Unsupervised Clustering for Classifying Clients

Group 1: Luiz Carvalho, Manoj Soman Nair, Stanislav Taov

29/02/2020

Abstract

To better analyze and understand the selling effectiveness and efficiency, it is necessary to investigate similarities between customer profiles to group them by interests. The ability to identify properly the adherence of the product to the customer needs is viewed as critical to the correct classification of sales leads into selling categories of the sales calls.

Differences in classification accuracy are proposed as key to explaining variations in sales performance. The differences in accuracy are posited to result from the attributes believed to identify customer requirements, the quantitative levels associated with the attributes, and the degree of emphasis given to attributes in ascertaining client needs.

Implications for the sales process are tremendous, not only in terms of the effectiveness of the sales but also in terms of the sales effort.

Problem statement /Business problem

Based on the client profile identify the chances of success in reaching the client and once reaching it the chances of selling to it. Based on this percentage we would be able to classify the potential clients by “chance to be reached” and “chance to sell”. These two percentages would allow us to group these clients. This grouping would allow better management of the sales effort.

In summary, the objective is to identify patterns associating with the profile of the client with success in reaching and selling.

The product being sold is a credit card and the client base is the pre-existing list of people having bank accounts in a large bank in Brazil.

The bank doesn´t operate the sales itself. A third part company is sub-contracted to do the actual sales campaign through the phone. The database with the client’s details is forwarded to this company by the bank.

From the perspective of this third part company, the objective is how to maximize sales with the minimum sales effort.

To address this issue the classification of the clients in terms of their propensity to be reached and to buy the product is crucial.

In addition to that, the data exploration also gives insights about possible strategies of pre-processing the database which could eliminate the need for some types of calls.

Another important point linked with the process is the fact that our concern here is how to increase the sales effectiveness, we assume that every client in the list, theoretically can have the credit card. There is no negative-classification in the sense that some profiles cannot get the card. The idea is just to identify if a profile is likely/unlikely to purchase a card. A priori eligibility for having the card was done by the bank when preparing the mailing sent to the third part company. Therefore we are not faced here with eventual ethical dilemmas (See responsible AI in consumer enterprise).

Approach /Analytical problem

The first issue was to define which strategy we could use to group potential clients. It is important because the whole process is based on our capability of putting people in groups where we suppose they will behave similarly regards being accessible to the salesforce and regards actually buying the service.

We used ten parameters and cluster these people using k-means, assuming the possibility of having 150 clusters. The number 150 was obtained from the decision tree. Note that we are mixing together parameters linked with the person and parameters linked with personal behaviour.

Datasets/Getting the data

We used a real sanitized and anonymized dataset representing a subset of the clients operated by the third part company during the months of August, September and October of 2019, totalizing 343630 clients.

Data dictionary

We have several information regards the clients:

• Age

• Sex

• Income

• Zipcode

• if the phone was a mobile or fix-line.

• Credit score of the client

• If the client has an active relationship with the bank (number of interactions within the month).

• How many products of the bank the client already have • How many phone numbers the client has

• How many times the client was contact by telephone during the month

We also have the results of the campaign:

• If the person was in fact reached

• If the sales actually happened.

