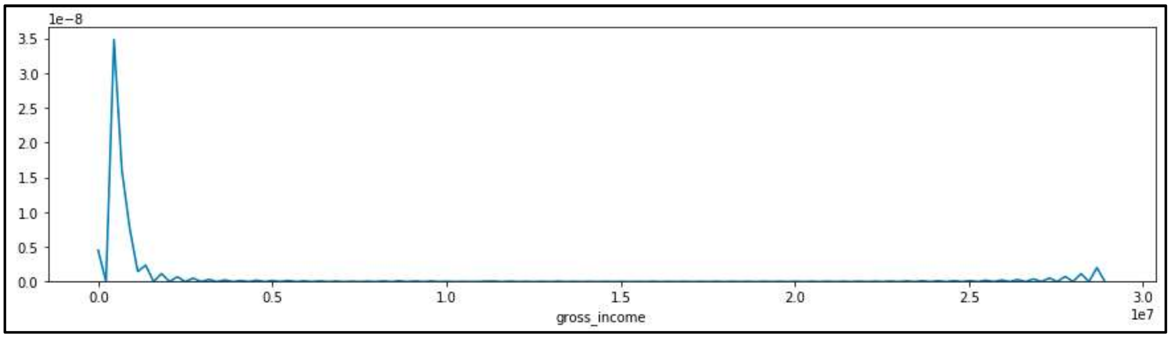


Figure 4: Distribution of 'gross_income'

The distribution of 'gross_income' showed that majority of the customers were in the lower income band of between $1000 to $200,000. This accounts for more than 99% of the customers. The 'rich' customer had a gross income of more than $22 million. While they appeared to be outliers, they were retained for further analysis.

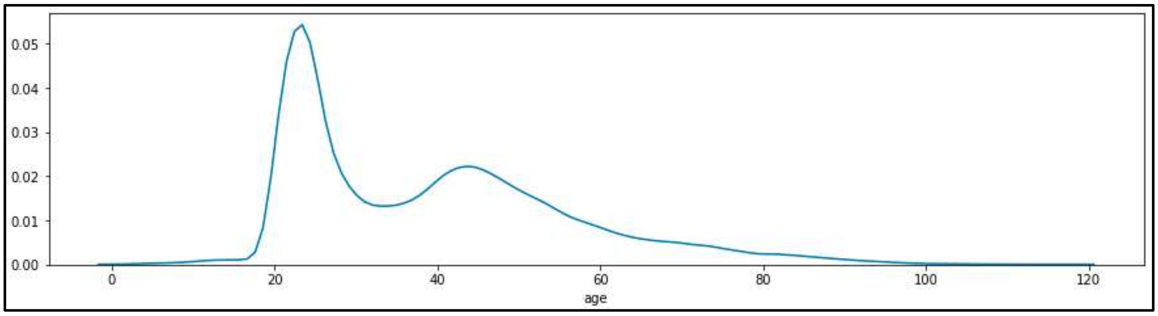


Figure 5: Distribution of 'age'

For the distribution of 'age', more customers were from the range of 20-60 years old. The distribution of 'seniority', below, showed that the customers' seniority from 5 to 50 months were of significance and the customers with seniority from 61 to 200 were of relative importance.

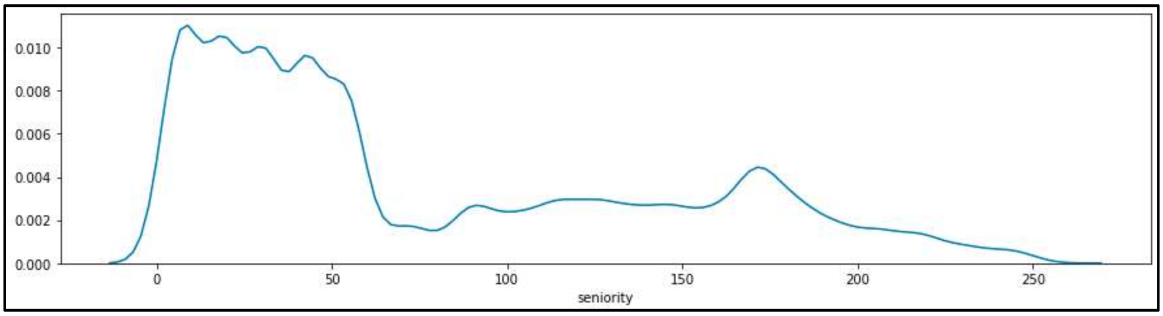


Figure 6: Distribution of 'seniority'

The 'no_of_accounts' and 'province_segment' variables were plotted below. The grouped province segments for A has the highest GDP and D having the lowest GDP. In other words, the richer customers come from 'province_segment' A while the customer who belonging to segment B were from poorer provinces.

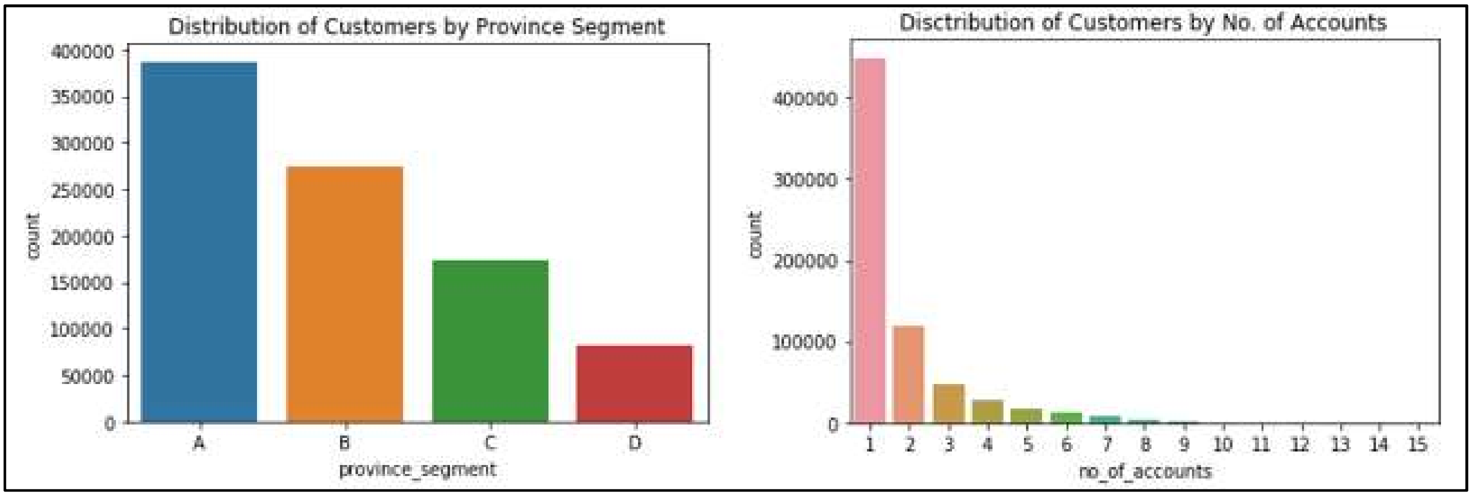


Figure 7: Distribution of province segment and No. of accounts

In the bank's interest, the gender may not be of significance as compared to gross income and province segments. As the bank work based on monetary terms, we further explored based on this fact. The 'gross_income' vs 'no_of_accounts' vs 'province_segment' were plotted below and it revealed that higher income customers were from segment A. However, the other segments showed similar characteristics.

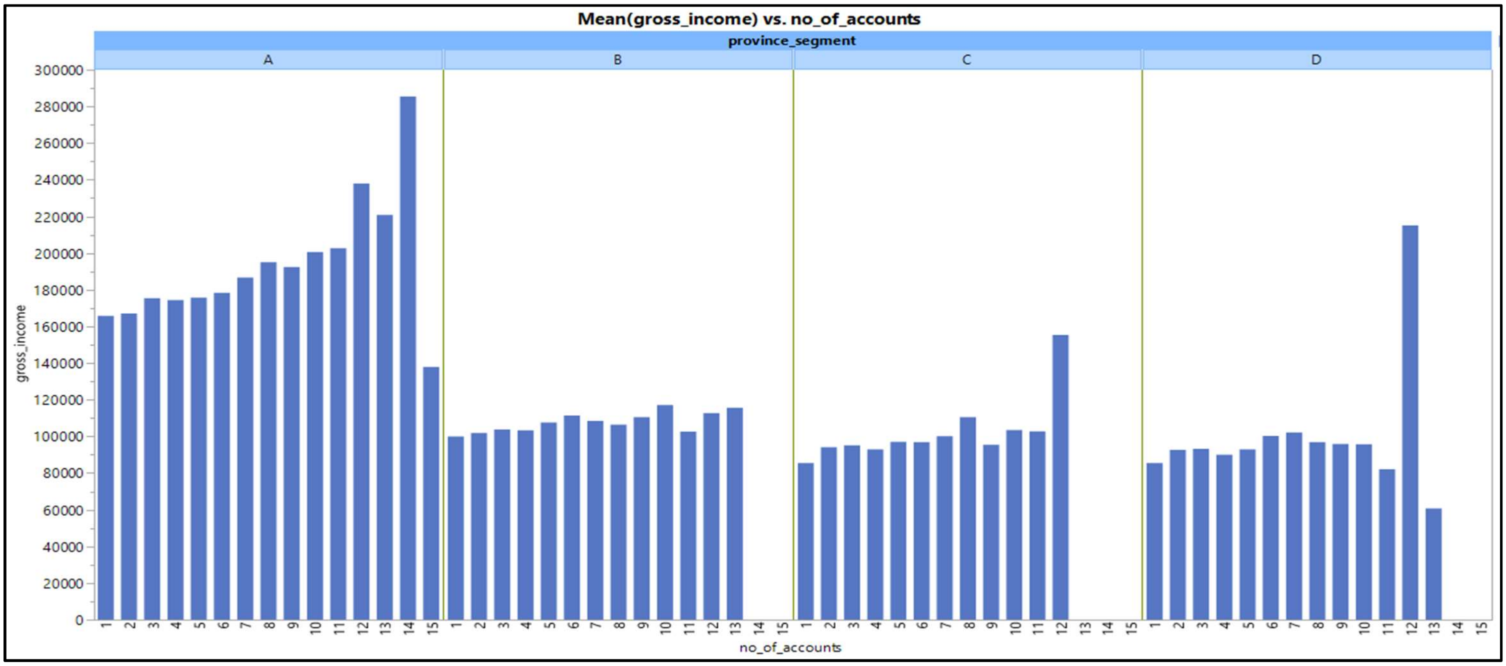


Figure 8: Mean Income v/s No. of accounts for each province

Exploring other variables did not reveal any significant insights as most of the nominal variables were highly skewed. While the customer could be categorized based on province segments, no of accounts and gross income, they had no conclusive outcome based on the accounts the customers hold. Customer segmentation will be further examined in the next section.

The popularity can be examined by counting the number of customers that own the account. The plot is shown in the figure below. The most popular account was the 'ind_current_account', which amounts to 80.6% of 689,449 customers. Thus, the top 5 popular accounts among the customers were:

1. Ind_current_account (80.6%)
2. Ind_direct_debits (16.3%)
3. Ind_particular_account (14.3)
4. Ind_eaccount (10.9%)
5. Ind_payroll_account (10.6%)

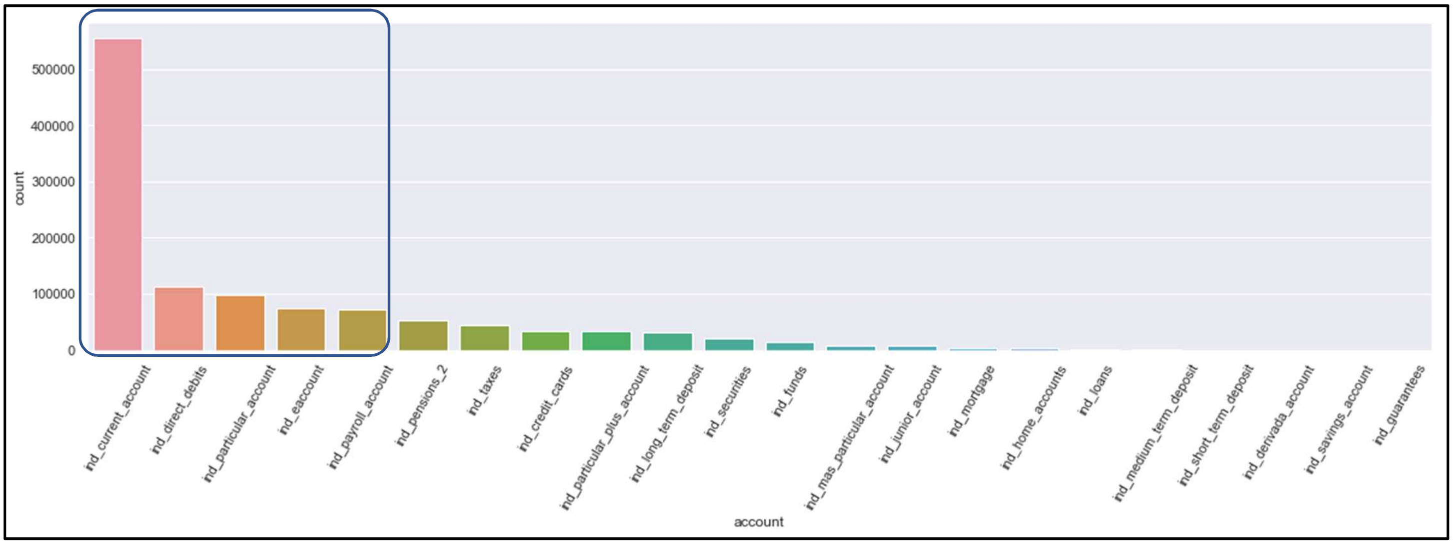


Figure 9: Most Popular Accounts of Santander

Since there were some distribution based on gross income, age and seniority. We can perform bivariate analysis with the top 5 accounts. Plotting the distribution with 'gross_income', it appears that all top 5 popular accounts have similar distribution to the 'gross_income' with the exception of 'ind_payroll_account' where there were more customers with income of $200,000 to $220,000 (as indicated).

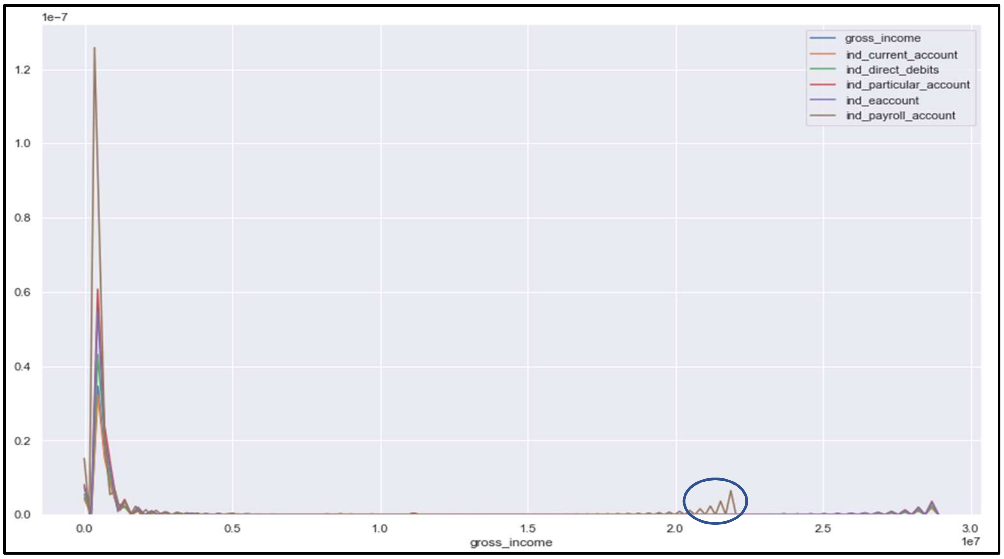


Figure 10: Top 5 Popular Accounts against Gross Income

Comparing with ‘age’, the most popular account resides with the age grouo of 18-30 years old. The other 4 popular accounts have more customers within the age group of 40-60 year sold.

