PRODUCT RECOMMENDATION FOR CUSTOMERS OF A BANK

BANA 8083 MS BANA Capstone

Abstract

Recommendation systems can enhance customer engagement by not only providing selective offers which can be highly appealing to the customer but also by adopting targeted marketing and advertising efforts towards potential customer segments and thereby achieving cost efficiency. The objective of this analysis is to look at customer purchasing behavior of financial products at a Bank and predict the new products that customers are likely to purchase thereby recommending those products to the customers. With a more effective recommendation system in place, the bank can better meet the individual needs of all customers and ensure their satisfaction. Different data mining classification algorithms are tried and compared to identify the best model for such a problem

Shreya Ghelani

(M10721450)

Under the guidance of:

Dr. Yan Yu and Dr. Edward Winkofsky

Table of Contents

1. [Introduction](#_Toc448352323) ……..2

2. [Data Source](#_Toc448352324) 2

3. Data Preparation ………………………………………………………………………………………………………………………………….2

3a. [Missing Data Analysis and Variable](#_Toc448352326) Elimination…………………………………………………………………………..2

3b. Data Imputation……………………………………………………………………………..…………….....………………………. 3

3c. Analysis of Product Popularity…………………………………………………………………………………………………….3

4. [Exploratory Data Analysis](#_Toc448352325) 4

4a. [Analysis of Customer Initiation by](#_Toc448352326) Month……………………………………………………………………………………4

4b. Analysis of Customer Age, Segments, Channel of Entry and Income …………….....……………………….5

4c. Analysis of Product Popularity……………………………………………………………………………………………………7

4d. Association Mining for a Market Basket Analysis of Products…………………………………………………….9

5. [Modeling and Performance Evaluation 10](#_Toc448352327)

5a. Multi Label Classification using [Binary Relevance 1](#_Toc448352328)1

5b. Multi Label Classification Using Random Forests [12](#_Toc448352330)

5c. Multi Label Classification Using Neural Networks [14](#_Toc448352331)

6. [Conclusion 1](#_Toc448352334)5

6. [Appendix](#_Toc448352335) 15

7. [Bibliography 1](#_Toc448352336)7

Figure 1 – Plot of Customer Income by City 3

Figure 2 – Histogram of Customer Age 4

Figure 3 – Plot of Customer Initiation into the Bank by Month 4

Figure 4 – Plot of Customer Age and Segment 5

Figure 5 – Plot of Customer Household Income and Segment 5

Figure 6 – Plot of Customer Age and Channel of Entry 6

Figure 7 – Plot of Customer Channel of Entry and Income 6

Figure 8 – Plot of Product Popularity 7

Figure 9 – Plot of Product Popularity by Segments 8

Figure 10 – Plot of Product Populaity by Sex 9

Figure 11 – Association Rules for Products 10

Figure 12 - Description of the Binary Relevance Problem Transformation Learner ……..11

Figure 13 – Description of the Binary Performance Measures for the Binary Relevance Leaner 12

Figure 14 – Description of the Random Forests Algorithm Adaptation Learner………………………………………………13

Figure 15 – Description of the Binary Performance Measures for the Random Forests Leaner………………………13

Figure 16 – Plot of Neural Networks Learner…………………………………………………………………………….……………………14

Figure 17 – Description of the Binary Performance Measures for the Random Forests Leaner………………………15

# Introduction

Today every industry is making use of recommender systems and predictive modeling with their own tailored versions. Banking is no different. The number of services and products that banks offer today to their customers have increased significantly in recent times. Recommendation systems are an approach towards personalization of banking services using prediction of anticipated customer purchasing behavior.

In a world strife with competition that is only a click away, offering your customers products and services best matched to their needs and preferences can go a long way towards gaining customer loyalty thereby making it a key business strategy. Predicting customer purchase behavior and recommending personalized products that have a high propensity of purchase can also prove to be a cost-effective strategy through targeted marketing. Customers are increasingly overwhelmed by marketing messages from web sites, emails, mobile apps, etc. Besides cost effectiveness for the organization, recommender systems also overcome the challenge of cutting through the noise and predict what customers are most likely to buy and make effective and useful recommendations.

Banks and financial institutions have always collected huge amounts of data, predominantly on customer profiles, investments, transactions, products and recently logs website clicks and app usage. This data is a gold mine to tap into for analytical use-cases and making intelligent decisions across marketing, risk management and operations.

# Data Source

The data for this analysis comes from Santander Bank which is a wholly owned subsidiary of the Spanish Santander Group. It is based in Boston and its principal market is the northeast United States. It offers products and services that cater to individual customers, Small and Medium Enterprises, institutions, students, non-profit and non-governmental organizations, etc. Under their current system, a small number of Santander Bank customers receive many recommendations while many others rarely see any, resulting in an uneven customer experience. The Bank wants to predict potential products that their new customers will use based on the product purchase and usage behavior of other similar customers. With a more effective recommendation system in place, the bank can better meet the individual demands of its customers.

Santander Bank provided this data to Kaggle which turned this problem into a Data Science Challenge. The bank has provided data about 1 lac customers. This data includes demographic information about the customer and a snapshot of his product purchase behavior for one month. Based on this data we predict the products a set of 10,000 new customers are likely to purchase in the next month. A customer can own/purchase more than one product which makes this a multi-label classification problem.

Features include age, sex, seniority, country and province of residence, customer segment etc. and the response variables are a set of 24 products that a customer either owns or does not own. The list of features and the description of the columns is provided in the Appendix section.

# Data Preparation

A couple of steps were performed to clean/prepare the data for further analysis. These include missing value treatments, character-value sanitization, data consolidation and splitting etc. The following steps were performed in particular –

1. Missing Data Analysis and Feature Elimination

The following columns had missing values. The variables “ult\_fec\_cli\_1t” (Last date as primary customer) and “conyuemp” (spouse index) are extremely sparse with more than 99% of the data missing. So, these two columns are dropped. Also, the columns “tipodom” (address type), “cod\_prov” (province code) and “segment” (customer segment) do not provide any meaningful additional information since we are already capturing the country and province information for customers. We discard these columns. The dimensionality of the data (feature space) is reduced by discarding seemingly unimportant customer information like “indrel\_1mes” (Customer type at the beginning of the month), “indext” (foreigner index) and “indfall” (deceased index) since these variables do not impact the product purchasing behavior of customers significantly.