Data exploration

The data used contained 16 variables with 344627 observations. The variables included birth, sex, age, creditscor, rescen, city, typephone - landline or mobile number, tel_o, tel_a, enrri, incomerang, lat, lon, areacode, spoke, sold. The details can be seen below.

```
head(df)
```

```
##        birth sex age creditscor rescen             city typephone tel_o
tel_a
## 1 03/01/1979   M  40          5      2        SAO PAULO         C     7
7
## 2 31/07/1992   M  27          8      1        CAMPO BOM         C     3
4
## 3 07/09/1952   M  67          8      1 LAURO DE FREITAS         C     7
10
## 4 07/07/1989   F  30          2      1           GOIANA         C     3
3
## 5 21/03/1959   F  60          2      1         BRAGANCA         C     4
4
## 6 30/09/1949   M  69          4      2           RECIFE         C     5
6
##   enrri incomerang    lat    lon areacode spoke sold
## 1     0          A -18.46 -50.57       64     1    0
## 2     1          A -30.51 -51.49       51     0    0
## 3     3          A -12.58 -38.31       71     0    0
## 4     0          A  -8.17 -35.59       81     0    0
## 5     0          A  -1.27 -45.43       98     0    0
## 6     1          A  -8.17 -35.59       81     1    0
```

```
str(df)
```

```
## 'data.frame':    343627 obs. of  16 variables:
##  $ birth     : Factor w/ 21976 levels "01/01/1940","01/01/1941",..: 1483
21790 4826 4743 14570 21326 18091 18245 17847 5586 ...
##  $ sex       : Factor w/ 2 levels "F","M": 2 2 2 1 1 2 2 2 2 1 ...
##  $ age       : int  40 27 67 30 60 69 36 63 37 28 ...
##  $ creditscor: Factor w/ 11 levels "1","10","2","3",..: 6 9 9 3 3 5 4 9 1
9 ...
##  $ rescen    : int  2 1 1 1 1 2 2 1 1 1 ...
##  $ city      : Factor w/ 6789 levels "","120033","12263",..: 5883 1237
3355 2369 973 4985 5268 6229 3891 3687 ...
##  $ typephone : Factor w/ 2 levels "C","F": 1 1 1 1 1 1 1 1 1 1 ...
##  $ tel_o     : int  7 3 7 3 4 5 6 4 3 5 ...
##  $ tel_a     : int  7 4 10 3 4 6 6 4 3 5 ...
```

```
##  $ enrri     : int  0 1 3 0 0 1 0 0 0 0 ...
##  $ incomerang: Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ lat       : Factor w/ 69 levels "-0.59","-02. 3",..: 19 54 12 64 5 64
12 11 61 45 ...
##  $ lon       : Factor w/ 70 levels "-35.53","-35.59",..: 48 53 8 2 30 2 8
27 6 59 ...
##  $ areacode  : int  64 51 71 81 98 81 71 77 84 45 ...
##  $ spoke     : int  1 0 0 0 0 1 0 0 1 1 ...
##  $ sold      : int  0 0 0 0 0 0 0 0 0 0 ...
```

Let's make a copy of data for further analysis

```
df1 <- df
```

We will convert latitude and longitude columns into integers

```
suppressWarnings(df1$lat <- as.numeric(df1$lat))
suppressWarnings(df1$lon <- as.numeric(df1$lon))
```

Checking for missing values

```
sum(is.na(df1))
```

We discovered that there were 184161 missing values locating within creditscor, lat, lon and city columns. We are going to replace missing values for creditscor with mean. Missing values for city, lat and lon columns will be removed.

Replacing all missing creditscor values with the mean value of creditscor column.

```
df1$creditscor <- sapply(df1$creditscor, is.numeric)
df1$creditscor <- lapply(df1$creditscor, na.aggregate)
```

Dropping all missing values.