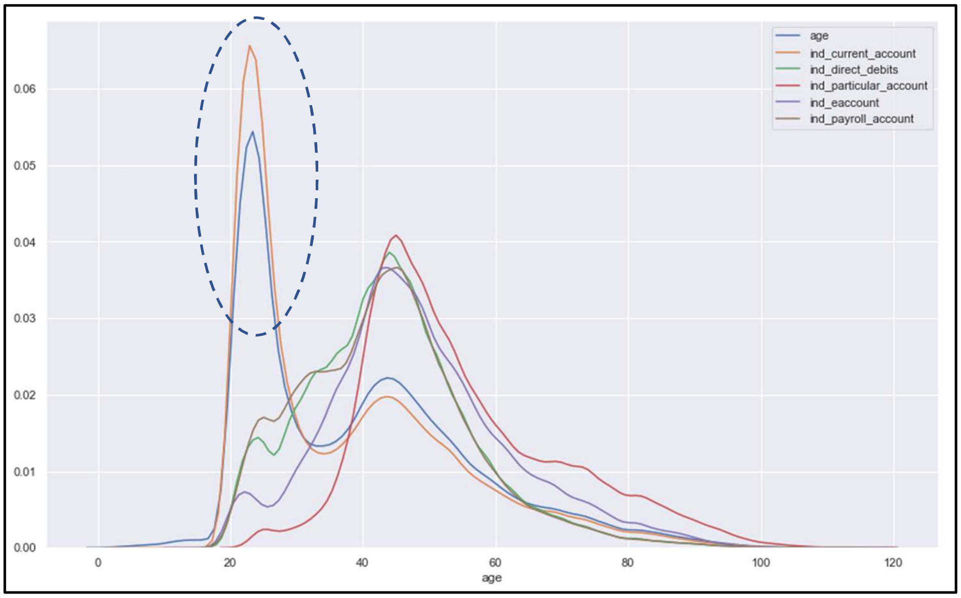


Figure 11: Top 5 Popular Accounts against Age

Examining against the ‘seniority’, the ‘ind_particular_account’ had the highest customer count with the seniority from 120-180 months. The inference would be to target customers within this range. The rest of the accounts are similar to the distribution of ‘seniority.

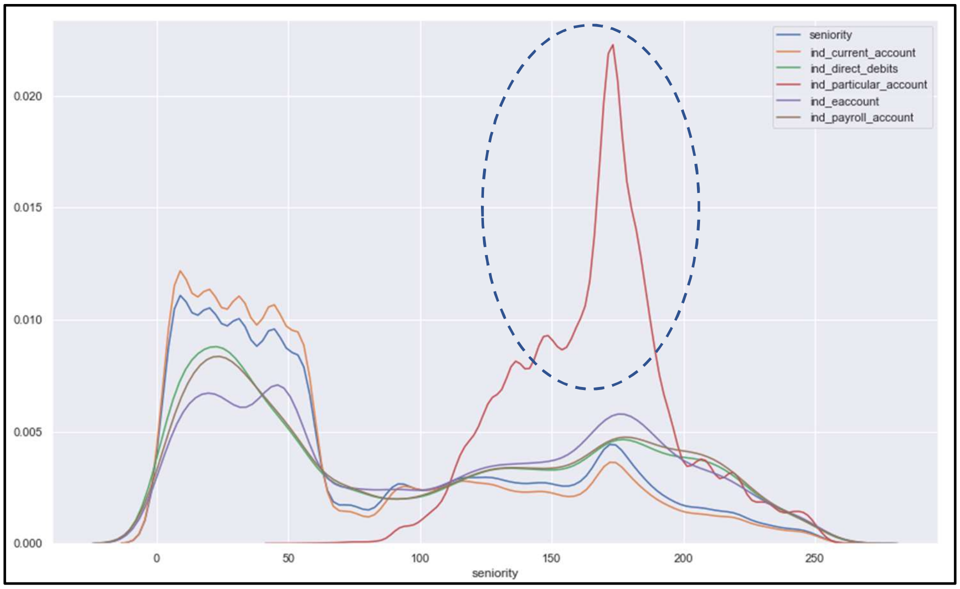


Figure 12: Top 5 Popular Accounts against Seniority

From the EDA results, we could further analyse the ‘age’, ‘seniority’ and ‘gross_income’ using customer segmentation, Market Basket Analysis and Collaborative Filtering techniques to develop marketing strategies to the customers.

9 Customer Segmentation

Banco Santander aims to reduce marketing cost by promoting banking products through mass-marketing and website design strategies. In order to achieve this objective, it is important to identify groups of individuals that are similar in specific ways such as age, income or product interests which will help the bank to do more target specific marketing and encourage customer to buy more.

Customer segmentation is the most common practice that is used to divide the customer base into small groups based on similarity of customer preferences and needs. By identifying key differences between the groups, it will be easier for the bank to discover what products or services are most valuable for each group of customers and also help in determining profitable and non-profitable segments.

Multiple details can be taken into consideration while doing segmentation as Santander owns a rich CRM data consisting of demographics, relation with *bank* and product ownership. As the data contains continuous variables like income, age and seniority we can simply perform K-Means Clustering to segment the customer on basis of these three variables. But in banking industry, customer segments are generally created by defining rules based on the business requirements or final goal, therefore we have segmented the customer using both the approaches considering different use cases.

9.1 Rule Based Segmentation

Rule based segmentation allows to manually define the condition for segmentation and the rules are usually based on the marketing strategy or business target. Rule based segmentation can be done using one or more variables. When the customers are segmented based on just one similarity like age or income it is single-level rule based segmentation whereas, if you first segment the customers by income and then for each income segment, you further segment by age it is multi-level rule based segmentation.

9.1.1 Single-level

The variable selection for doing single-level rule based segmentation depends on the marketing objective. If the bank wants to market value added services and products with high premium the target customers will be of high income, in that case the single-level rule based segmentation will be done based on income. Similarly, if the bank launches a new junior account, they will segment customers by age and promote it to the young customers. Therefore, considering all possible business requirements we have tried various single-level segmentations.

Continuous variables like income, age and seniority have been binned in three equal quantiles and for number of accounts equal width bins are created. The distribution of customers across segments for these variables can be seen below:

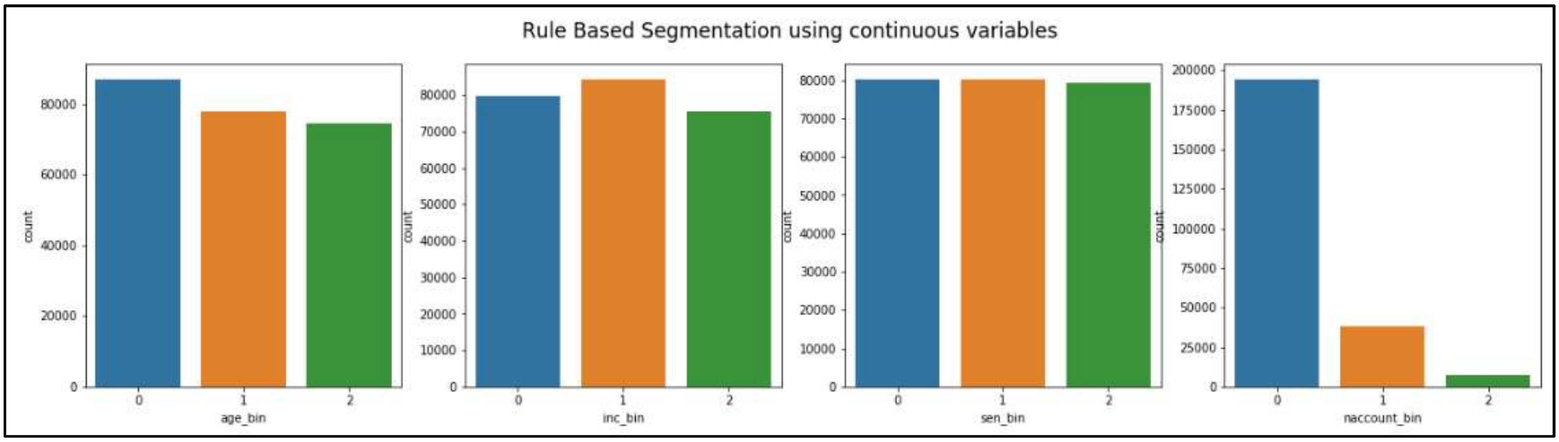


Figure 13: Segment distribution using continuous variables (single – level rule based)

The CRM data also contains various nominal variables out of which we have selected customer type, activity index (*0 – not active; 1 – active*), new customer index (*1 – joined in last 6 months; 0 – more than 6 months*) and foreign index (*N – not foreigner; S – foreigner*) as other nominal variables were highly skewed. So, while using these variables for single-level rule based segmentation, the customer segments were made based on the categories in each variable. The customer segments for these variables are as below:

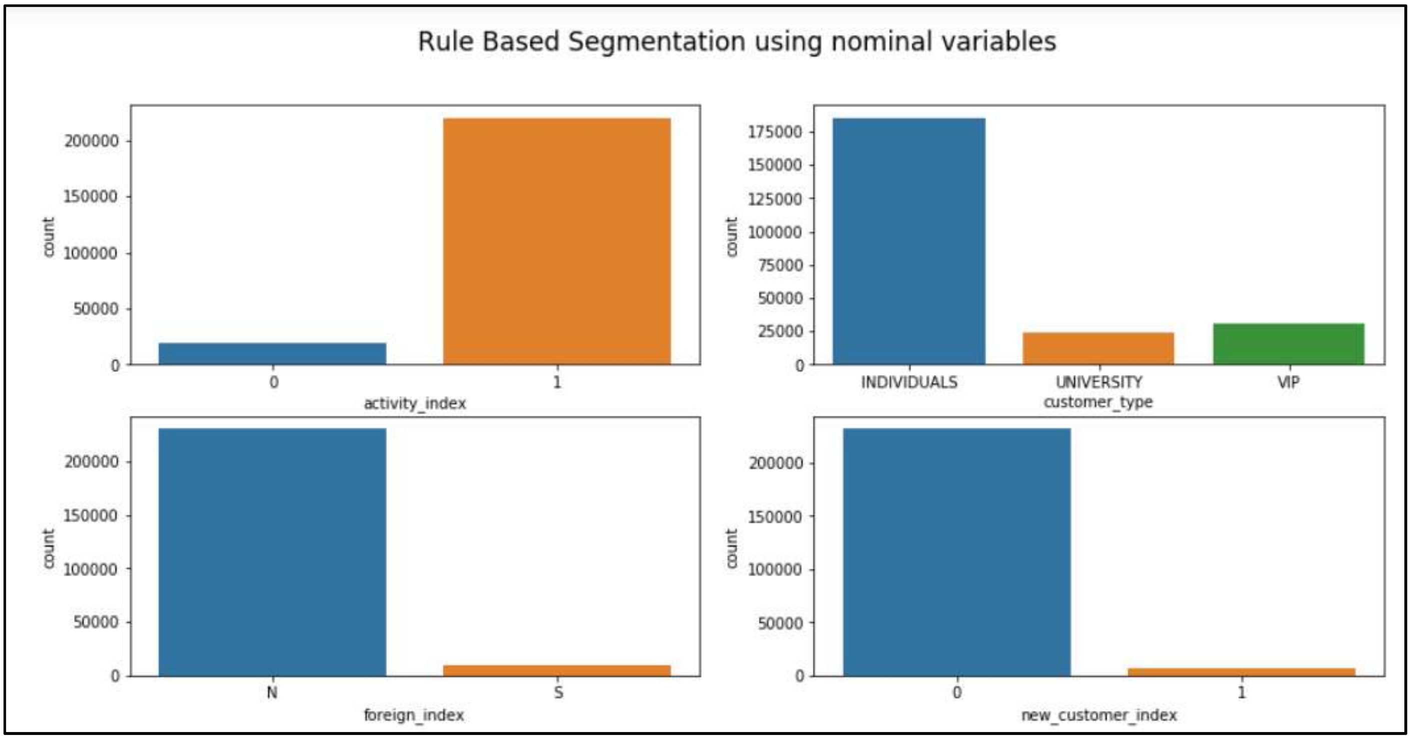


Figure 14: Segment distribution using nominal variables (single – level rule based)

9.1.2 Multi-level

Multi-level rules based segmentation uses at least 2 variable to determine the similarity between customers and create segments. There are different types of deposit account (*short-term, medium-term, long-term*) available at Santander, and the minimum requirements to own these accounts may vary. So for example, the target customers for long-term deposit account would be rich and young customers, so that they deposit more and be committed with the bank for a longer time. In this case, the bank would use multi-level rule based segmentation and use income along with age to segment their customer base. Similarly, multi-level segmentation using income along with no. of accounts owned can help the bank identify the customer who earn more but does not hold more no. of accounts, so that specific marketing strategies can be designed for such group of customers to make them own more number of products. The selection of variable and level of segmentation can vary based on the marketing requirement, but here we have considered the above discussed use cases and performed multi-level rule based segmentation using Income along with Age as one method and Income along with No. of accounts as another.