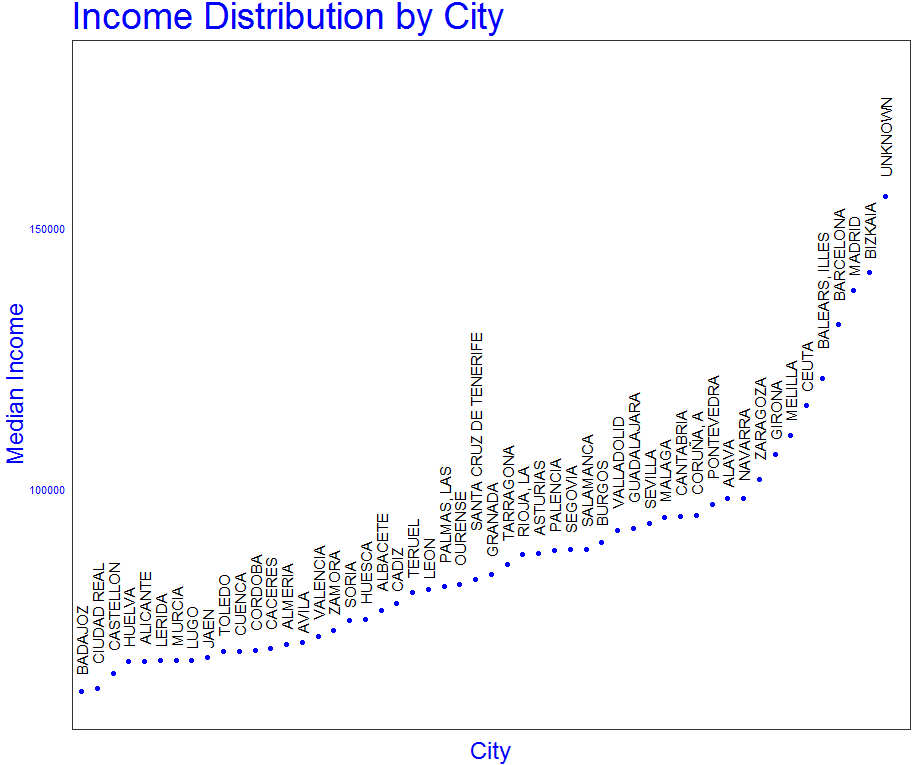
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ult\_fec\_cli\_1t | indrel\_1mes | tiprel\_1mes | conyuemp | canal\_entrada | cod\_prov | nom\_ov | renta |
| 97211 | 415 | 402 | 97376 | 675 | 403 | 403 | 24163 |

1. Data Imputation

Missing values for the columns “sexo” (Sex), “canal\_entrada” (channel of entry) and “nomprov” (province name) are assigned the value “UNKNOWN”.

Missing values for the column “tiprel\_1mes” (Customer relation type at the beginning of the month) are assigned the value “A” since that is the majority status (active).

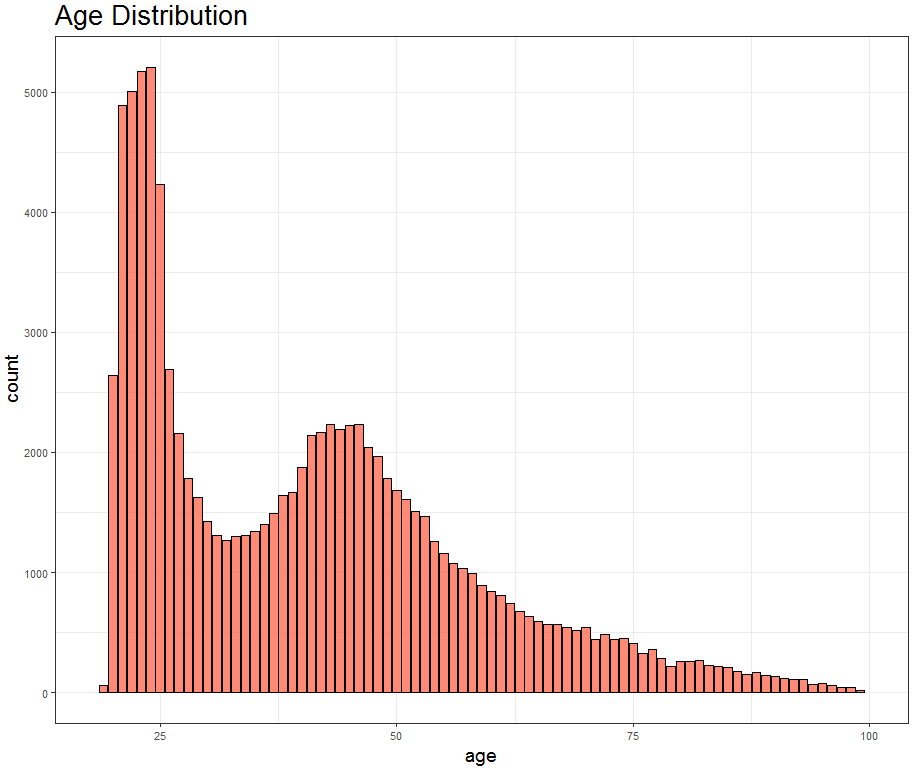
Missing values for the column “renta” (Income) are imputed by examining the income distribution per city and imputing the median income of the city the customer belongs to.



**Figure 1 - Plot of Customer Income by City**

1. Data Sanitization

The distribution of the column “age” is shown below. We can see that the distribution is bimodal with a lot of values around 25 and 50. It also has some values that do not make logical sense like people below the age of 18 and above the age of 100. These values are sanitized to replace the values below 18 with the median age between 18 and 30 and the values above 100 with the median age between 50 and 100.



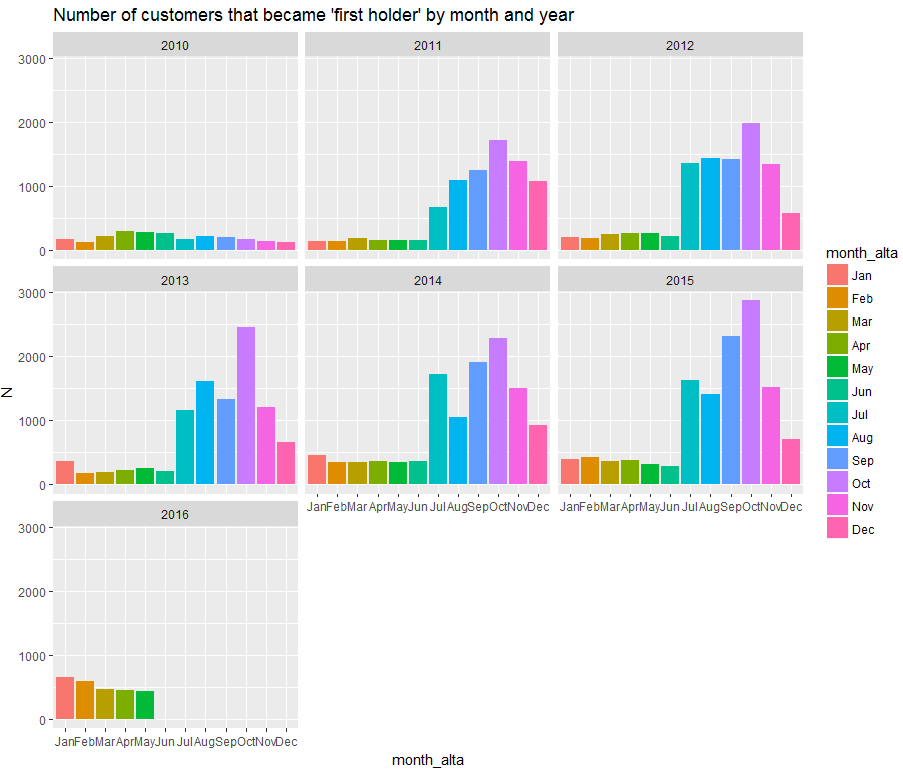
**Figure 2 –Histogram of Customer Age**

Some measures are performed to reduce the categorical values for “nomprov” (Province Name), “pais\_residencia” (Country of Residence) and “canal\_entrada” (Channel of Entry). There are 53 different provinces and 50% of the customers come from Madrid, Barcelona and Valencia. Therefore, I assign the province name for the remaining customers as ‘Other’. Similarly, there are 66 different nationalities and 97% of the customers come from Spain. Therefore, I assign the country to residence for the remaining customers as ‘Other’. Also, there are 143 different channels of entry and 70% of the customers come via ‘KHE’, ‘’KAT’ and ‘KFC’. Therefore, I assign the channel of entry for the remaining customers as ‘Other’.