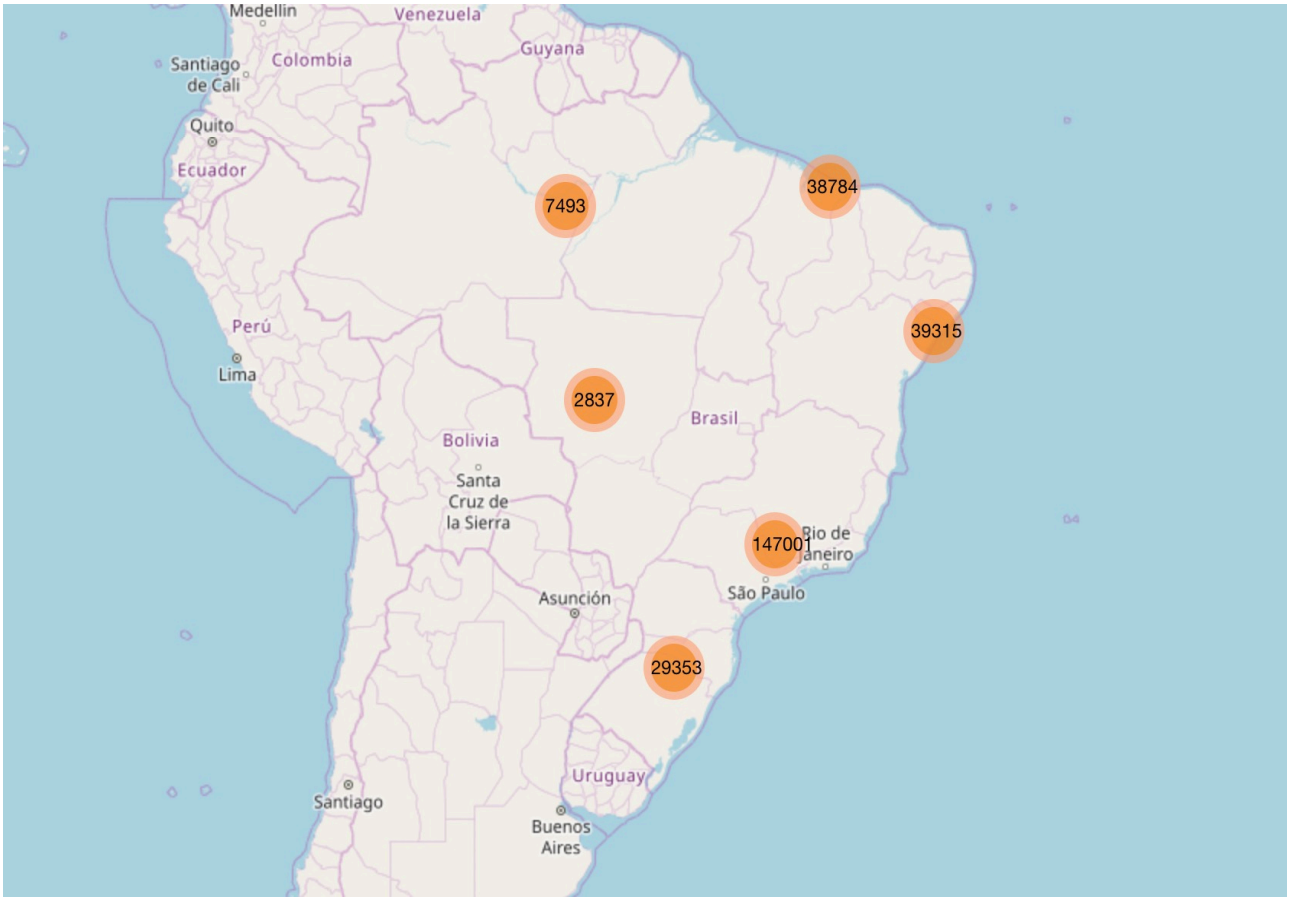
```
df1 <- na.omit(df1)
```

Let's check the frequency of city column

```
questionr::freq(df1$city, cum = TRUE, sort = "dec", total = TRUE)
unique(df1$city)
```

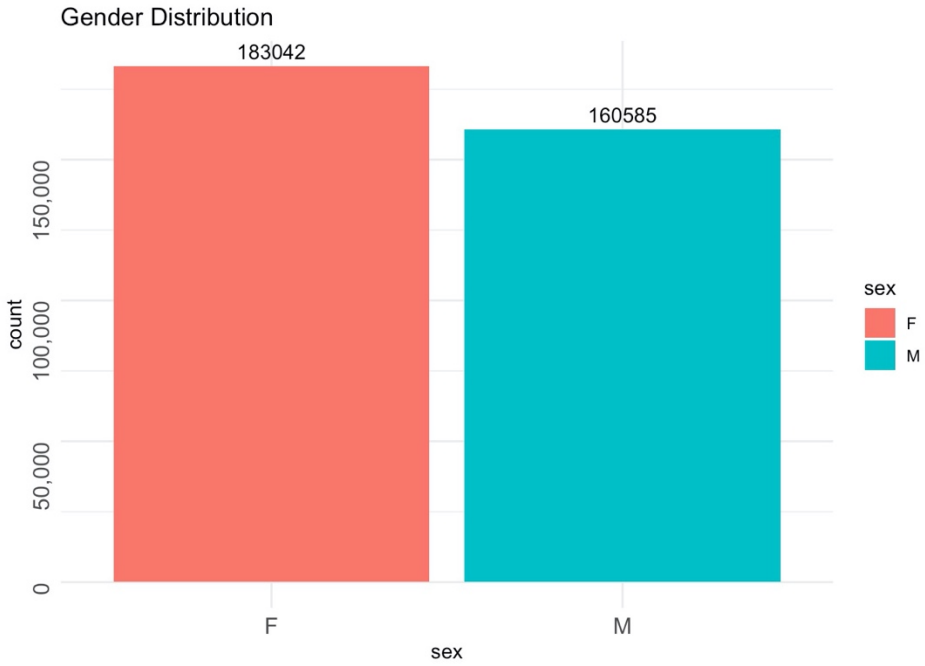
There are 5203 unique cities in the database with the most frequent cities are Sao Paulo, Rio De Janerio and Fortaleza combining around 15% of all cities in the dataset.