The CRM data is first divided into 3 quantiles of income (*low; medium; high*), 2 quantiles of age (*young; old)* and 3 equal width bins for no. of accounts (*2-4; 5-7; more than 7*). Then the segments are formed by selecting relevant bins. Below is the list of customer segments created from multi-level rules based segmentation

Multi-Level Rule Based Segmentation using *Income and Age*	Low Income – Young Age Low Income – Old Age Average Income – Young Age Average Income – Old Age High Income – Young Age High Income – Old Age
Multi-Level Rule Based Segmentation using *Income and No. of Accounts*	Low Income – 2-4 accounts Low Income – 5-7 accounts Low Income – more than 7 accounts Average Income – 2-4 accounts Average Income – 5-7 accounts Average Income – more than 7 accounts High Income – 2-4 accounts High Income – 5-7 accounts High Income – more than 7 accounts

Figure 15: Customer segments from multi-level rule based segmentation

The customer distributions across segments in multi-level rule based segmentation for income along with age and income along with number of accounts is as below:

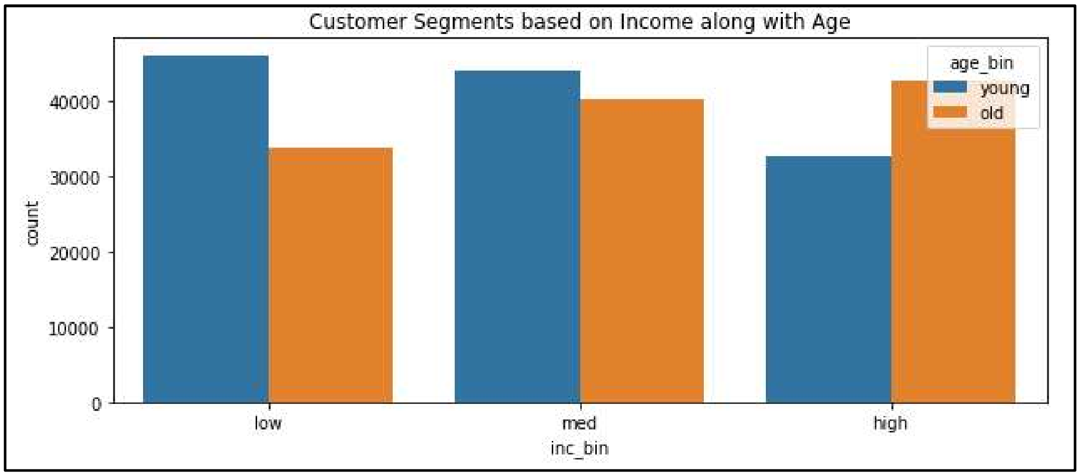


Figure 16: Segment distribution using Income and Age (multi-level rule based)

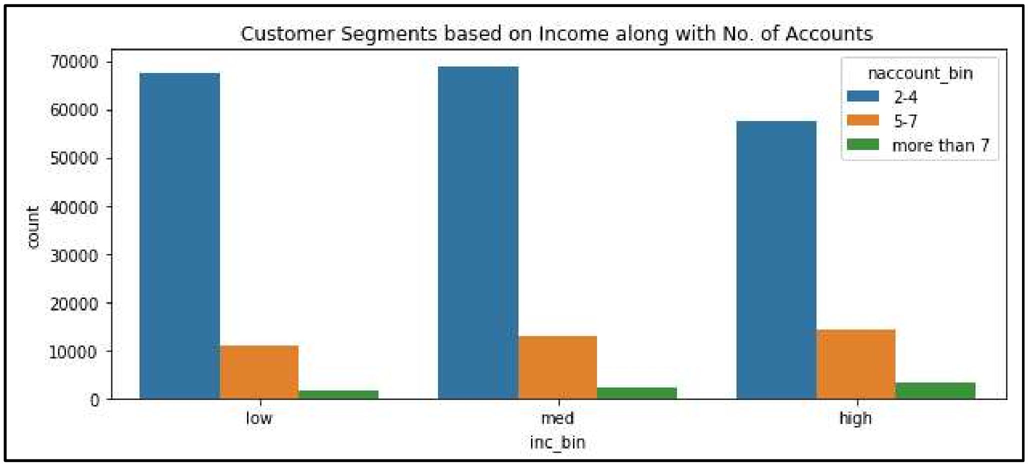


Figure 17: Segment distribution using Income and No. of accounts (multi-level rule based)

9.2 Segmentation by K-Means Clustering

K-Means clustering is an unsupervised learning algorithm that is used to find cluster of data points within the data and discover underlying patterns. The number of centroids(K) needs to be defined and then the algorithm iterates between assigning data points to the centroid and updating the centroids until the within-cluster Euclidian distance is minimized. Instead of segmenting the customer by manually deciding the rules, we have used K-means clustering to find the similarity between customer based on their gross income, age and seniority and defined the number of clusters(K) as 3.

Once the K-means clusters are formed, we have analyzed the characteristics of these clusters based on the three variables used to cluster similar customers. The distribution plots for income, age and seniority across the segments shows that among the three features, seniority is the most differentiating feature.

The first cluster consists of customer with higher seniority (>130 months) therefore, we have named this segment as *Long Term Customers*. The second cluster consists of customers with lower seniority (0 – 50 months) so, we name this cluster as *Recent Customers* and the third cluster consists of customers with an average seniority (50 – 120 months) so we name this cluster as *Loyal Customers*. The variable distribution across segments is as below:

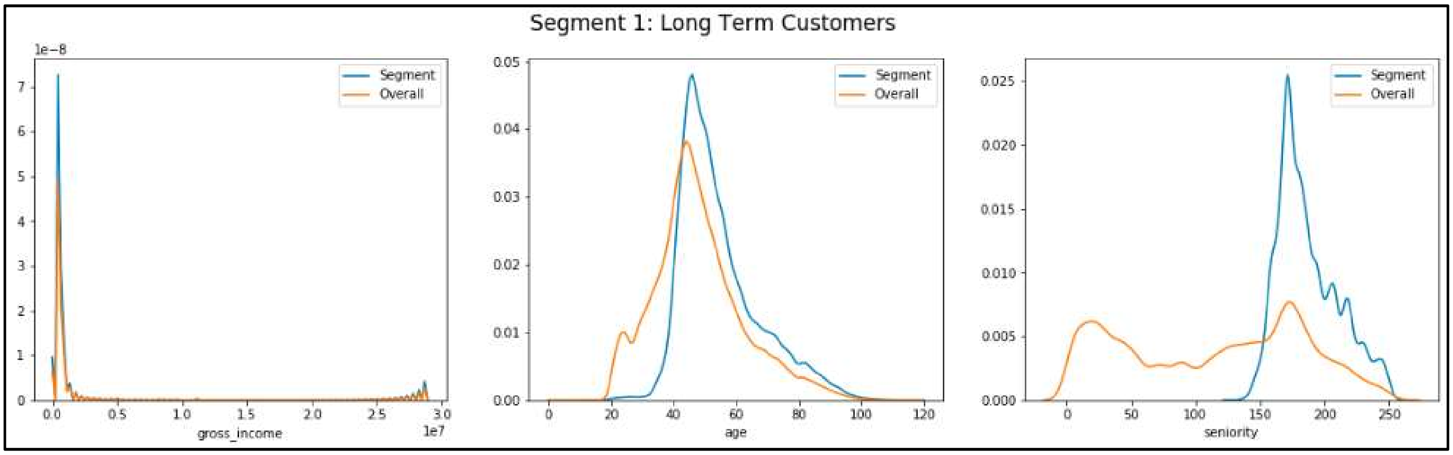


Figure 18: Variable distribution for Long Term Customers segment (K-Means clustering)

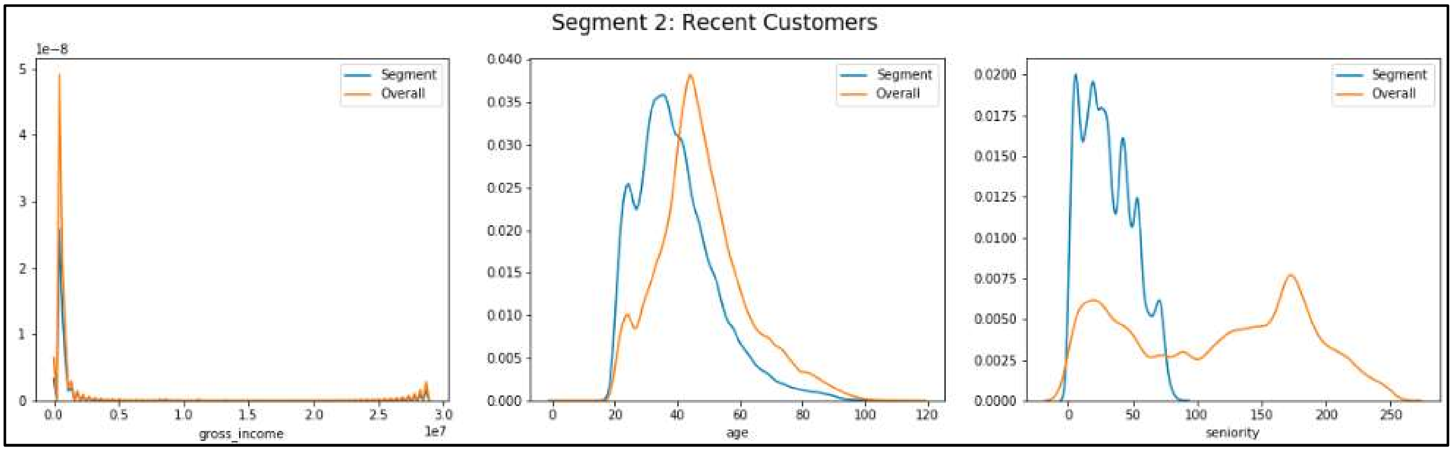


Figure 19: Variable distribution for Recent Customers segment (K-Means clustering)

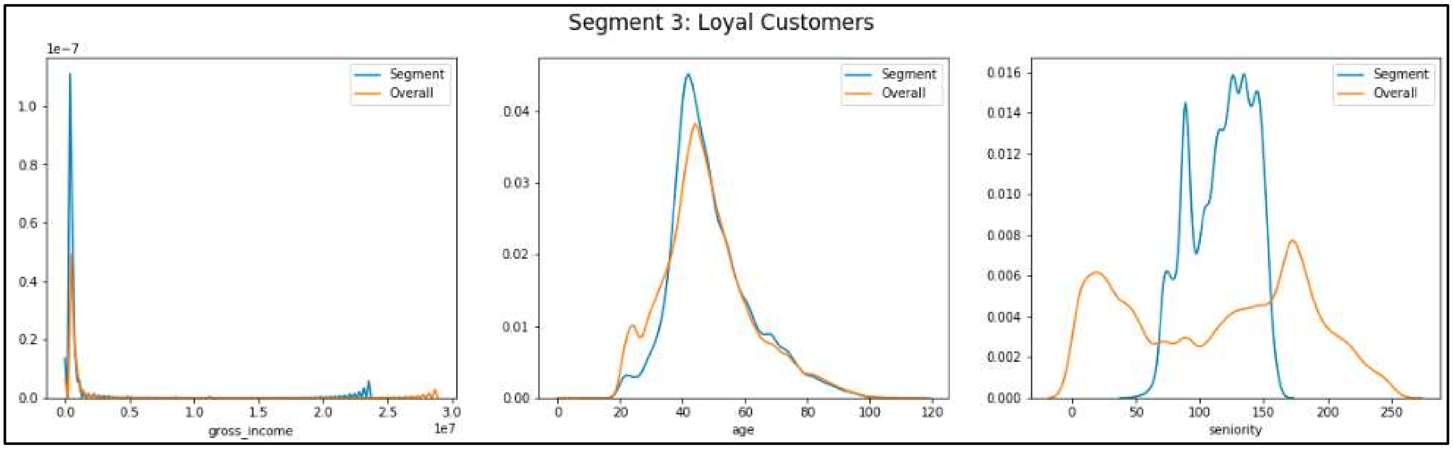


Figure 20: Variable distribution for Loyal Customers segment (K-Means clustering)

9.3 Segmentation Method Selection

As discussed above, the segmentation approach can vary depending on the business requirement. But based on our research, banks usually follow a single-level rule based segmentation using customer demographics like income and age. Therefore, instead of following the banks traditional method we decided to perform segmentation using no. of accounts owned by each customer.

The objective behind selecting this variable for segmentation was to check if the no. of relations a customer hold with a bank reflects any pattern in the product ownership. Also, if we segment the customers based on products owned, we can market potential products to the lower segment and migrate them to the next segment. We have used single-level rule based method using no. of accounts for segmenting Santander's customer base. Equal width bins are created for no. of accounts and the customers are divided into three segments: 2-4 accounts, 5-7 accounts and 8-10 accounts. The distribution of customers across segments is as below:

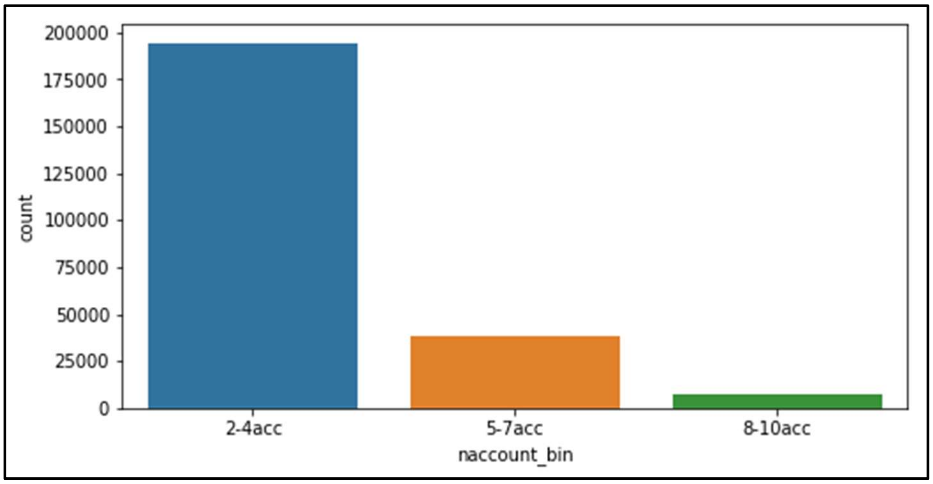


Figure 21: Segment distribution using No. of Accounts (Single-level rule based)

Once the segments were created, we explored the characteristics of the segments to see if there was enough separability across segments in terms of other features. The distribution of age and seniority was checked to identify if people who are committed to the bank for a longer time actually own more products or were just using one products over the years. We also checked the activity index of customers across segments. The distribution plots of these features can be seen below:

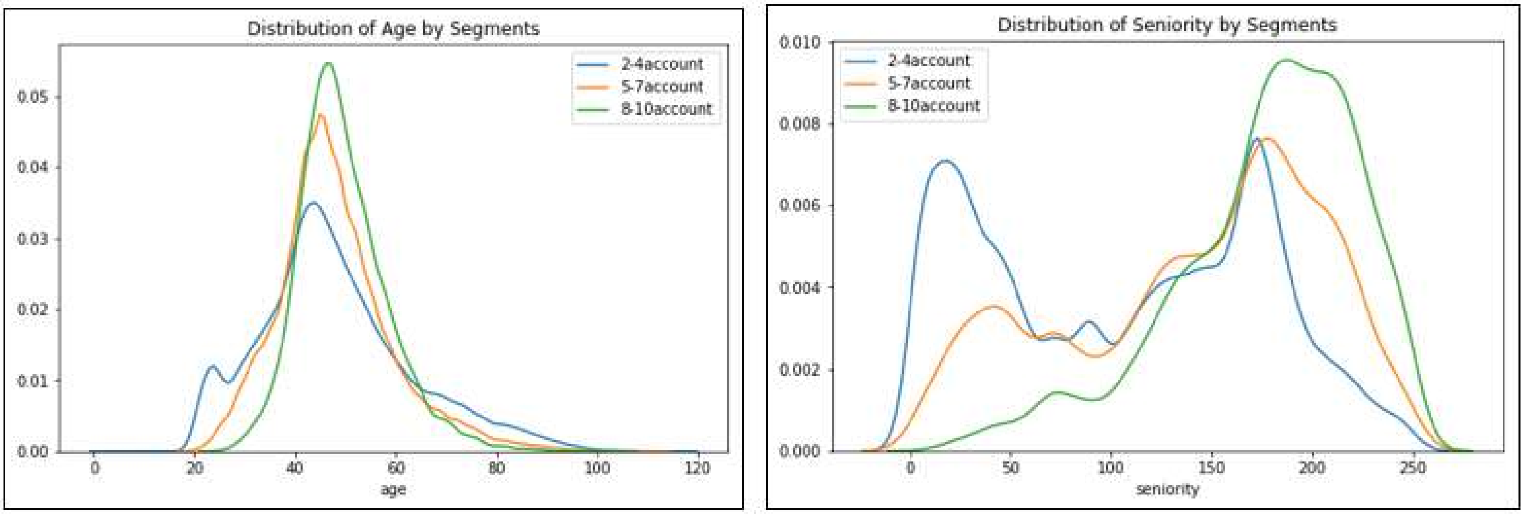


Figure 22: Distribution of Age and Seniority across segments (single-level - no. of accounts)

From the above plots, it can be inferred that if the customers belong to the 8-10accounts segment, they have been committed with the bank(seniority) for a long time and also that, young customers(age) tend to own less number of accounts and mostly belong to the 2-4accounts segment.