# Exploratory Data Analysis

1. Analysis of Customer Initiation into the Bank by Month

We can see that there is a significant rise in July that remains until Autumn and the most number of holders are between September and October. this might be because July is the first month in Spain with vacations, and the academic calendar starts in September-October and is considered like a “new year” when people very prone to doing new things like opening a new bank account.

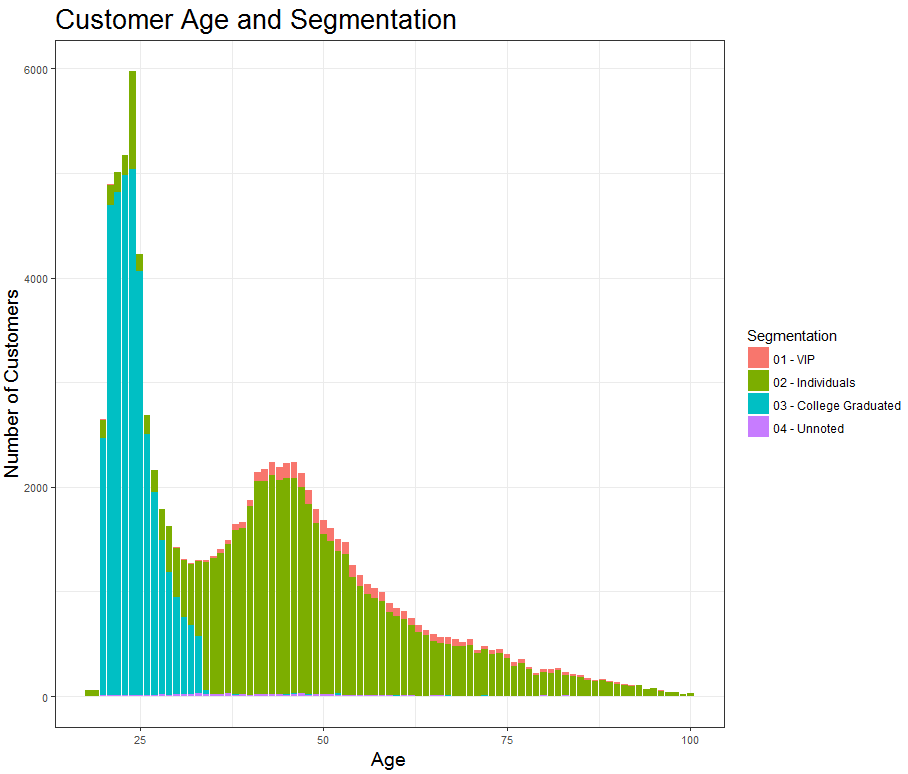


**Figure 3 - Plot of Customer Initiation into the bank by month**

1. Analysis of Customer Age, Segments, Channel of Entry and Household Income

Histogram of Customer Age and Segments

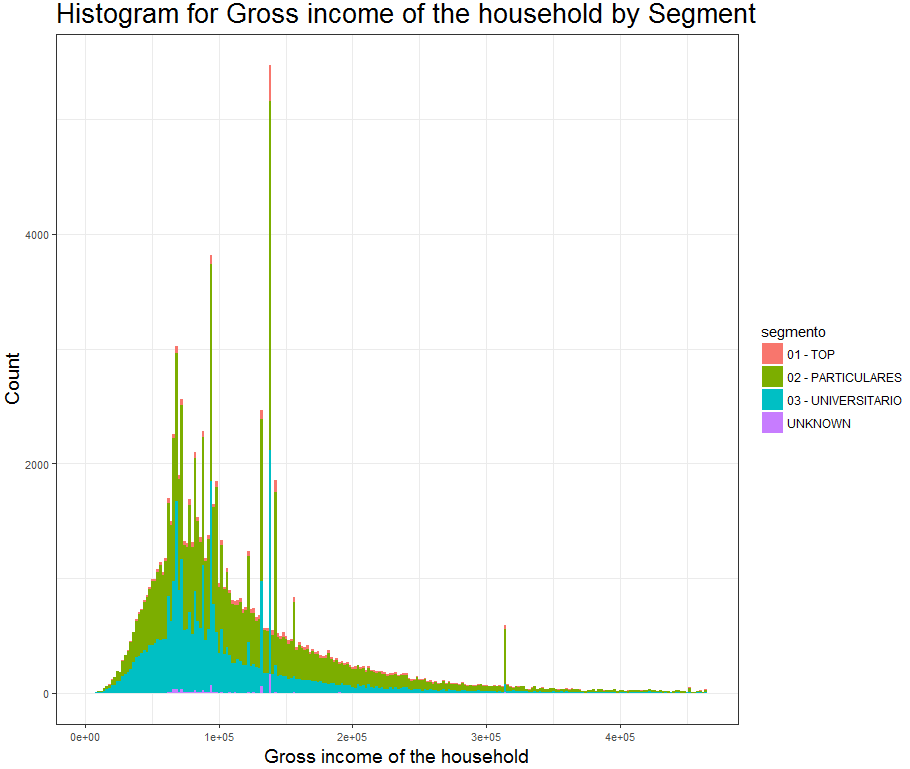
As can be seen from the plot, college graduates are young people whereas VIP and Individuals are middle aged which seems logical and expected.