To visualize the distribution of customer locations we used a leaflet library to plot each customer on the map. The graph below shows that customers are distributed in 6 main areas.

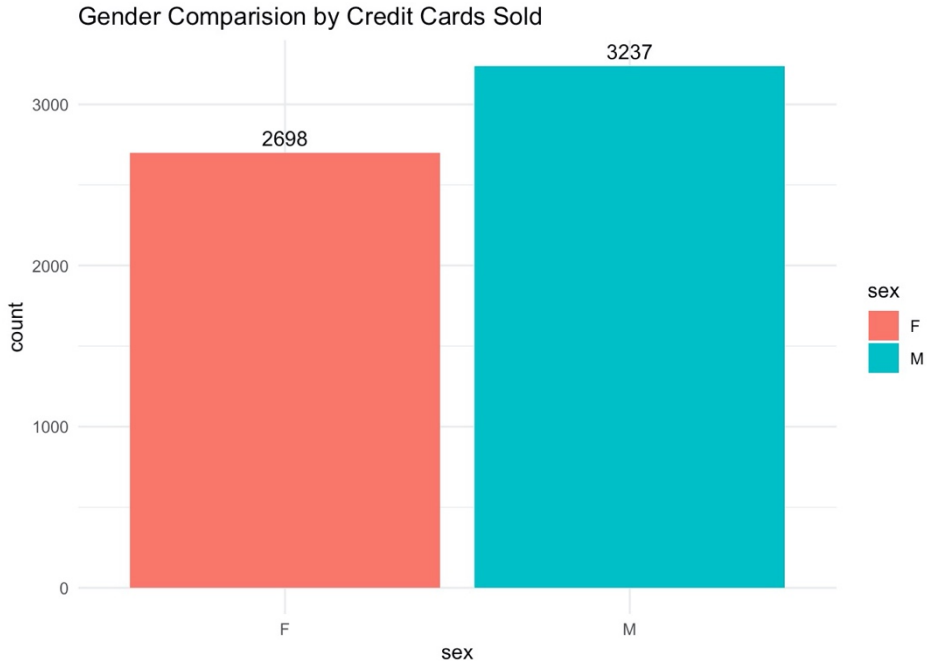


Map Brazil

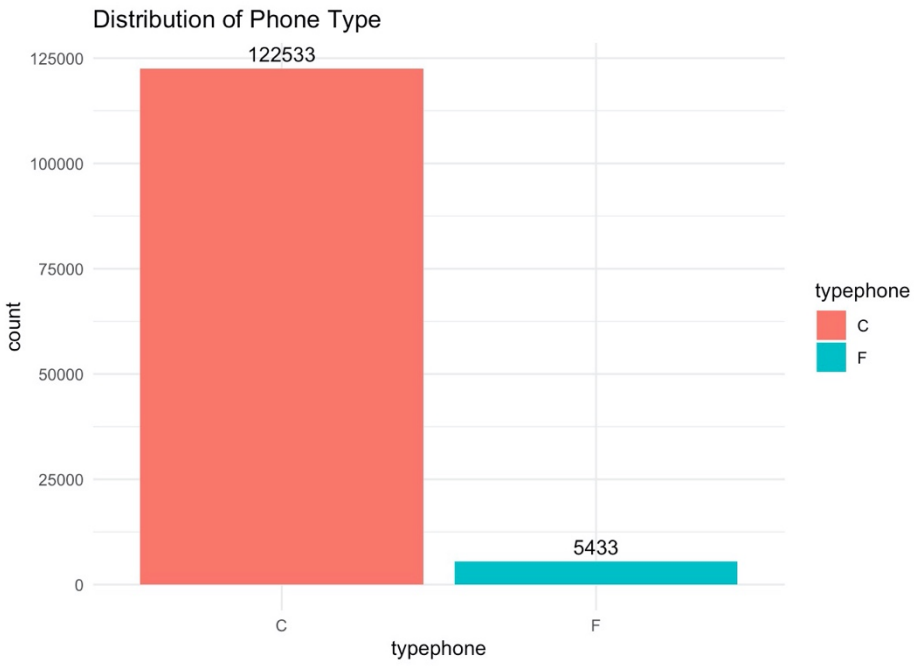
We discovered that there are more female clients in the database.