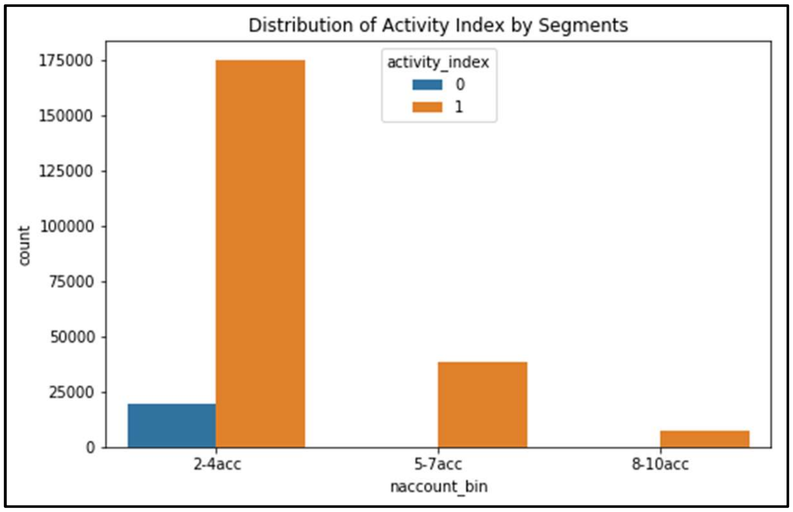


Figure 23: Distribution of Activity Index across segments (single-level - no. of accounts)

From the count plot above, we could see that almost all the inactive customers belonged to 2-4 account segment and from the previous distribution plot (*Figure 13*) we know that these are the young and recent customers, so the bank can allot more marketing budget for this segment and attract them to buy more products.

10 Market Basket Analysis for Mass Marketing

Market basket analysis was performed to understand associations between different accounts that a customer owns. This strategy is widely used to define products for mass marketing campaigns like website flyers, brochures or promotion e-mails.

Customers usually have different portfolio of accounts with a bank and we cannot generalize association rules for customers, therefore we used the results of segmentation (based on no. of accounts) to add a level of personalization for similar customers.

Market Basket Analysis was performed using apriori algorithm executed in python. Separate rules were generated of each segment – 2-4account (low), 5-7account (medium), 8-10account (high).

Data Preparation and steps for MBA:

- The data was already in the user-product format, where each row is one customer, and the product columns have 0’s and 1’s which depicts if the customer owns the product or not. To perform MBA we don’t require other demographic and relations variables, therefore we remove all the columns other than the product columns and create the transactions data to perform MBA.
- Market Basket Analysis was performed on each segment using apriori algorithm, and ‘lift’ is used as the evaluation metric. We have defined a minimum support = 0.2 for the products to appear in the association rules, also the minimum lift is set to 1.2, as we want to analyze the top rules.

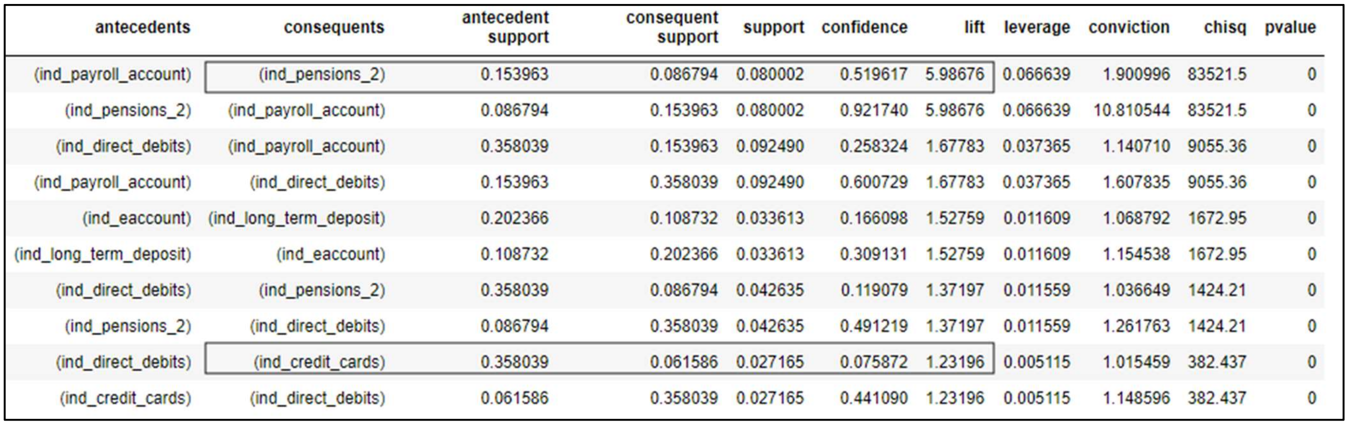
- Chi Square test was used to find the most significant association rules generated from MBA. The chi-square test will give a p-value for each rule, and we consider rules with p-value greater than 0.05 as significant rules to be used for further analysis.

$$\chi^2 = n\ (lift - 1)^2 \frac{supp\ conf}{(conf - supp)\ (lift - conf)}$$

Figure 24: Formula used to calculate the chi-square value

To attract customers to buy more accounts, the bank must market the most potential products to each segment. To find the most potential products in each segment, the association rules generated for each segment was used. The product that appears on the RHS in most of the rules and has a lower support (consequent support) is considered to be the most potential product for that segment . The rationale behind this approach is to promote products which are less selling among the customers (low consequent support) of that segment but have good potential to be bought by the customers (high lift).

We will see at the association rules for the lower segment and go through the process of selecting the most potential products for this segment, below are the top association rules generated from MBA for 2-4account segment:

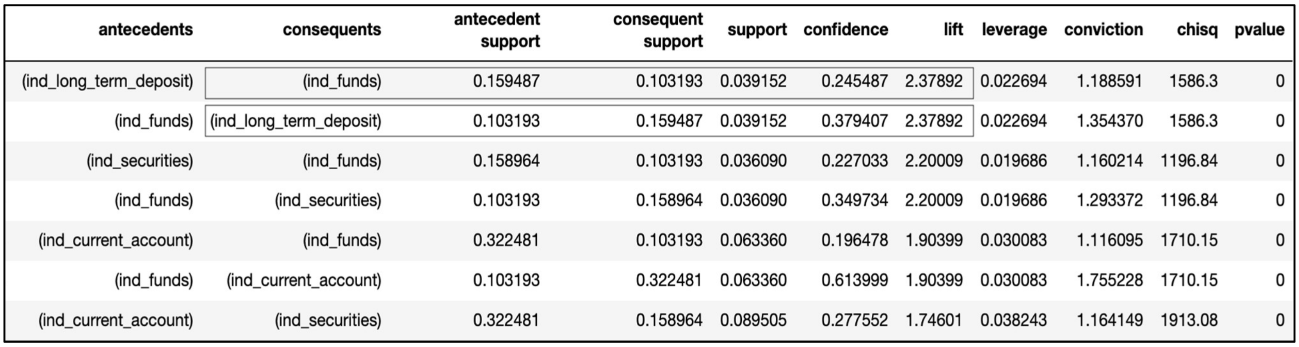


antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	chisq	pvalue
(ind_payroll_account)	(ind_pensions_2)	0.153963	0.086794	0.080002	0.519617	5.98676	0.066639	1.900996	83521.5	0
(ind_pensions_2)	(ind_payroll_account)	0.086794	0.153963	0.080002	0.921740	5.98676	0.066639	10.810544	83521.5	0
(ind_direct_debits)	(ind_payroll_account)	0.358039	0.153963	0.092490	0.258324	1.67783	0.037365	1.140710	9055.36	0
(ind_payroll_account)	(ind_direct_debits)	0.153963	0.358039	0.092490	0.600729	1.67783	0.037365	1.607835	9055.36	0
(ind_eaccount)	(ind_long_term_deposit)	0.202366	0.108732	0.033613	0.166098	1.52759	0.011609	1.068792	1672.95	0
(ind_long_term_deposit)	(ind_eaccount)	0.108732	0.202366	0.033613	0.309131	1.52759	0.011609	1.154538	1672.95	0
(ind_direct_debits)	(ind_pensions_2)	0.358039	0.086794	0.042635	0.119079	1.37197	0.011559	1.036649	1424.21	0
(ind_pensions_2)	(ind_direct_debits)	0.086794	0.358039	0.042635	0.491219	1.37197	0.011559	1.261763	1424.21	0
(ind_direct_debits)	(ind_credit_cards)	0.358039	0.061586	0.027165	0.075872	1.23196	0.005115	1.015459	382.437	0
(ind_credit_cards)	(ind_direct_debits)	0.061586	0.358039	0.027165	0.441090	1.23196	0.005115	1.148596	382.437	0

Figure 25: Association rules for 2-4account (low) segment

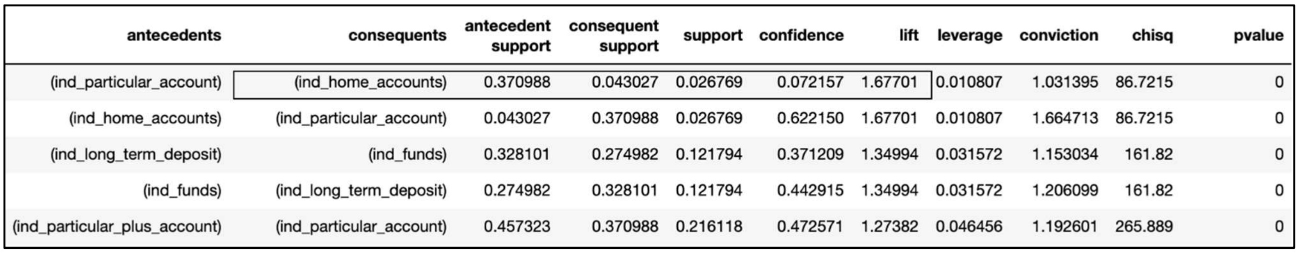
For the above list of association rules, ind_pensions_2 and ind_credit_cards are the two product with low consequent support as compared to others, also the rules containing these product have high lift and the antecedent product (direct debit = 0.358) has a higher support. This means that most of the customers in this segment owned a direct debit account and as per the rules, pensions and credit card account were the next best accounts. Therefore, for this segment our potential products would be pensions account(best) and credit card account(second best).

Similarly, the top rules were generated for each segment by setting the threshold of lift and support and then calculating the p-value. The list of top association rules for the other two segment is as below:



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	chisq	pvalue
	(ind_long_term_deposit)	(ind_funds)	0.159487	0.103193	0.039152	0.245487	2.37892	0.022694	1.188591	1586.3	0
	(ind_funds)	(ind_long_term_deposit)	0.103193	0.159487	0.039152	0.379407	2.37892	0.022694	1.354370	1586.3	0
	(ind_securities)	(ind_funds)	0.158964	0.103193	0.036090	0.227033	2.20009	0.019686	1.160214	1196.84	0
	(ind_funds)	(ind_securities)	0.103193	0.158964	0.036090	0.349734	2.20009	0.019686	1.293372	1196.84	0
	(ind_current_account)	(ind_funds)	0.322481	0.103193	0.063360	0.196478	1.90399	0.030083	1.116095	1710.15	0
	(ind_funds)	(ind_current_account)	0.103193	0.322481	0.063360	0.613999	1.90399	0.030083	1.755228	1710.15	0
	(ind_current_account)	(ind_securities)	0.322481	0.158964	0.089505	0.277552	1.74601	0.038243	1.164149	1913.08	0

Figure 26 : Association rules for 5-7account (medium) segment



antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	chisq	pvalue
(ind_particular_account)	(ind_home_accounts)	0.370988	0.043027	0.026769	0.072157	1.67701	0.010807	1.031395	86.7215	0
(ind_home_accounts)	(ind_particular_account)	0.043027	0.370988	0.026769	0.622150	1.67701	0.010807	1.664713	86.7215	0
(ind_long_term_deposit)	(ind_funds)	0.328101	0.274982	0.121794	0.371209	1.34994	0.031572	1.153034	161.82	0
(ind_funds)	(ind_long_term_deposit)	0.274982	0.328101	0.121794	0.442915	1.34994	0.031572	1.206099	161.82	0
(ind_particular_plus_account)	(ind_particular_account)	0.457323	0.370988	0.216118	0.472571	1.27382	0.046456	1.192601	265.889	0

Figure 27 : Association rules for 8-10account (high) Segment

The same approach as stated above was used to find the top two potential products for the other two segments. The potential products that can be used for marketing purposes by bank on segment level are listed below:

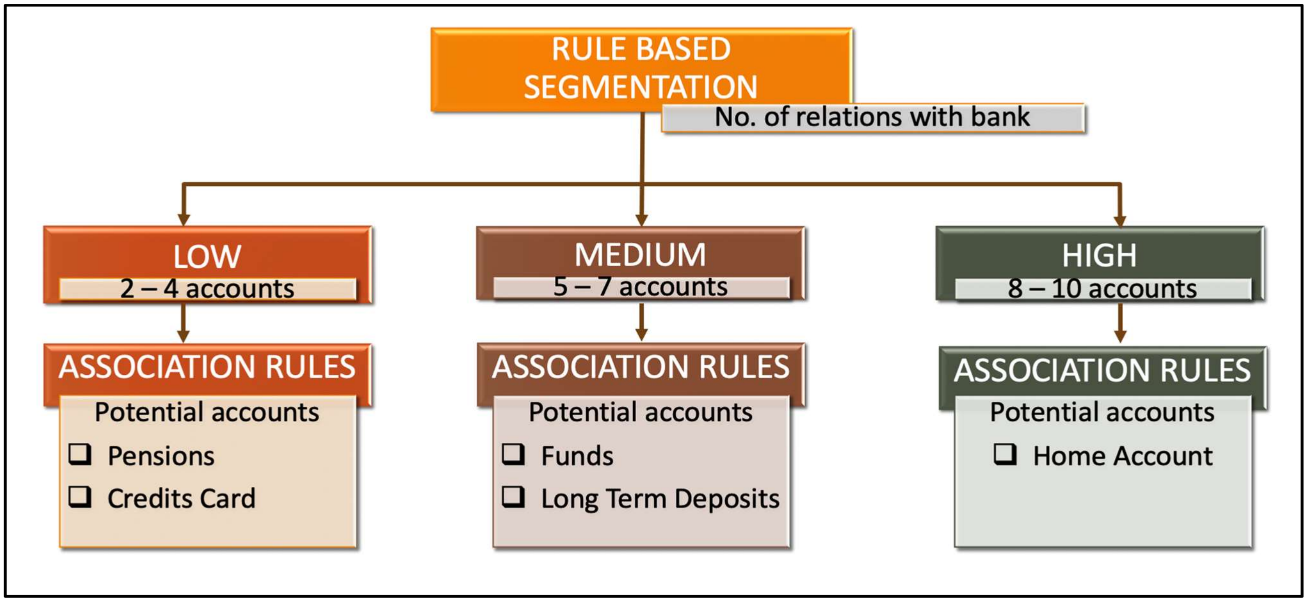


Figure 28: Top two potential products for each segment

11 Recommender Systems for Personalized Marketing

11.1 Collaborative Filtering for Implicit Feedback datasets

A common task of recommender systems is to improve customer experience through personalized recommendations based on prior implicit feedback. These systems passively track different sorts of user behavior, such as purchase history, watching habits and browsing activity, in order to model user preferences.