**Figure 4 - Plot of Customer Age and Segments**

Plot of Household Income by Segments

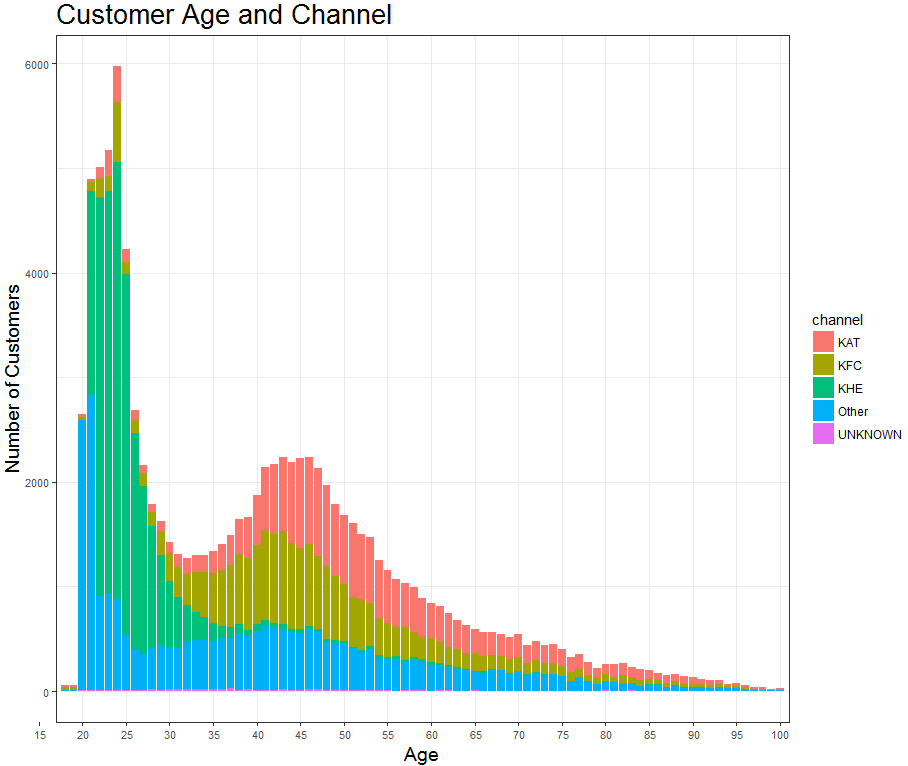
This plot also aligns with realistic expectations. VIP customers have the highest income, then individuals and towards the end, graduates.



**Figure 5 - Plot of Customer Household Income and Segments**

Plot of Customer Age and Channel of Entry

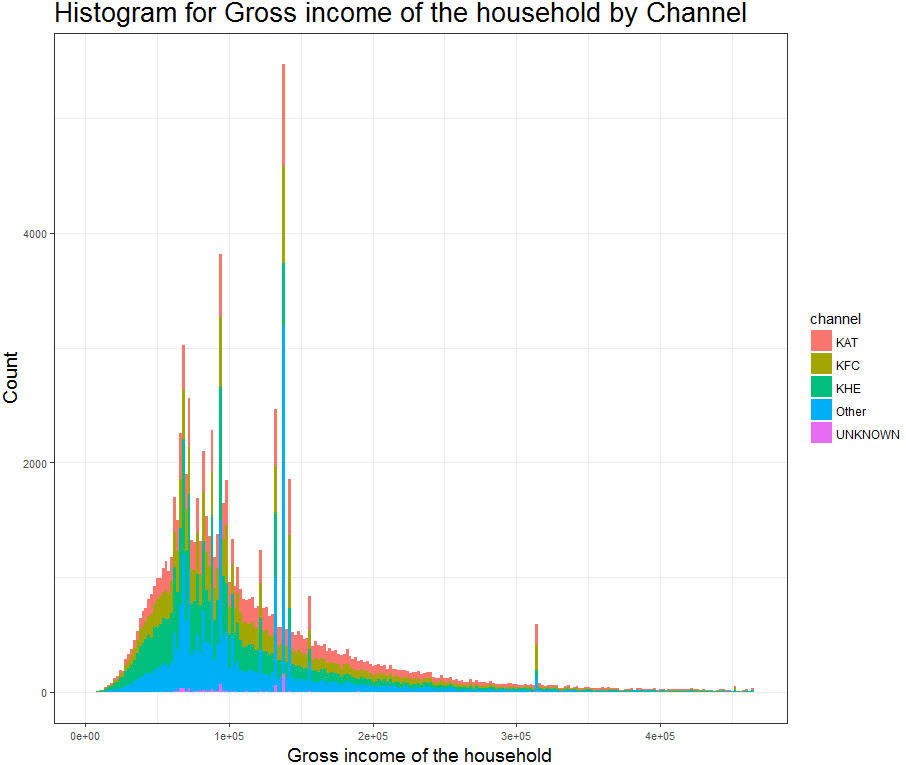
It can be inferred that the channel of entry “KHE” was mostly used by young university goers.



**Figure 6 - Plot of Customer Age and Channel of Entry**

Plot of Customer Household Income and Channel of Entry

It can be seen that the channel of entry is correlated with the gross household income. Customers with highest income have entered via the KAT channel whereas the customers entering via the channel KHE (university goers) have incomes lesser than those entering via KFC and KAT (mostly middle aged people belonging to the “Individuals and Top” classes). Thus, entry of channel, household income, age and segments seem strongly correlated.