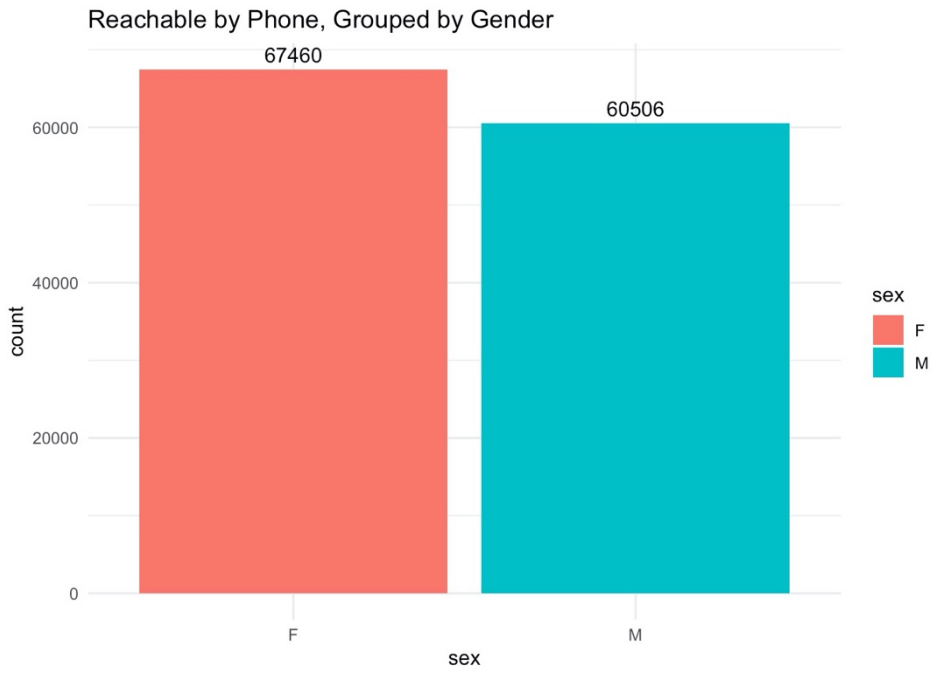
The graph below shows the difference in distribution by gender and filtered by sold credit cards. Even though, there are more female customers in the database the more conversions were made by male customers.