For the implicit feedback in product ownership, we will use the matrix factorization methodology using ALS to perform the recommendations. To perform this, we start with a USER-ITEM matrix R of size *u x i* with our users, items, which is required to be decomposed into matrices with users and hidden features of size *u x f* and one with items and hidden features of size *f x i.* In *U* and *V* we have weights for how each user/item relates to each feature as below:

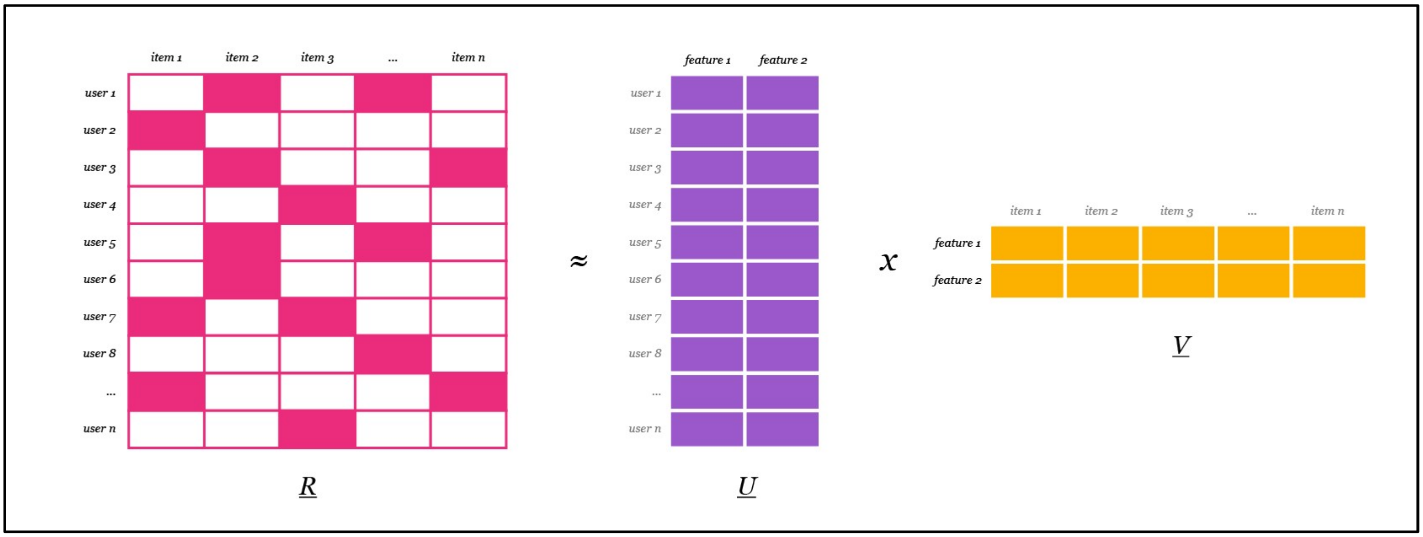


Figure 29: Latent Factorization of User-Item matrix

The implicit feedback is modelled as function of PREFERENCE and CONFIDENCE on the PREFERENCE that a user likes a product based on his purchase. We start with missing values, which imply user-account cells of the accounts that a user has never interacted with as a negative preference with a low confidence value and existing values which are the accounts a user owns a positive preference but with a high confidence value. The preference P, is set as a binary representation of our feedback data r. If the feedback is greater than zero we set it to 1:

$$p_{ui} = \begin{cases} 1 & r_{ui} > 0 \\ 0 & r_{ui} = 0 \end{cases}$$

As a next step, we calculate the confidence (c) on the preference (p) as below using the magnitude of r:

$$c_{ui} = 1 + \alpha r_{ui}$$

The rate of which our confidence increases is set through a linear scaling factor α. We also add 1 so we have a minimal confidence even if α x r equals zero*.* Now the next step is to find the vector for each user (xu) and account (yi) in latent factor dimensions in a way that it minimizes the loss function:

$$\min_{y_*, y_*} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda \left(\sum_u \| x_u \|^2 + \sum_i \| y_i \|^2 \right)$$

Where the first term is the sum squares of the difference between the actual interaction between a user and an account and the calculated rating as a product of the preference and confidence from the user and item factor matrices. The second term is the squared magnitude of the factor vectors multiplied by the regularization parameter to maintain the bias-variance balance and prevent overfitting. The presence of the X transpose time Y matrix term renders this function non-convex, so conventional optimization methods like Gradient descent will take a large amount of memory for fitting on the 0.68M x 24 sized matrix, so as suggested by the research paper, if we fix the user factors or item factors we can calculate a global minimum as one of terms being constant makes the objective function convex. The derivative of the above equation produces the following equation for minimizing the loss to produce the user factors is:

$$x_u = (Y^T C^u Y + \lambda I)^{-1} Y^T C^u p(u)$$

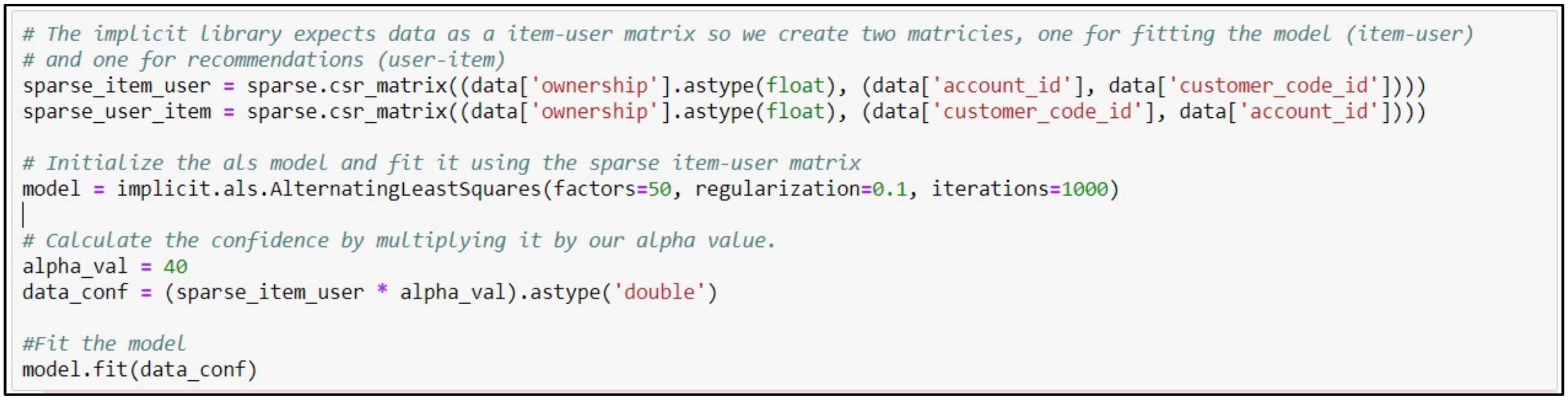
The loss function to be reduced to produce the factors for the accounts is:

$$y_i = (X^T C^i X + \lambda I)^{-1} X^T C^i p(i)$$

By iterating between computing the two equations above we create one matrix with user vectors and one with item vectors that we can then use to produce recommendations or find similarities. To make recommendations for a given user we calculate the dot product between our user vector and the transpose of our item vectors. This gives us a recommendation score for our user and each item which is later scaled to a 0 to 1 scale for sanity:

$$score = U_i \cdot V^T$$

We can iterate between users and items to identify the scores of every item for each user and rank them by order of preference and confidence. Then the top items are used to make recommendations to the user.
The training of the model using the IMPLICIT library's fit function, is done with 50 latent factors, the regularization factor which is the inverse of **λ** is chosen as 0.1, and the alpha value is taken to 40 as suggested in the paper.



```
# The implicit library expects data as a item-user matrix so we create two matricies, one for fitting the model (item-user)
# and one for recommendations (user-item)
sparse_item_user = sparse.csr_matrix((data['ownership'].astype(float), (data['account_id'], data['customer_code_id'])))
sparse_user_item = sparse.csr_matrix((data['ownership'].astype(float), (data['customer_code_id'], data['account_id'])))

# Initialize the als model and fit it using the sparse item-user matrix
model = implicit.als.AlternatingLeastSquares(factors=50, regularization=0.1, iterations=1000)

# Calculate the confidence by multiplying it by our alpha value.
alpha_val = 40
data_conf = (sparse_item_user * alpha_val).astype('double')

#Fit the model
model.fit(data_conf)
```

Figure 30: Collaborative Filtering for Implicit Feedback using ALS

From the produced recommendations, it can be seen that the model recommends most, some of the very popular products among the customers like the particular account, direct debits account and the e-account.

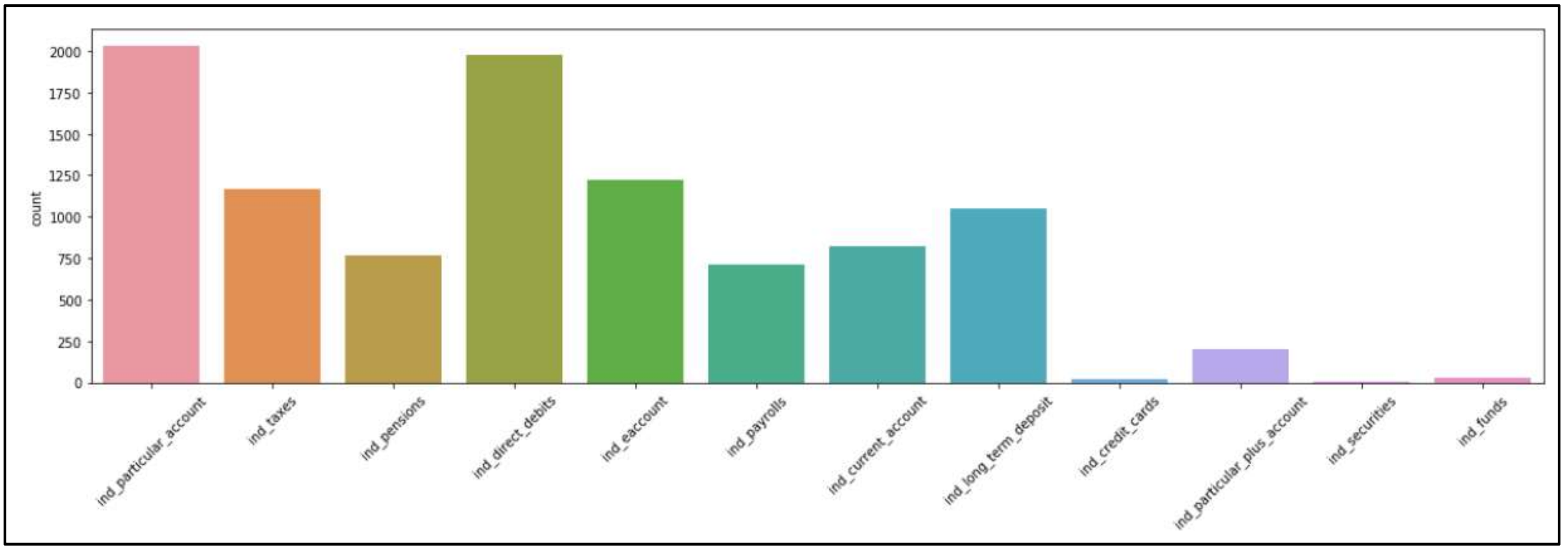
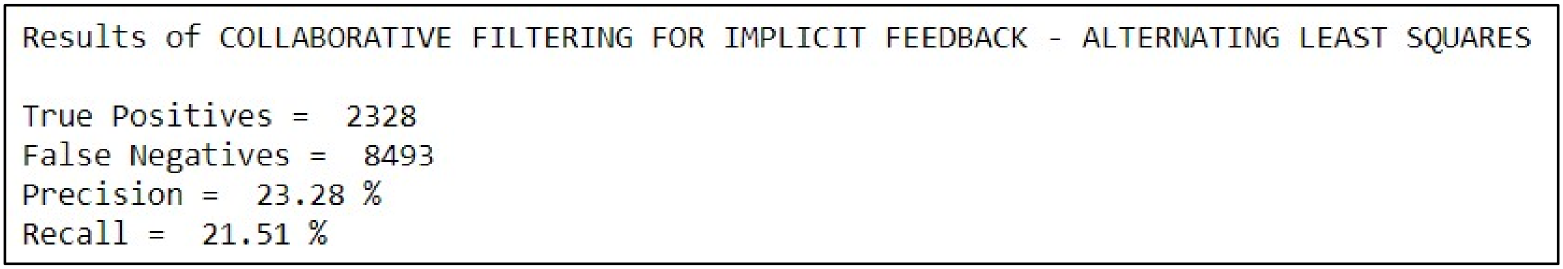


Figure 31: Top recommended products

Performing model validation by applying the top recommendation to each user for a sample of 10000 users and scoring the results against the last month data of the actual purchases by the customers produces below results:



```
Results of COLLABORATIVE FILTERING FOR IMPLICIT FEEDBACK - ALTERNATING LEAST SQUARES

True Positives =  2328
False Negatives =  8493
Precision =  23.28 %
Recall =  21.51 %
```

Figure 32: Collaborative Filtering (ALS) - Model Results

The precision and recall of the model are considerably low, with it only being able to capture 2328 of the 10000 purchases. We will also build other recommender system models and compare them on a similar scale by their performance on the score data.

11.2 User-User similarity-based CF

User based Collaborative filtering uses cosine similarity to find similar users to the target user and compares the products which both users own and recommends the product which target user doesn't have and will most likely buy due to the similarity in purchase behavior between the users.