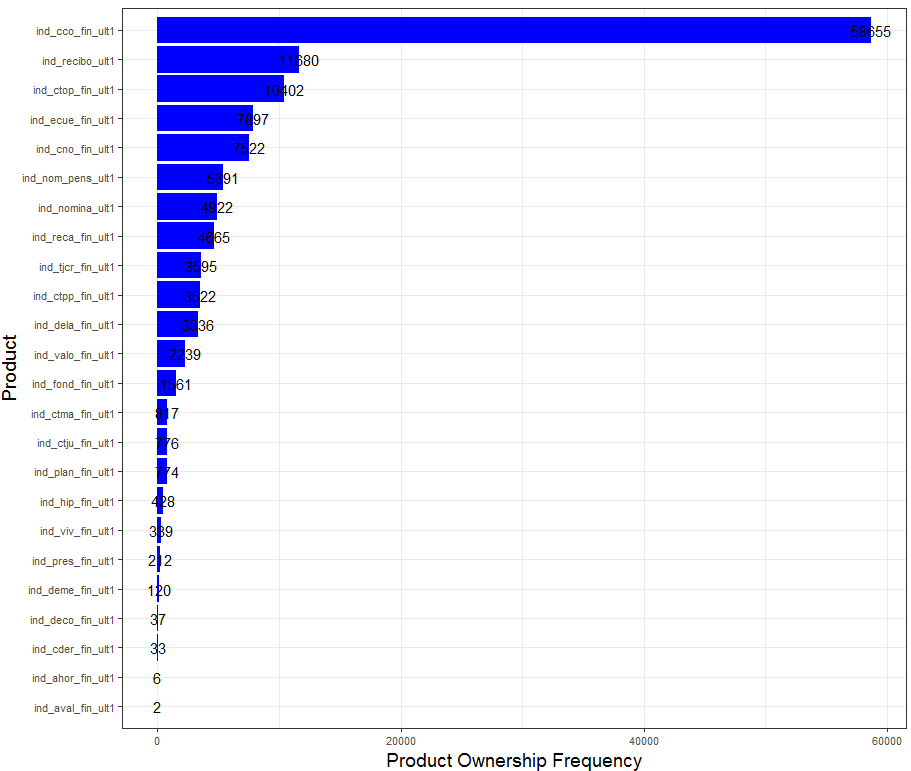


**Figure 7 - Plot of Customer Channel of Entry and Income**

1. Analysis of Popularity of the Products

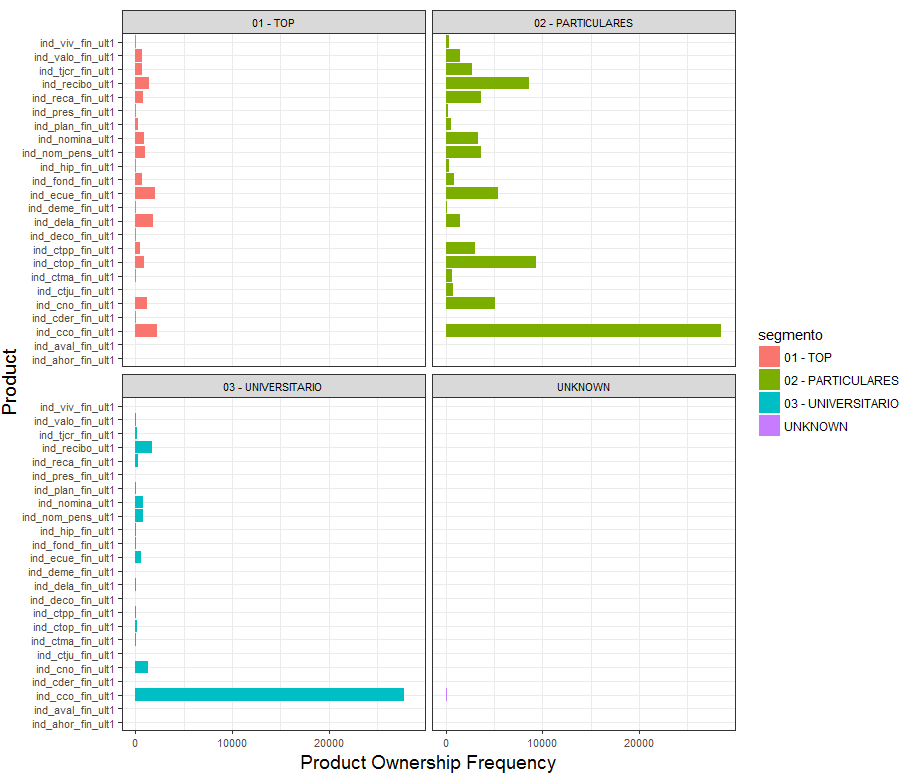
Plot of the most popular products

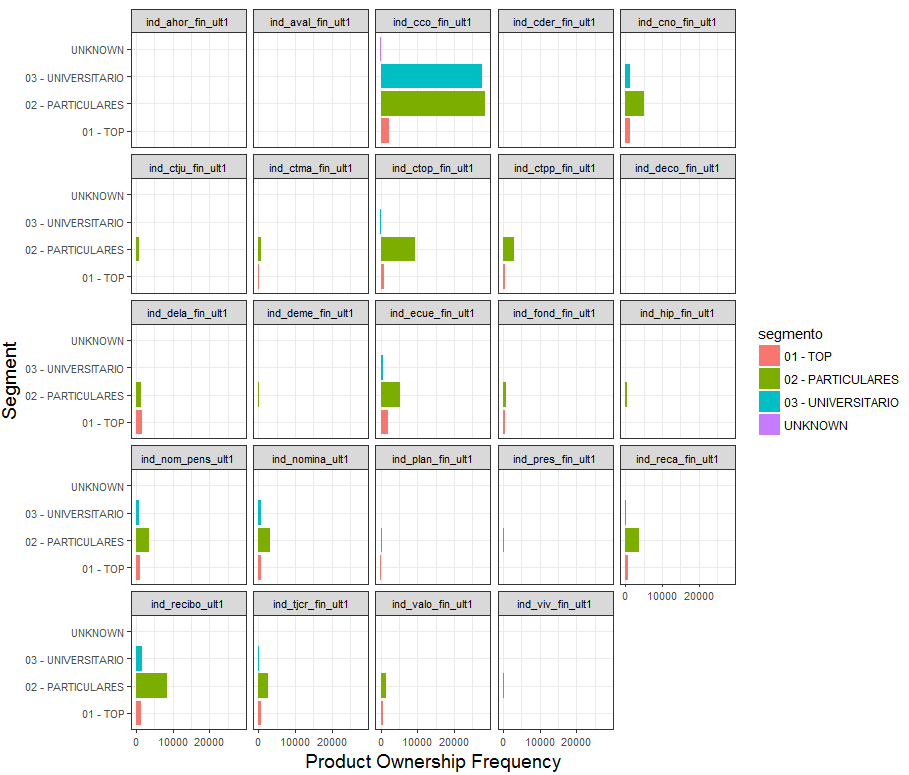
The bar chart below shows the number of customers that own a product. This gives us an indication of product popularity among customers. We can see that the product “ind\_cco\_fin\_ult1” (Current Accounts) has highest customer ownership (>50%) followed by the products “ind\_ctop\_fin\_ult1” (Particular Account) and “ind\_recibo\_ult1” (Direct Debit).