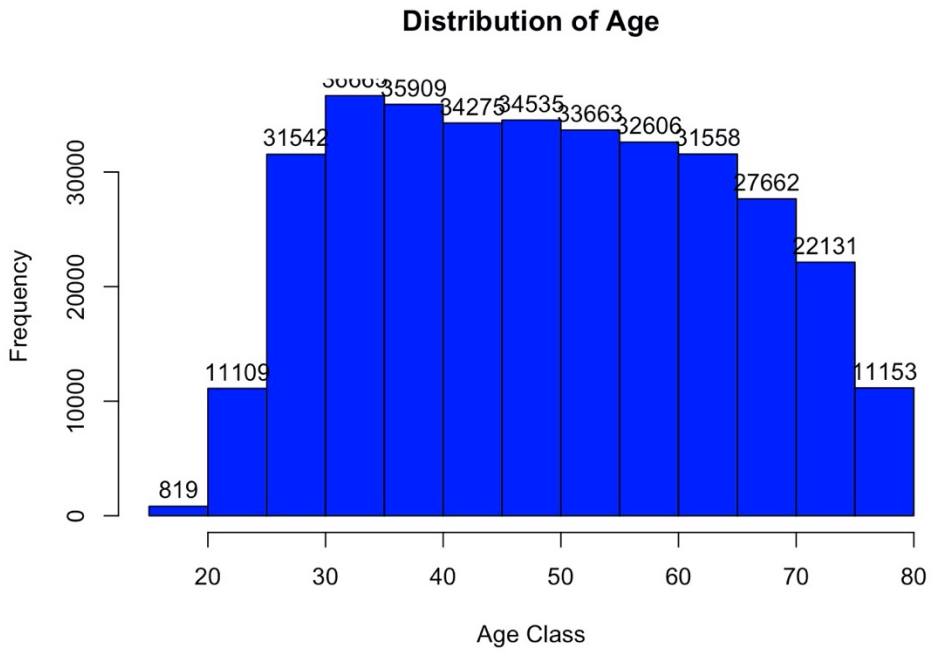
There are more customers with primer cell phone numbers this most likely means that these customers are more reachable over phone.



The plot below reveals that female customers are more reachable on the phone than male customers.