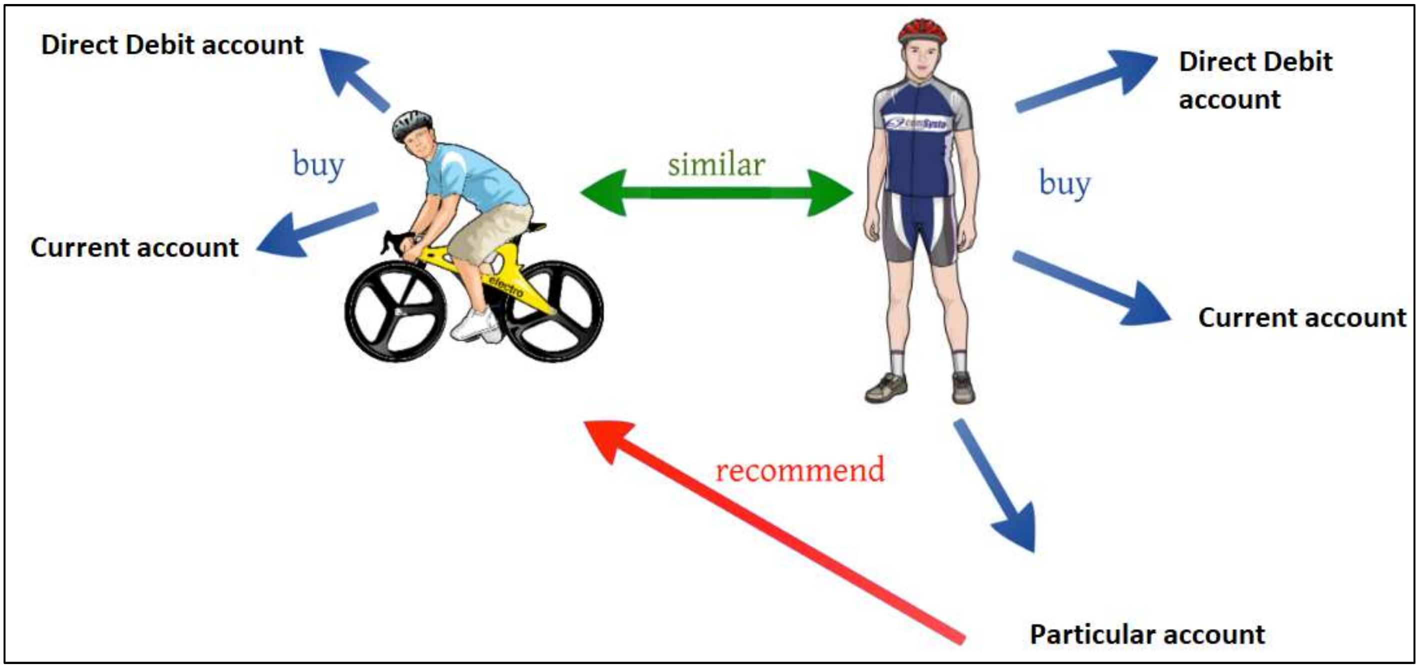
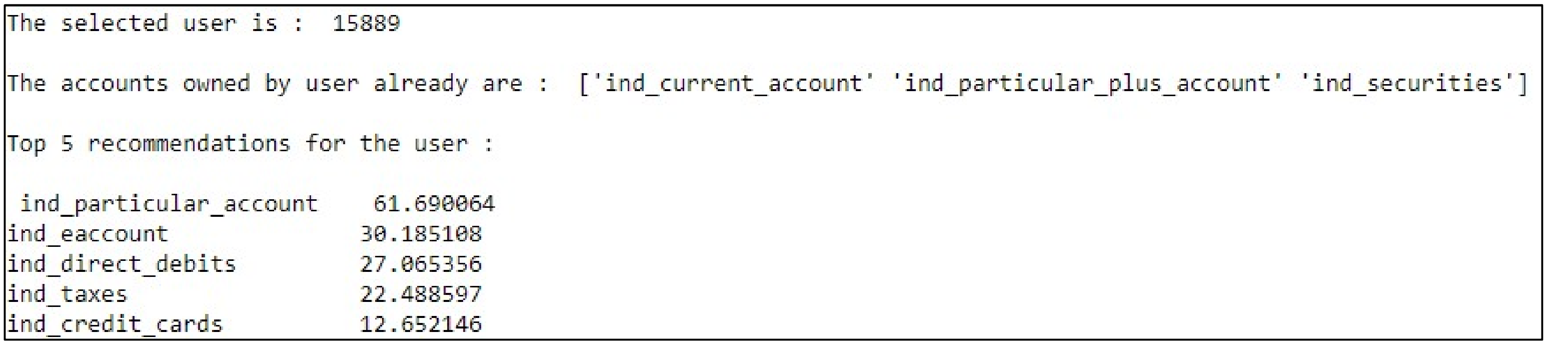


Figure 33: Example of USER-USER based CF

For Santander dataset, A data-item matrix is created for the one but last month data. Then, USER-USER cosine similarity matrix for each user pair is constructed. The system is tested for a single user first and then expanded to the entire dataset. For a single user, 500 most similar users and their products are identified. Considering the products the user already owns, the most similar items to the ones the user has liked from the neighborhood are identified. The top 5 most similar items are recommended to the user sorted by score. The below image shows the example for a selected user: 15889 who owns current account, particular plus and securities and the top 5 recommendations for the user are shown too.



```
The selected user is :  15889

The accounts owned by user already are :  ['ind_current_account' 'ind_particular_plus_account' 'ind_securities']

Top 5 recommendations for the user :

 ind_particular_account    61.690064
ind_eaccount              30.185108
ind_direct_debits         27.065356
ind_taxes                 22.488597
ind_credit_cards          12.652146
```

Figure 34: Top 5 recommendations for user 15889

This process is performed for all the users in dataset and the most recommended products are found out. From the produced recommendations, the model recommends most, some of the very popular products among the customers like the particular account, direct debits account and the e-account.

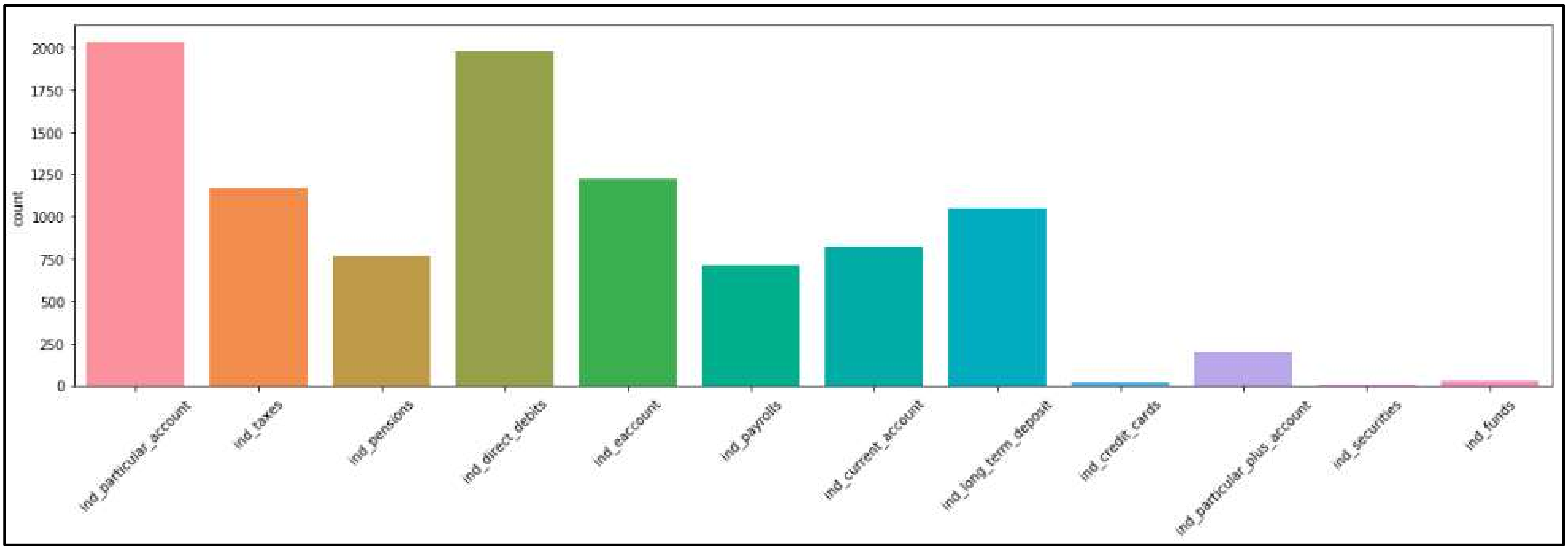
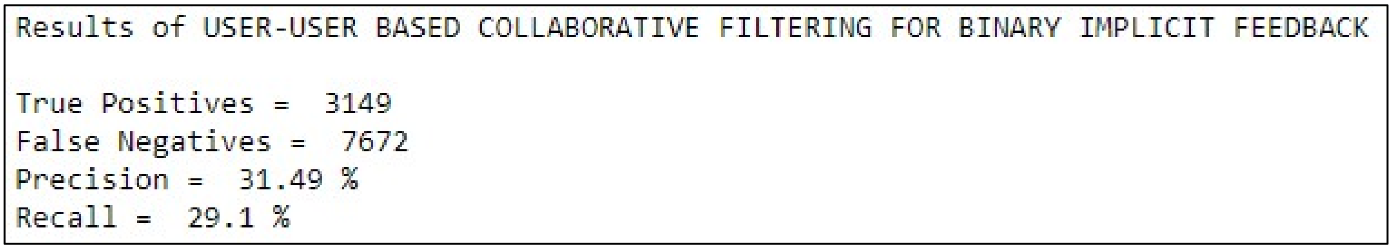


Figure 35: Top recommended products

Performing model validation by applying the top recommendation to each user for a sample of 10000 users and scoring the results against the last month data of the actual purchases by the customers produces below results:



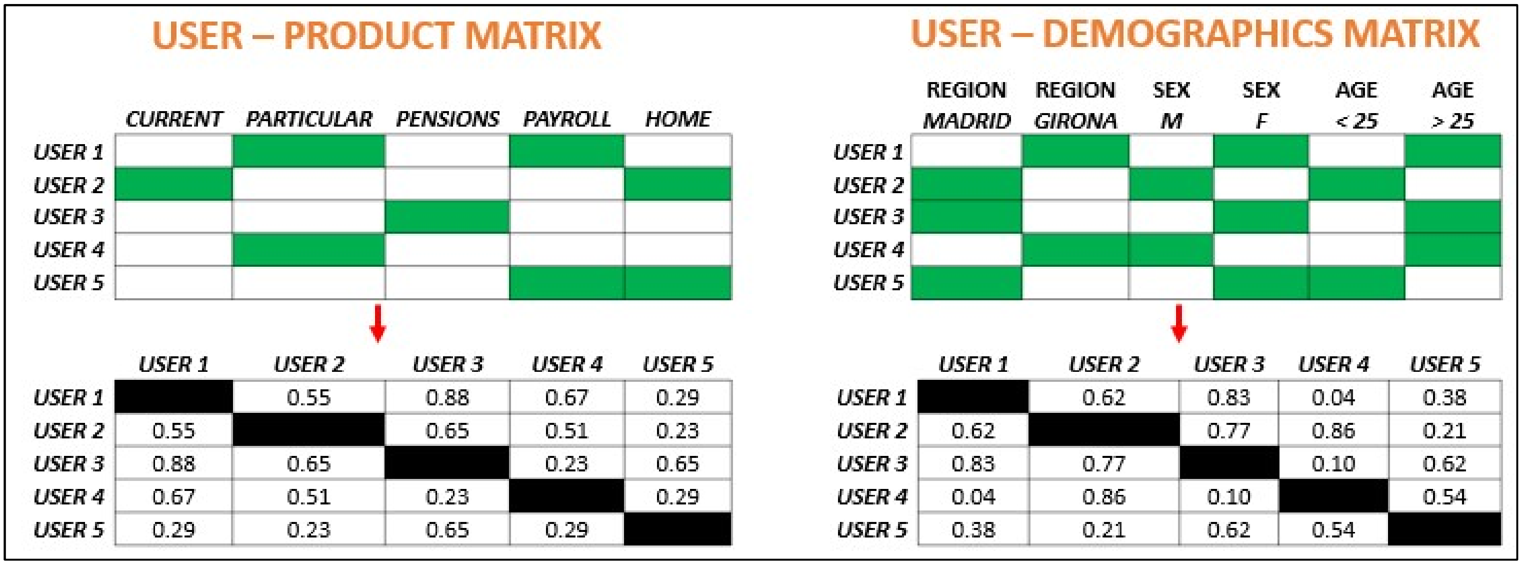
```
Results of USER-USER BASED COLLABORATIVE FILTERING FOR BINARY IMPLICIT FEEDBACK

True Positives =  3149
False Negatives =  7672
Precision =  31.49 %
Recall =  29.1 %
```

Figure 36: CF for Binary Implicit Feedback - Model Results

The precision and recall of the model are 30% more than CF-ALS model, with it being able to capture 3149 of the 10000 purchases. We will enhance this recommender system by incorporating demographic correlations.

11.3 User-User similarity-based CF enhanced by Demographic Correlations

Instead of focusing on users and products alone, we decided to add demographic features in the existing user-based CF model by using a hybrid algorithm that keeps the core idea of the existing User-based CF recommender system and enhances them with relevant information extracted from demographic data.
The following key demographic attributes were considered and one-hot encoded: sex, age bin, new customer index, seniority bin, foreign index, province name. User profiles were expressed as vectors constructed solely from demographic data and similarities among those user profiles were calculated for final predictions to be generated.



USER – PRODUCT MATRIX

	CURRENT	PARTICULAR	PENSIONS	PAYROLL	HOME
USER 1		🟩		🟩	
USER 2	🟩				🟩
USER 3			🟩		
USER 4		🟩			
USER 5				🟩	🟩

	USER 1	USER 2	USER 3	USER 4	USER 5
USER 1	■	0.55	0.88	0.67	0.29
USER 2	0.55	■	0.65	0.51	0.23
USER 3	0.88	0.65	■	0.23	0.65
USER 4	0.67	0.51	0.23	■	0.29
USER 5	0.29	0.23	0.65	0.29	■

USER – DEMOGRAPHICS MATRIX

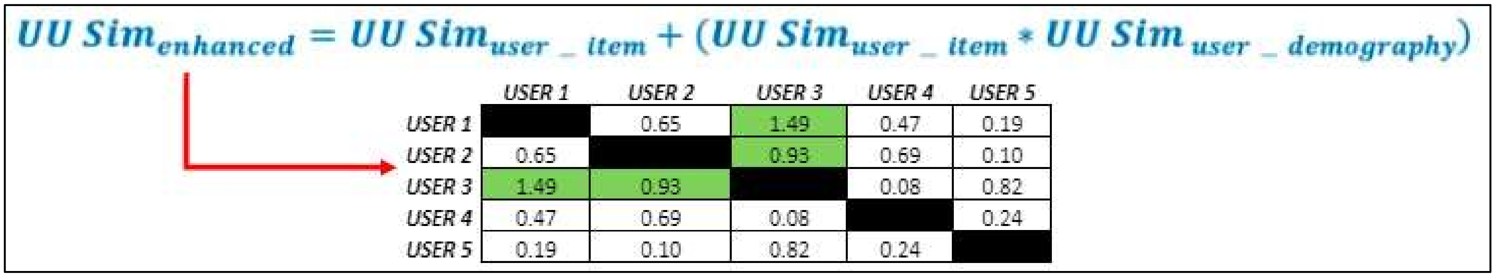
	REGION MADRID	REGION GIRONA	SEX M	SEX F	AGE < 25	AGE > 25
USER 1		🟩		🟩		🟩
USER 2	🟩		🟩		🟩	
USER 3	🟩			🟩		🟩
USER 4		🟩	🟩			🟩
USER 5	🟩			🟩	🟩	

	USER 1	USER 2	USER 3	USER 4	USER 5
USER 1	■	0.62	0.83	0.04	0.38
USER 2	0.62	■	0.77	0.86	0.21
USER 3	0.83	0.77	■	0.10	0.62
USER 4	0.04	0.86	0.10	■	0.54
USER 5	0.38	0.21	0.62	0.54	■

Figure 37: User-Product & Demographics Matrix

Demographic correlations between two users are defined by the similarity of the vectors which represent the specific users. That similarity is calculated by the cosine similarity of the two vectors. In the above image, the first matrix shows the calculation of similarities between users based on products and the second matrix shows similarities between users based only on demographics.

Using the above 2 user-user similarity matrices, enhanced similarity matrix is obtained using the below formula where UU Sim $_{user\_item}$ = User – Product matrix, UU Sim $_{user\_demography}$ = User- Demography matrix



$$UU\ Sim_{enhanced} = UU\ Sim_{user\ \_\ item} + (UU\ Sim_{user\ \_\ item} * UU\ Sim_{user\ \_\ demography})$$

	USER 1	USER 2	USER 3	USER 4	USER 5
USER 1	■	0.65	1.49	0.47	0.19
USER 2	0.65	■	0.93	0.69	0.10
USER 3	1.49	0.93	■	0.08	0.82
USER 4	0.47	0.69	0.08	■	0.24
USER 5	0.19	0.10	0.82	0.24	■

Figure 38: User-based CF enhanced by demographic correlations

- After the enhanced matrix is generated, 1000 nearest neighbors were identified.
- Items to be recommended were predicted as the weighted average of the preferences from each user's neighborhood.
- Items already owned by user were removed and the remaining recommendations were ranked by score.

From the produced recommendations, the model recommends most, some of the very popular products among the customers like the direct debits account, particular account and the taxes account.