**Figure 8 - Plot of Product Popularity**

Plot of Product Popularity by Segments

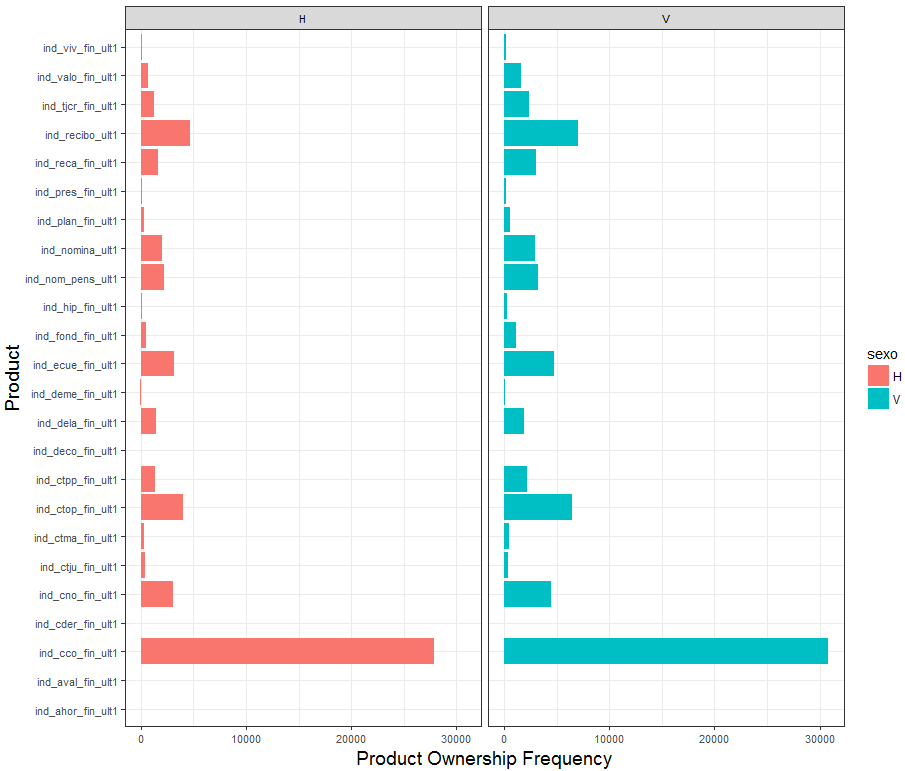


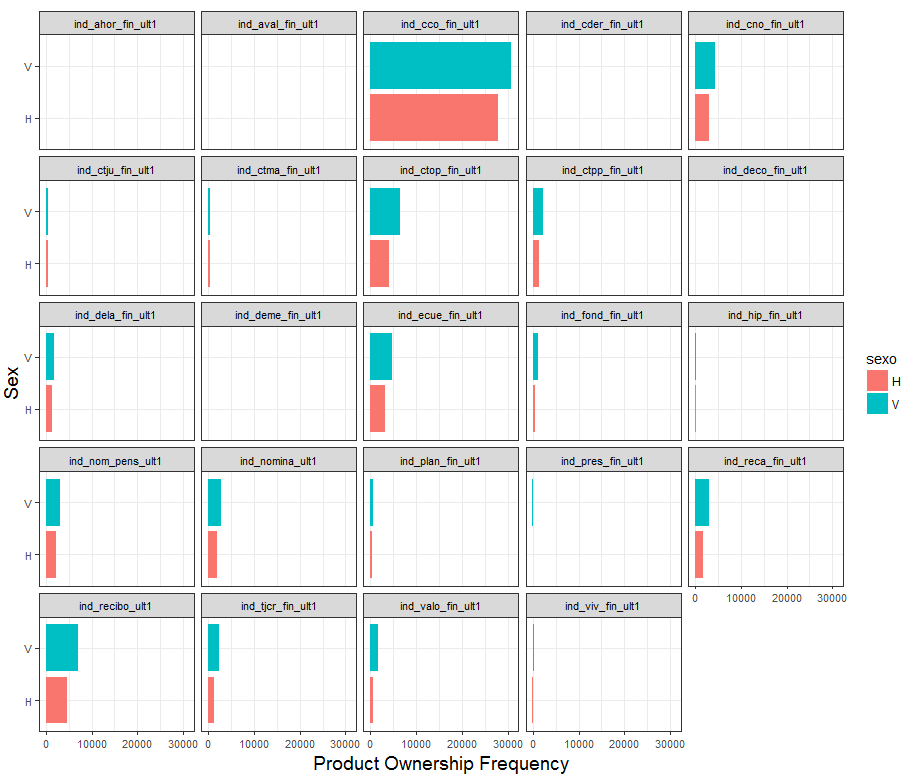


**Figure 9 - Plot of Product Popularity by Segments**

It is clear that the Current Accounts product is the most popular amounts University Students and the Particulars groups (both of which have a gross household income of below average and average). In the Top income bracket group, the products e-account, long-term deposits and direct debit are also popular along with current account. This is logical since this group has more capital to invest at their disposal. The direct debit, payroll and particular products are also popular with the Particulars group which falls in the middle-income bracket (maybe working class). The short-term deposits product is also particularly popular with the Particulars group.

Plot of Product Popularity by Sex





**Figure 10 - Plot of Product Popularity by Sex**

There is no significant difference between the products owned by males and females. The proportion of males is higher in the dataset and that is reflected in the plots.

1. Association Rule Mining and Market Basket Analysis of Products

Association mining is commonly used to make product recommendations by identifying products that are frequently bought together. The Apriori algorithm is used to generate the most relevant set of rules from a given transaction data set. A rule is a notation that represents which items are frequently bought together. Support, Confidence and Lift are three measures that are used to decide the relative strengths of the rules.

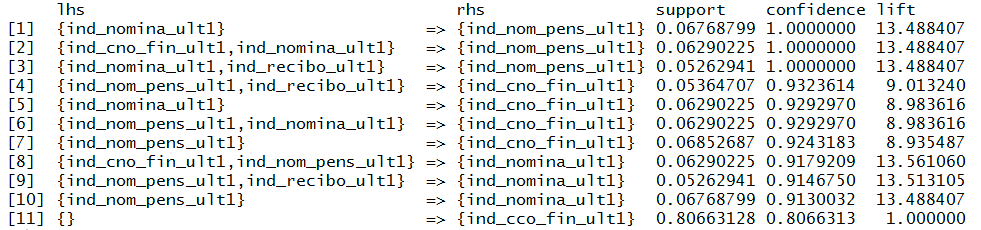
Consider, the rule A => B,

Support = , Confidence =

Expected Confidence = , Lift =

Lift is a factor by which the co-occurrence of A and B exceeds the probability of A and B co-occurring had they been independent. So, higher the lift, higher the chances of A and B occurring together.

Using a support argument of 5% and a confidence argument of 80%, the Apriori algorithms generates the following rules for our dataset. A confidence of 1 implies that whenever an LHS item was purchased, the RHS item was purchased 100% of the time. A rule with a lift of 13.5 indicates that the items in LHS and RHS are 13.5 times more likely to be purchased together compared to the purchases when they are assumed to be independent.



**Figure 11 – Association Rules for Products**

We can infer that, the product combinations of {Payroll, Pensions}, {Payroll Account + Payroll , Pensions} and {Payroll + Direct Debit , Pensions} have always been bought together. From the association rules, it is clear that the Payroll, Payroll Account, Pensions, Direct Debit and Current accounts products are most likely to be purchased together.

# Modeling and Performance Evaluation

The methods for multi-label classification can be grouped into two main categories (Ref: Multi Label Classification – An Overview) –

1. Problem Transformation Methods – Methods that transform the multi-label transformation problem into one or more single-label classification problems
2. Algorithm Adaptation Methods – Methods that extend specific learning algorithms in order to handle multi-label data directly

Evaluation Metrics for Multi Label Classification Problems –

1. Hamming Loss – Hamming Loss measures how many times on average, the relevance of an example to a class label is incorrectly predicted. It takes into account the prediction error(an incorrect label predicted) and the missing error(a relevant label not predicted), normalized over total number of classes and total number of examples.

It’s formula is given by: HammingLoss (xi,yi) =

where, |D| is the number of samples

|L| is the number of labels

yi is the true value

xi is the predicted value

1. Accuracy – Accuracy for each instance is defined as the proportion of the predicted correct labels to the total number (predicted and true) of labels for that instance. Overall Accuracy is the average across all instances.
2. Precision – Precision is the proportion of predicted correct labels to the total number of actual labels averaged over all instances.
3. Recall – Recall is the proportion of predicted correct labels to the total number of predicted labels averaged across all instances.
4. F1-Measure – F1 measure is nothing but the harmonic mean of precision and recall.

As in single label classification task, the higher the value of accuracy, precision, recall and F-1 score, the better the performance of the learning algorithm.

This analysis considers one Problem Transformation Method – The Binary Relevance Method and two Algorithm Adaptation Methods – Random Forests and Neural Networks.