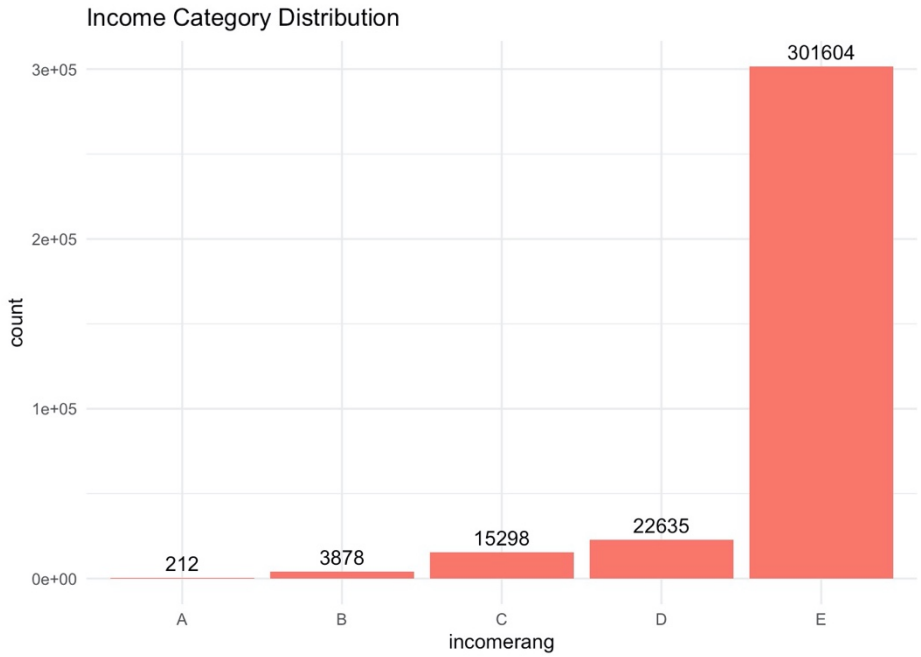


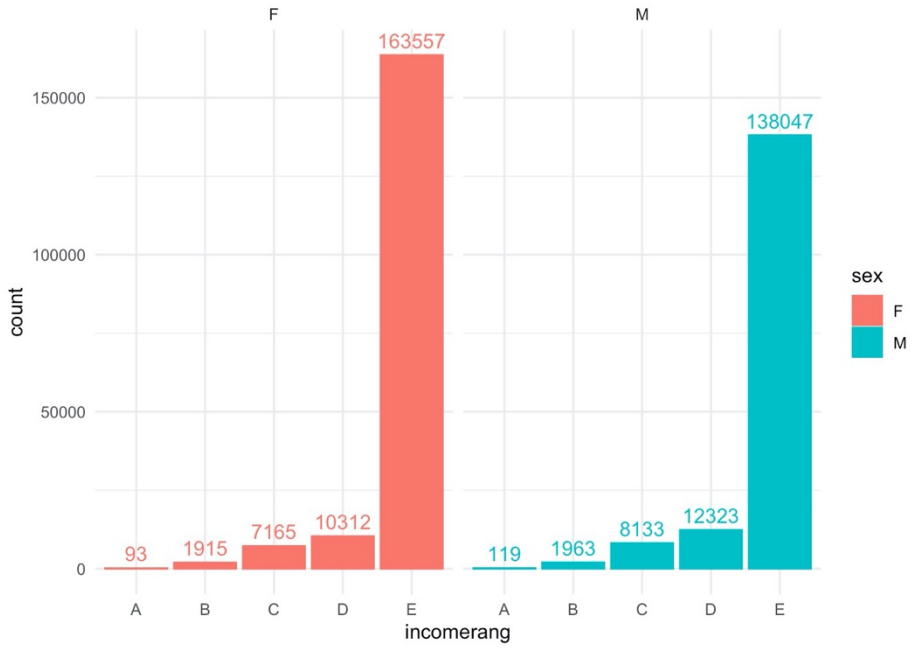
The histogram is shown below, displays age distribution of the dataset. We see that age is distributed quite evenly, with a few customers having age below 25 and above 75.



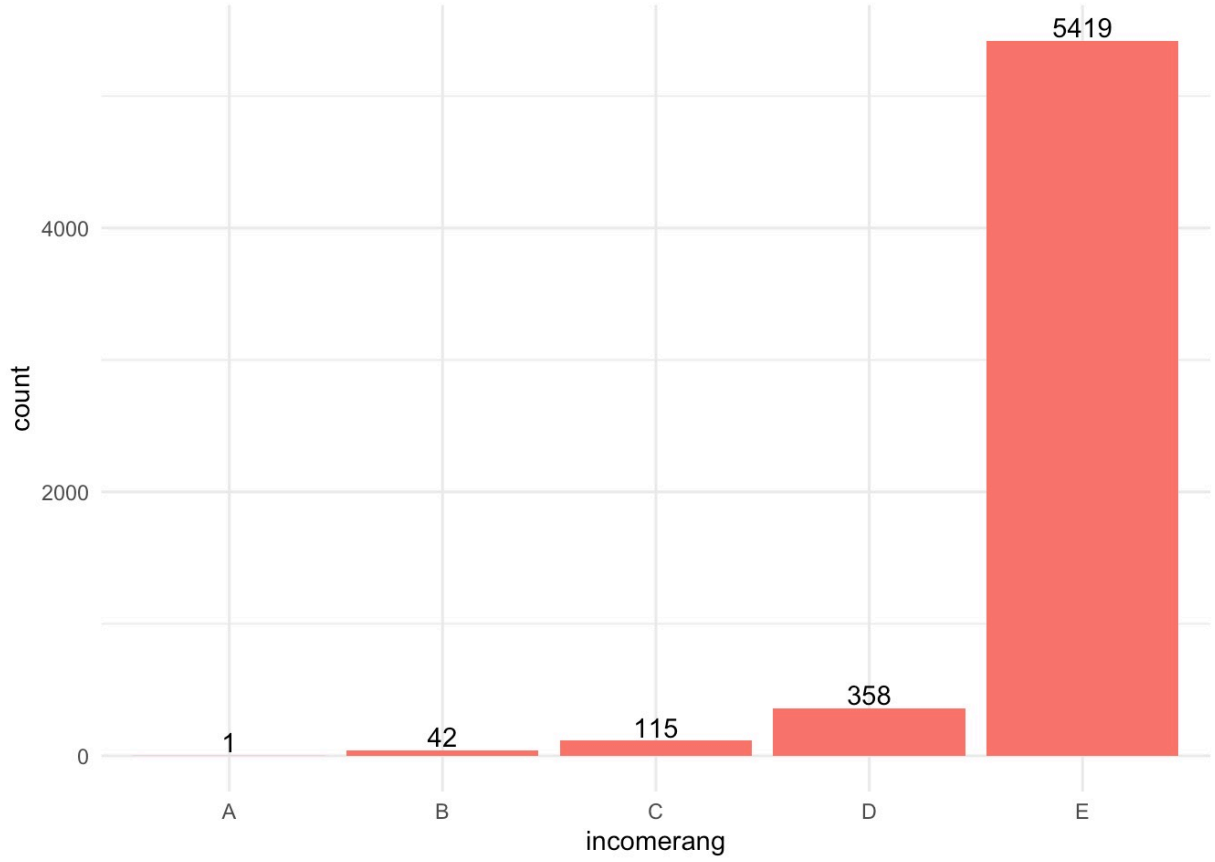
However, by grouping age with gender, we can see that males have more right-skewed distribution meaning that male customers tend to be younger.