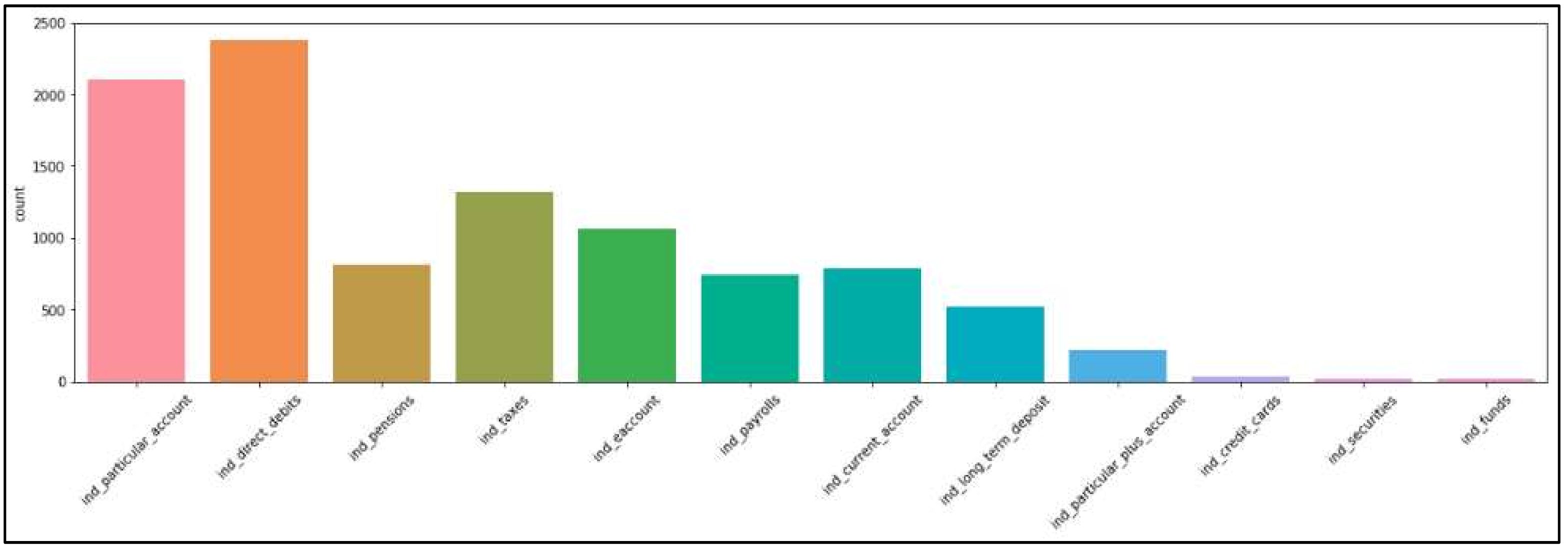
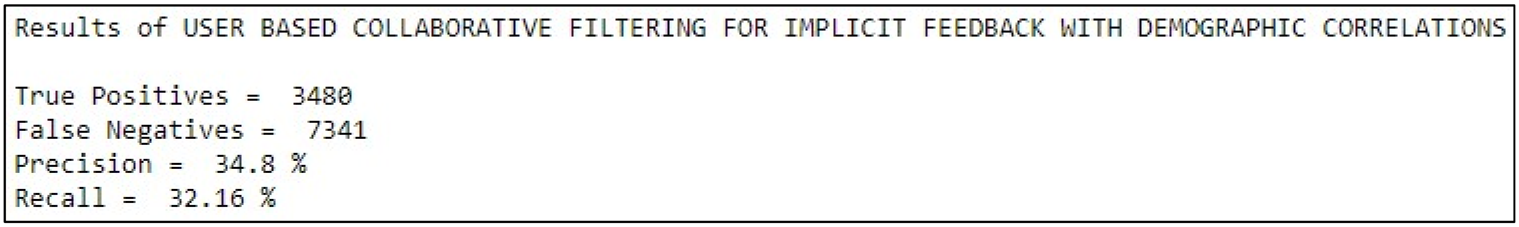


Figure 39: Top recommended products

Performing model validation by applying the top recommendation to each user for a sample of 10000 users and scoring the results against the last month data of the actual purchases by the customers produces below results:



```
Results of USER BASED COLLABORATIVE FILTERING FOR IMPLICIT FEEDBACK WITH DEMOGRAPHIC CORRELATIONS

True Positives =  3480
False Negatives =  7341
Precision =  34.8 %
Recall =  32.16 %
```

Figure 40: CF with demographic correlations - Model Results

The precision and recall of the model are the highest among all models, with it being able to capture 3480 of the 10000 purchases. There is a slight improvement in the metrics in this enhanced model over the base User-User model.

11.4 Model comparison

We have built three Collaborative filtering-based recommender systems of which two are enhanced and one baseline model to compare the uplift. Now we will perform model comparison based on their performance on the score data to select one for implementation in Santander production. To compare the models, we will use @K metrics because in the context of recommendation systems we are interested in recommending top-N items to the user and expect the recommendation to turn into conversion. So, the evaluation is valid to compute precision and recall metrics in the first N items instead of all the items. Thus, we've tended to precision and recall at k, where k is a user definable integer to match the top-N recommendations objective:

- Precision @ K: Precision at k is the proportion of recommended items in the top-k set that are relevant i.e. that a user ends up purchasing. It's interpreted as, if precision at 5 in a top-5 recommendation problem is 65%. This means that 65% of the recommendations made are relevant to the user.

  Precision@k = (# of recommended items @k that are relevant) / (# of recommended items @k)

- Recall @ K: Recall at k is the proportion of relevant items found in the top-k recommendations. It's interpreted as, if recall at 5 is found to be 45% in our top-5 recommendation system, it means that 45% of the total number of the relevant items are captured in the top-k recommended items.

  Recall@k = (# of recommended items @k that are relevant) / (total # of relevant items)

The recommendation is done on a score cut off basis, so if the top recommended account falls below the cut off, there is no recommendation offer made. So, if there are no items recommended. i.e. number of recommended accounts at k is zero, we cannot compute precision at k since we cannot divide by zero. In that case we set precision at k to 1. This makes sense because in that case we do not have any recommended account that is not relevant. Similarly, when computing Recall@k, when the total number of relevant items is zero, i.e. if a customer never purchased a new product in the next month, we set recall at k to be 1. This is because we do not have any relevant/purchased account that is not identified in our top-k results. So, the final Precision and Recall is calculated as the average of individual Precision and Recall metrics per customer.

Owing the business scenario in a bank, there is a high chance of customer fatigue if the number of offers/number of times a relationship manager reaches out to a customer with an offer to take up a new engagement is high, causing customer satisfaction indexes to fall or worst case, customer churn, so we set K=3, as the maximum applicable for the use-case, but iterate from K=1 to K=5 to compare the model performance.

PRECISION at K metrics for each model:

	K=1	K=2	K=3	K=4	K=5
IMPLICIT FB LATENT FACTORIZATION BY ALS	0.233	0.2238	0.183	0.15	0.143
USER BASED CF FOR BINARY IMPLICIT FB	0.315	0.28	0.232	0.209	0.188
USER BASED CF WITH DEMOGRAPHIC CORRELATION	0.344	0.289	0.242	0.21	0.189

RECALL at K metrics for each model:

	K=1	K=2	K=3	K=4	K=5
IMPLICIT FB LATENT FACTORIZATION BY ALS	0.22	0.44	0.51	0.55	0.66
USER BASED CF FOR BINARY IMPLICIT FB	0.29	0.53	0.64	0.77	0.866
USER BASED CF WITH DEMOGRAPHIC CORRELATION	0.32	0.54	0.67	0.78	0.87

Before choosing a final model, we can see that the Latent Factorization model optimized using ALS performs poorest on the Santander data. This could be because of the implicit feedback being only unary which makes the modelling of the product ownership data into preference and confidence no different from using the raw data in itself as both Ru-i and Cu-I terms are into [0,1] binary scale. Also, it's evident that the introduction of the demographic correlations improves on the baseline USER-USER similarity collaborative filtering model, so there exists an effect between factors like the user's personal profile and location of residence with the preference of banking products purchased by them. So the user similarity based collaborative filtering enhanced by demographic correlations recommender system model is chosen as the final model as it's able to capture almost 70% of the relevant accounts for the users within the allowable limit of K=3. To put this into perspective, there are 0.68M customers in the Santander bank's database and only 23K of them ended up purchasing an additional account in the score data month, so a recall of 70% is considerably good for production implementation.

Choice of an ideal K represents these metrics as inputs for a typical elbow graph problem to identify the trade-off. We can see the elbow is sharpest at K=2, however we can choose the ideal K=3 for production implementation of the recommender engine as in the bank's scenario the cost of the false positive (which indicates a recommendation not ending up in a purchase) is much lower than a false negative (which indicates the failure to predict an actual made purchase).

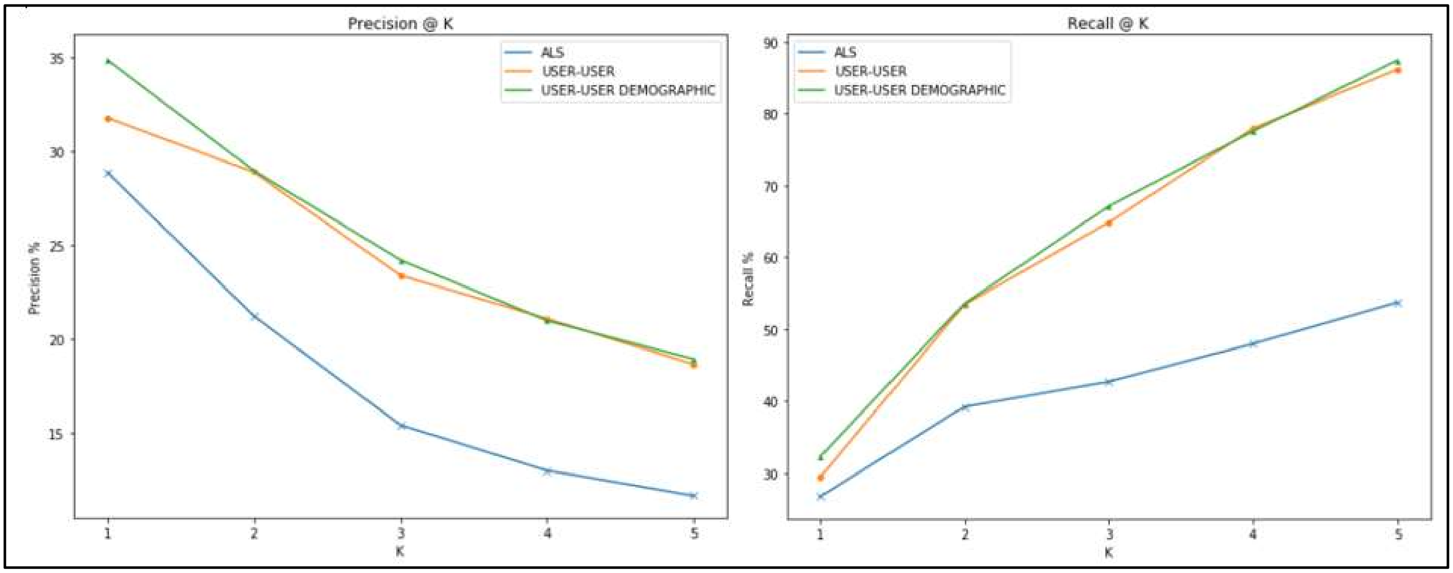


Figure 41: Elbow graph for PRECISION and RECALL @ K

12 Survival Analysis

Survival Analysis was performed on the data to check and compare the time until the survival probability of top selling accounts dropped below 80% (considered as a threshold in this case).

According to Harvard Business School report, on average a 5% increase in customer retention rates results in 25% – 95% increase of profits. Therefore, to improve the retention rates of Santander customers, we performed survival analysis to help Santander management analyze survival probability of the products and take responsible actions to reduce the percentage of churn.

12.1 Analysis using Kaplan Meier method

How Kaplan Meier method works for survival analysis?

In this analysis, Kaplan-Meier test was used to understand the cumulative survival probabilities and hazard rates over time for top accounts with Santander. Kaplan-Meier estimate was widely used in clinical test and is one of the best options used to measure the fraction of subjects living for a certain amount of time after treatment.

At each time interval and account, it is important to know

1. # Active Customers – Customers relation still active
2. # Stopped Customers – Customers who de-activated or stopped
3. # Censored Customers – Customers whose status is unknown or incomplete

Approach for calculating the Survival Curve -

1. Calculate hazard probability - the hazard at a given tenure is the number of customers who 'Stopped' divided by the total customers who are either 'Active' or 'Stopped'

   h(t) = # dead(t)/# at risk(t)

2. Calculate survival function - It is defined as

$$\prod_{j:t_j \leq t} [1 - \frac{\#\ \text{dead at time j}}{\#\ \text{at risk at time j}}]$$

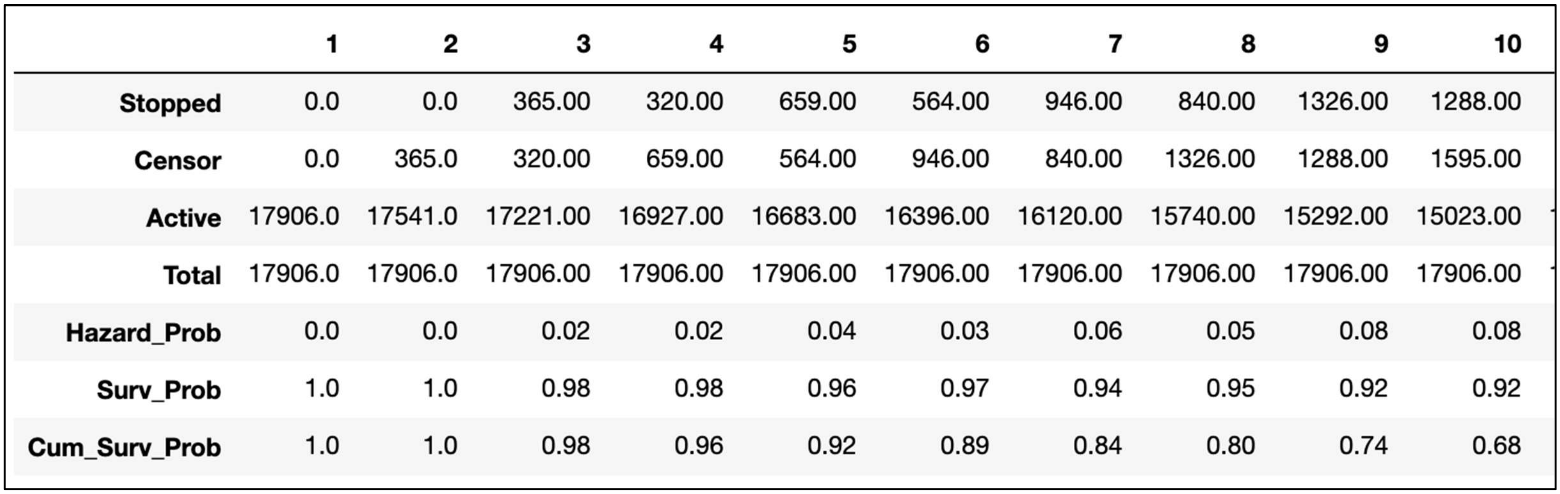
To arrive at this value, interval survival probability was first calculated as S(t) = 1 - h(t), S(t) at time t = 0 is assumed to be 1

Cumulative survival probability was then calculated as S'(t) = S'(t-1) * S(t), where S'(0) is considered 1

Methodology:

For testing our use-case, we performed survival analysis on sample data on only one province.