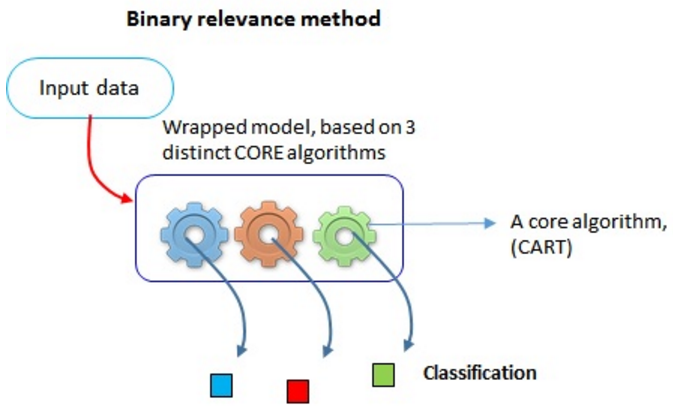
1. Binary Relevance Problem Transformation Technique for Multi Label Classification

The binary relevance problem transformation method converts the multilabel problem to binary classification problems for each label and applies a simple binary classifier on these. To classify a new instance, this method outputs as a set of labels the union of labels that are output by each simple binary classifier.

This binary classifier can be any of the widely-used classifiers – K-NN, Naïve Bayes, Decision Trees, Random Forests, Logistic Regression etc. In this analysis, we use the Classification RPART (Recursive Partitioning and Regression Trees).

The package “mlr” has an implementation of the Binary Relevance technique.

The below figure demonstrates how the Binary Relevance Method works -



**Figure 12 – Description of the Binary Relevance Problem Transformation Learner**

For the predictions, a threshold value of 0.5 is used.

The performance of the Binary Relevance Technique on the Training Data can be measured by the following metrics –

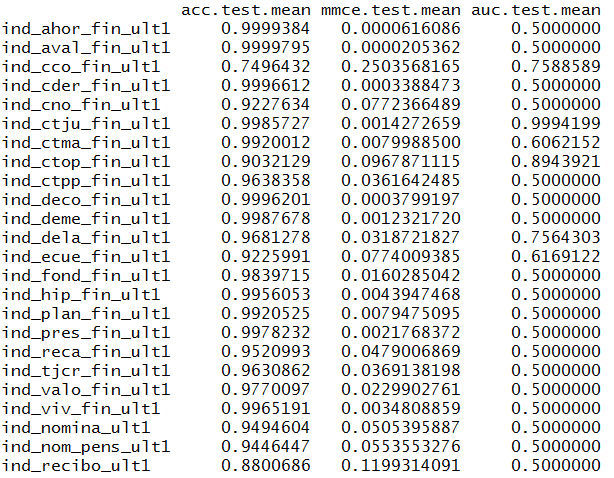
|  |  |  |
| --- | --- | --- |
| Hamming Loss | Accuracy | F1 |
| 0.0395 | 0.627 | 0.654 |

In addition to the multi label classification evaluation metrics, we can also see the following binary performance measures for each label –

Accuracy: mean (response == truth)

Area under the Curve or AUC: Integral over the graph that results from computing false positive rate and true positive rate for many different thresholds

Mean Classification Error or MMCE: mean(response != truth)



**Figure 13 - Description of the Binary Performance Measures for the Binary Relevance Leaner**

Upon performing 5-fold cross-validation, an aggregate measure of Hamming Loss was – 0.0396

1. Random Forests Algorithm Adaptation Technique for Multi Label (Ref: *Classification and Regression by randomForest*)

The Random Forests Algorithm Adaptation method extends the implementation of Random Forests to accommodate multi-label classification task. For this task, the definition of entropy is modified to be –

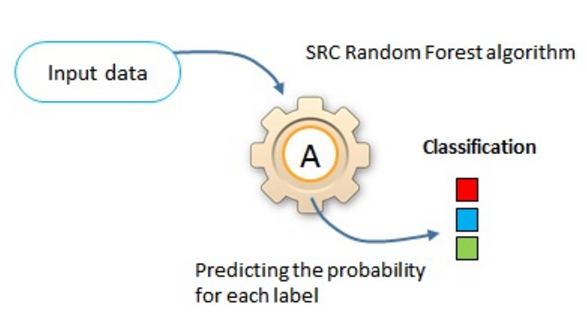
Entropy = -

where p(ci) = relative frequency of class ci and q(ci) = 1−p(ci).

In addition to this, multiple labels are allowed in the leaves of the tree.

The package “randomForestSRC” has a similar implementation and is used for this analysis.

The below figure demonstrates how the Random Forest method works -



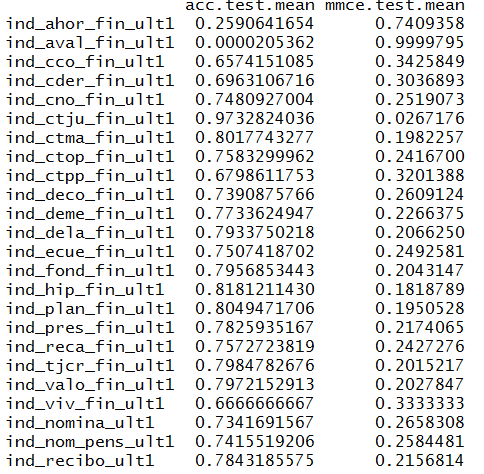
**Figure 14 – Description of the Random Forests Algorithm Adaptation Method Learner**

For the predictions, a threshold value of 0.5 is used.

The performance of the Random Forests Technique on the Training Data can be measured by the following metrics –

|  |  |  |
| --- | --- | --- |
| Hamming Loss | Accuracy | F1 |
| 0.0003 | 0.9947 | 0.9949 |

In addition to the multi label classification evaluation metrics, we can also see the binary performance measures Accuracy and Mean Classification Error for each label –