- Accounts considered – Current Account, Direct Debit, Particular Account, E-Account, Payroll Account
- Time Duration - Dataset consists of monthly updates of accounts activation and de-activation status of all customers over 17 months
- Data Manipulation – Manipulated data to create a matrix for each unique customer id and account tenure in months. Each account column represents tenure of that account for each customer. Tenure ranges from 0-17 months. 0 for an account implies customer never had an account and 17 implies customer was active for complete 17 months analysis period. The final matrix had 25,211 unique customers tenure data for 24 accounts.
- Survival Probability Matrix – Using hazard and survival probability function, calculated the below survival probability matrix for ‘current account’. Total number active customers (for current account) at least for 1 month are 17,906.



	1	2	3	4	5	6	7	8	9	10
Stopped	0.0	0.0	365.00	320.00	659.00	564.00	946.00	840.00	1326.00	1288.00
Censor	0.0	365.0	320.00	659.00	564.00	946.00	840.00	1326.00	1288.00	1595.00
Active	17906.0	17541.0	17221.00	16927.00	16683.00	16396.00	16120.00	15740.00	15292.00	15023.00
Total	17906.0	17906.0	17906.00	17906.00	17906.00	17906.00	17906.00	17906.00	17906.00	17906.00
Hazard_Prob	0.0	0.0	0.02	0.02	0.04	0.03	0.06	0.05	0.08	0.08
Surv_Prob	1.0	1.0	0.98	0.98	0.96	0.97	0.94	0.95	0.92	0.92
Cum_Surv_Prob	1.0	1.0	0.98	0.96	0.92	0.89	0.84	0.80	0.74	0.68

Figure 42:Event table for survival analysis

Assumptions:

- As the start date or the activation date of the account for a customer is not defined in the data, number of active months was considered to calculate tenure for a product in the given time period of 17 months.
- Few customers deactivated product in middle and again activated. This case is ignored in our analysis.

12.2 Results and Insights

The below survival curve was plotted using the survival probability matrix generated for each account.
It provides information about number of customers who survived for a certain amount of time or tenure. In this case, we considered 0.8 as threshold to understand and compare customers survival probability for an account.

- Particular account has 80% survival probability at the end of 14 months whereas direct debits reaches 80% survival probability at the end of 4 months only.
- 70% of the customers have in the selected province have Current accounts and it has 80% probability of surviving at the end of 8 months, 60% probability of surviving at the end of 12 months. Therefore, special attention should be given to customers after 12 months like sending offers or rewards to ensure retention.
- E-Account and Payroll account have similar survival curve with 80% of customers surviving at the end of 6 months.

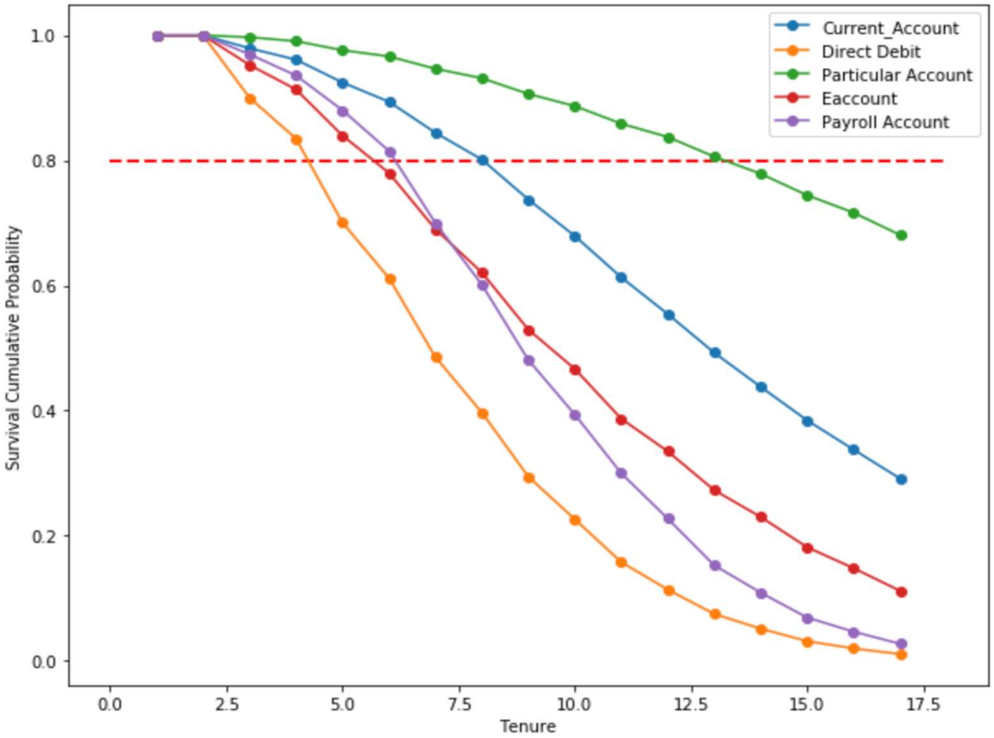


Figure 43: Cumulative survival probability curve for top 5 products

13 Findings and recommendations

13.1 Mass marketing strategies

Campaign management is an important task for marketing team in Santander. Mass marketing campaigns usually have less marketing budget as compared to other personalized campaigns. Therefore, to reduce costs of the marketing budget, same template e-mails/flyers/brochures are promoted for customers. The key point here is to send right product brochure to right customer. Segmentation followed by Market Basket Analysis is used to understand potential product ads to be sent to the customers.

In this case-study, we considered 2 business scenarios of mass marketing campaigns.

1) Marketing team has to monthly update website flyers on each customer's Santander website home page. Based on business decisions, the team strategically decides the potential products that should be updated for website flyers. In this case, we considered to push products that are least selling but have good potential in each segment (as discussed in section 10).

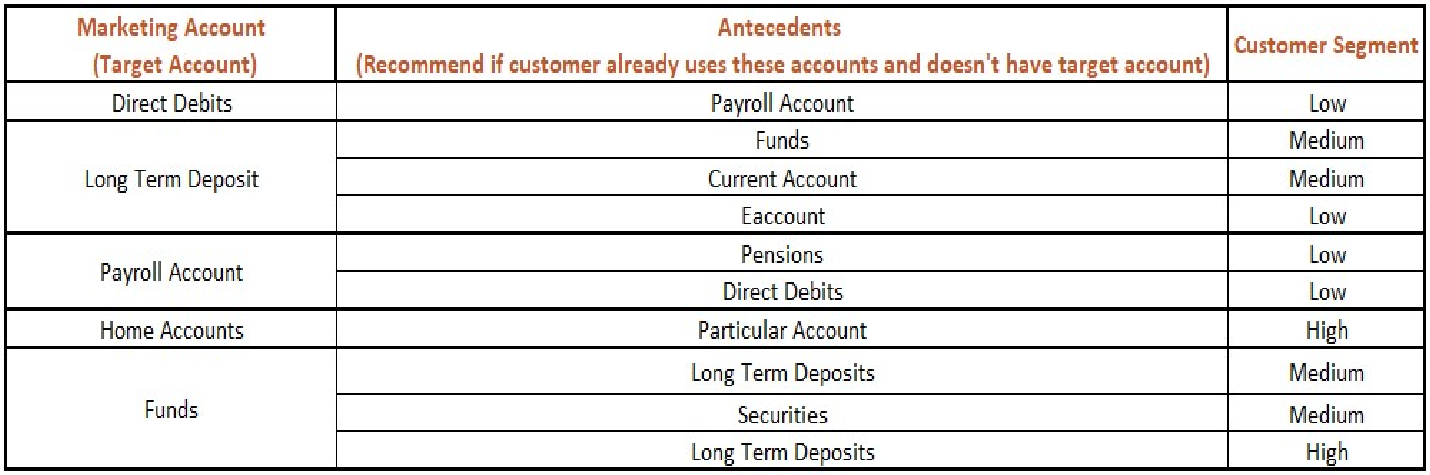
   Therefore, as an example for practical implementation the marketing team could update Pensions and Credit Cards Flyers for all customers in Low segment.

2) Santander Product Manager for Direct Debits decides to promote his product to increase number of accounts by 5% in Q2. In this scenario, the marketing team is aware of what product to market or campaign.

Potential customers were selected based on the following rule table. From the association rules generated for each segment, we first shortlisted all rules across segments which had RHS = 'Direct Debits". These rules were then sorted in the descending order of lift and support and top few significant rules were picked for each RHS or target account. The below table consists of rules for a sample of 5 target accounts.

As practical implementation, the targeted customers for ‘Direct Debits’ are those customers who belong to Low customer segment and have ‘Payroll Account’ but don’t have ‘Direct Debits’ (rule 1 in the below table).

Similarly, for other target accounts like Long Term Deposits, the significant rules are from both Medium and Low segments. Therefore, the targeted customers for ‘Long Term Deposit’ are those who belong to Medium Segment and have Funds/Current Account but don’t have Long Term Deposit and also customers who belong to Low segment and have e-account but don’t have Long Term Deposit.



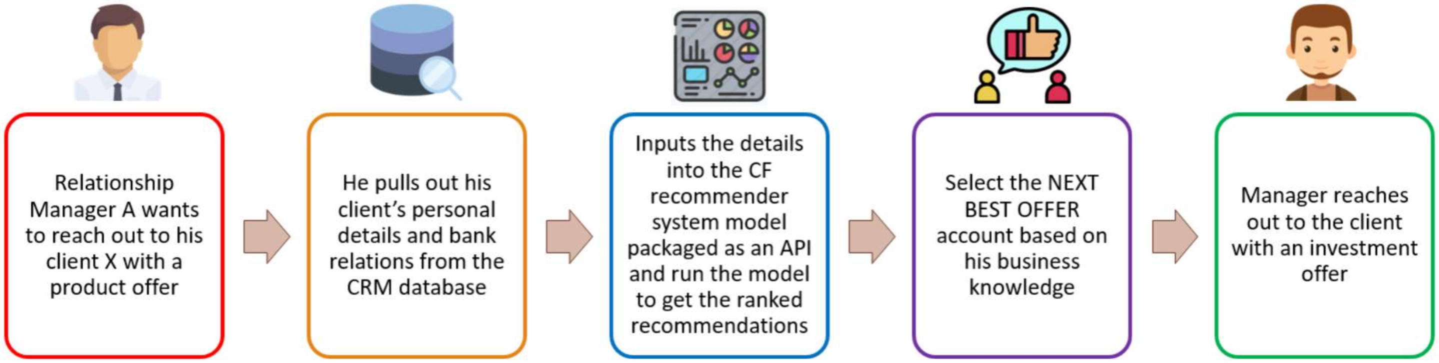
Marketing Account (Target Account)	Antecedents (Recommend if customer already uses these accounts and doesn't have target account)	Customer Segment
Direct Debits	Payroll Account	Low
Long Term Deposit	Funds	Medium
	Current Account	Medium
	Eaccount	Low
Payroll Account	Pensions	Low
	Direct Debits	Low
Home Accounts	Particular Account	High
Funds	Long Term Deposits	Medium
	Securities	Medium
	Long Term Deposits	High

Figure 44:Rules where target account appears on the RHS

13.2 Personalized marketing strategies

- Santander bank’s core business involves a client relationship manager guiding a client’s investment decisions
- The built recommender system model can aid relationship managers with making personalized and automated selection of next best products for private banking clients
- The chosen final model is wrapped onto an API and provided to the relationship managers for use from the front end

The process flow of making a recommendation to user is as below:



Relationship Manager A wants to reach out to his client X with a product offer

He pulls out his client’s personal details and bank relations from the CRM database

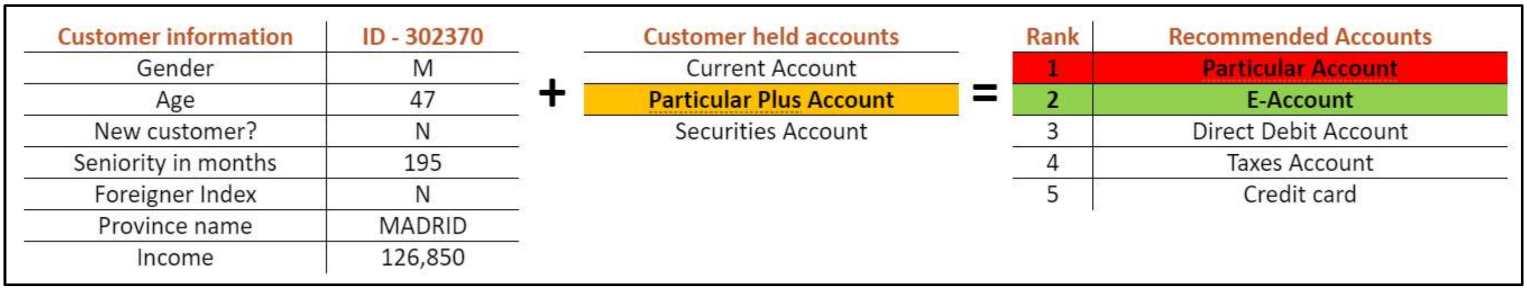
Inputs the details into the CF recommender system model packaged as an API and run the model to get the ranked recommendations

Select the NEXT BEST OFFER account based on his business knowledge

Manager reaches out to the client with an investment offer

Figure 45: Personalized Recommendation process flow

- The decision making is required to be assisted and not automatic, as the manager picks the top product from the model's results as there is a need to tailor the recommendations using business rules of applicability of an account for a user (For e.g., a junior account cannot be recommended to a customer of age 18+)
- Another reason for the above scenario is mentioned below, where the recommender system recommends the Particular account as the top for the given user's demographics and owned accounts, but the relationship manager can validate that the client already holds a Particular plus account, so it doesn't make sense to recommend a lower class account in Particular account which the client will never buy and is also a loss to the bank, so he/she can go for the next account in e-account to be the one to recommend to the client.



Customer information	ID - 302370
Gender	M
Age	47
New customer?	N
Seniority in months	195
Foreigner Index	N
Province name	MADRID
Income	126,850

+

Customer held accounts
Current Account
🟧 Particular Plus Account
Securities Account

=

Rank	Recommended Accounts
🟥 1	🟥 Particular Account
🟩 2	🟩 E-Account
3	Direct Debit Account
4	Taxes Account
5	Credit card

Figure 46: Recommender system - Business use case

14 References

i. VOZALIS, M. and MARGARITIS, K. (2007). Using SVD and demographic data for the enhancement of generalized Collaborative Filtering. *Information Sciences*, 177(15), pp.3017-3037.

ii. USER-USER Collaborative filtering Recommender System in Python. (2019). Retrieved from https://medium.com/@tomar.ankur287/user-user-collaborative-filtering-recommender-system-51f568489727

iii. Hu, Y., Koren, Y. and Volinsky, C. (2008). Collaborative Filtering for Implicit Feedback Datasets. *2008 Eighth IEEE International Conference on Data Mining*.

iv. *ALS Implicit Collaborative Filtering*. (2019) [online] Available at: https://medium.com/radon-dev/als-implicit-collaborative-filtering-5ed653ba39fe [Accessed 21 Apr. 2019].

v. *Recall and Precision at k for Recommender Systems*. (2019) [online] Available at: https://medium.com/@m_n_malaeb/recall-and-precision-at-k-for-recommender-systems-618483226c54 [Accessed 21 Apr. 2019].

vi. *Item-item collaborative filtering with binary or unary data*. (2019) [online] Available at: https://medium.com/radon-dev/item-item-collaborative-filtering-with-binary-or-unary-data-e8f0b465b2c3 [Accessed 21 Apr. 2019].