**Figure 13 - Description of the Binary Performance Measures for the Random Forests Leaner**

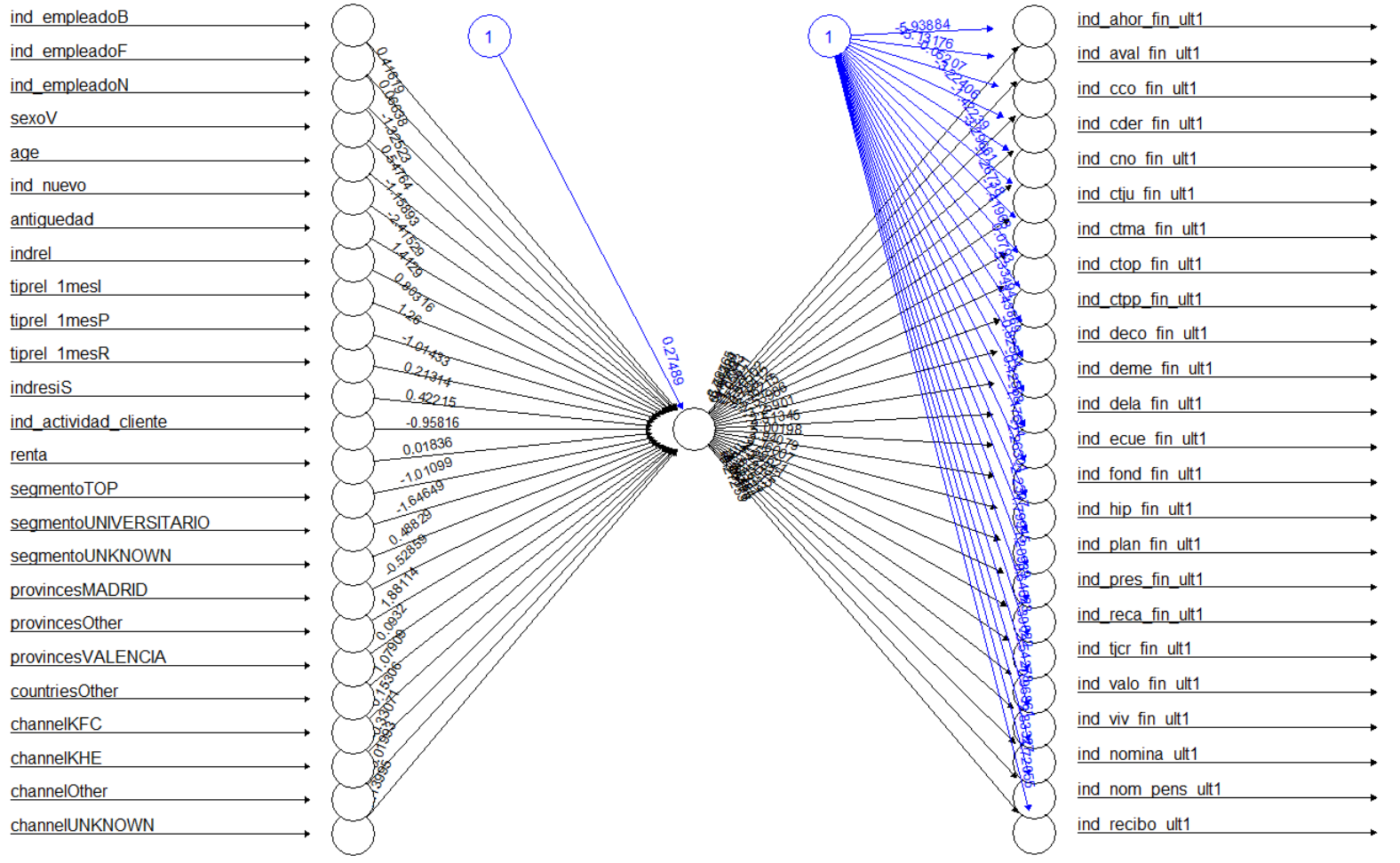
Upon performing cross-validation, an aggregate measure of Hamming Loss was – 0.2953

SRC random forest seems to be a very powerful algorithm, it can handle multiple target labels and resolve most missing value problems.

1. Neural Networks Algorithm Adaptation Technique for Multi Label Classification (Ref: Multi Label Classification with Neural Net Package am)

The BPMLL Algorithm – Back Propagation Multi Label Learner with multiple outputs is an implementation of neural networks in the multi label classification space. The package “neuralnet” in R is used for this analysis.

The activation function is set to Logistic. A visual-take on what is happening inside the model is shown below -



**Figure 14 – Plot of Neural Networks Learner**

The mean Accuracy achieved by Neural Networks is 0.9063.

Upon performing cross-validation, an aggregate measure of Accuracy was – 0.907, again no better than a random classifier.

# Conclusions

In Sample Performance (Training Data) – 70,000 customers

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Hamming Loss** | **Accuracy** | **F1** | **Hamming Loss (Cross Validation)** | **Accuracy (Cross Validation)** |
| **Binary Relevance Model** | 4% | 63% | 65.5% | 4% | 63% |
| **Random Forests Model** | 0.03% | 99% | 99% | 29.5% | 15% |
| **Neural Networks Model** | - | 90% | - | - | 91% |

Out Sample Performance (Testing Data) – 30,000 customers

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Hamming Loss** | **Accuracy** | **F1** |
| **Binary Relevance Model** | 4% | 63% | 66% |
| **Random Forests Model** | 0.1% | 97% | 96% |
| **Neural Networks Model** | - | 91% | - |

So, we can see that the performance of the Random Forests Algorithm Adaptation Technique is the best and that of the Binary Relevance Problem Transformation Method is the worst (slightly better than a random classifier) in this analysis. The Neural Network Algorithm Adaptation Technique performs well too. The performance of the learners is comparable in both the test and training data which shows that the techniques are robust and are not prone to overfitting.

# Appendix

List of variable names and their descriptions-

|  |  |
| --- | --- |
| **Variable 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 passive |
| 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) |
| indrel | 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month) |
| ult\_fec\_cli\_1t | Last date as primary customer (if he isn't at the end of the month) |
| 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) |
| tiprel\_1mes | Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential) |
| indresi | Residence index (S (Yes) or N (No) if the residence country is the same than the bank country) |
| 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 | Address type. 1, primary address |
| cod\_prov | Province code (customer's address) |
| nomprov | Province name |
| ind\_actividad\_cliente | Activity index (1, active customer; 0, inactive customer) |
| renta | Gross income of the household |
| segmento | segmentation: 01 - VIP, 02 - Individuals 03 - college graduated |
| ind\_ahor\_fin\_ult1 | Saving Account |
| ind\_aval\_fin\_ult1 | Guarantees |
| 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 |

# Bibliography

1. Liaw and M. Wiener (2002). *Classification and Regression by randomForest*. R News 2(3), 18—22

Grigorios Tsoumakas, Ioannis Katakis: Multi Label Classification – An Overview : Dept. of Informatics, Aristotle University of Thessaloniki, 54124, Greece

Guanwen Yao and Lifeng Cai : User-Based and Item-Based Collaborative Filtering Recommendation Algorithms Design

Jesse Read : Multi-label Classification : Universidad Carlos III de Madrid. Department of Signal Theory and Communications Madrid, Spain

Jinseok Nam 1 , 2 , Jungi Kim 1 , Eneldo Loza Menc ´ıa 1 , Iryna Gurevych 1 , 2 , and Johannes Furnkranz ¨ 1 : Large-scale Multi-label Text Classification — Revisiting Neural Networks : Department of Computer Science, Technische Universitat Darmstadt, Germany ¨ 2 Information Center for Education, German Institute for Educational Research, Germany

Min-Ling Zhang, Zhi-Hua Zhou: Ml-knn: A Lazy Learning Approach to Multi-Label Learning: National Laboratory for Novel Software Technology Nanjing University, Nanjing 210093, China

Miron Bartosz Kursa rFerns: An Implementation of the Random Ferns Method for General-Purpose Machine Learning : University of Warsaw

Tsung-Hsien Chiang, Hung-Yi Lo hungyi, Shou-De Lin : A Ranking-based KNN Approach for Multi-Label Classification : Graduate Institute of Computer Science and Information Engineering National Taiwan University

## [Xiaoyuan Su](https://www.hindawi.com/21346153/) and [Taghi M. Khoshgoftaar](https://www.hindawi.com/19583120/) : A Survey of Collaborative Filtering Techniques

Machine Learning in R : Multi Label Classification : [https://mlr-org.github.io/mlr tutorial/devel/html/multilabel/index.html#binary-performance](https://mlr-org.github.io/mlr%20tutorial/devel/html/multilabel/index.html#binary-performance)

Multi Label Classification with Neural Net Package : <https://www.r-bloggers.com/multilabel-classification-with-neuralnet-package/>