Next, we explore income categories to see what group of income category has the highest propensity to convert. Our dataset consists of five income categories where A category having the highest income and E having the lowest income. We see that the database is dominated by the lowest income customers, we will explore it further to find the percentage of conversions for each category.





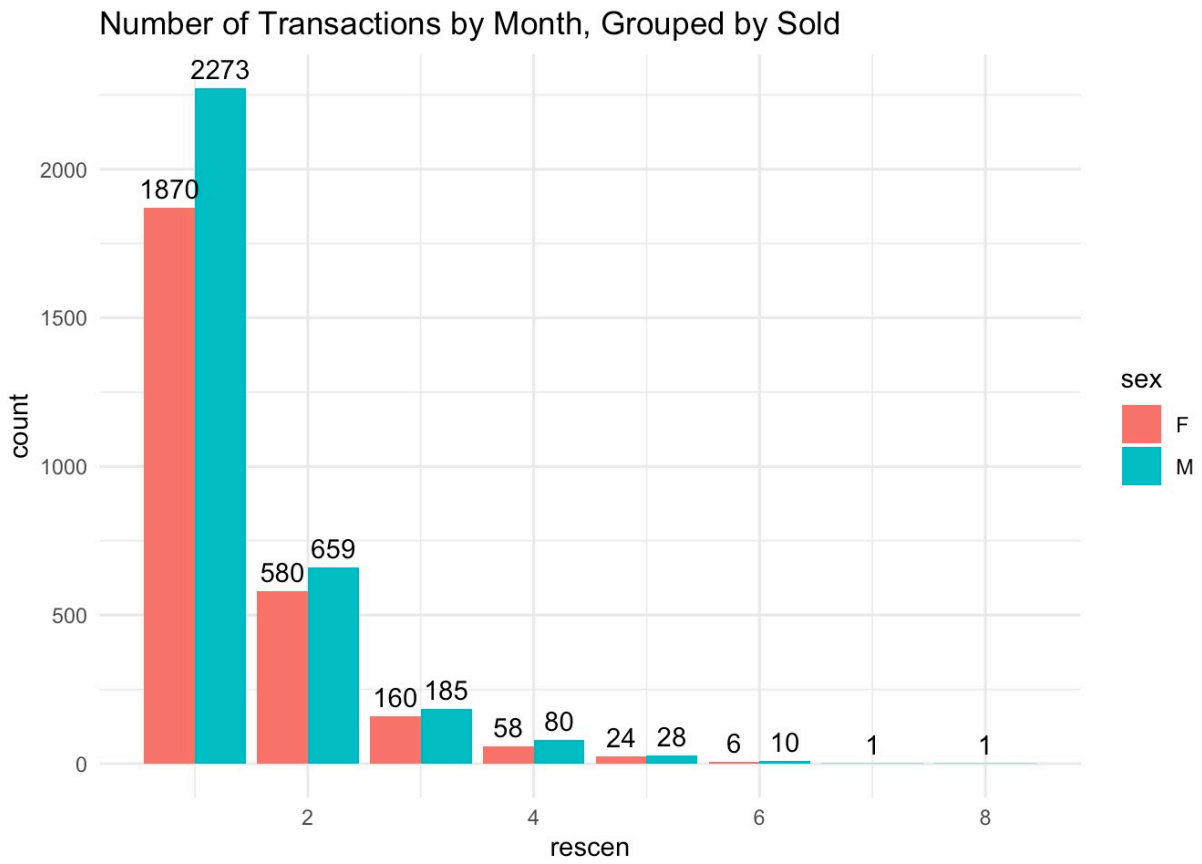
The plot above shows that income categories are distributed quite evenly among genders.



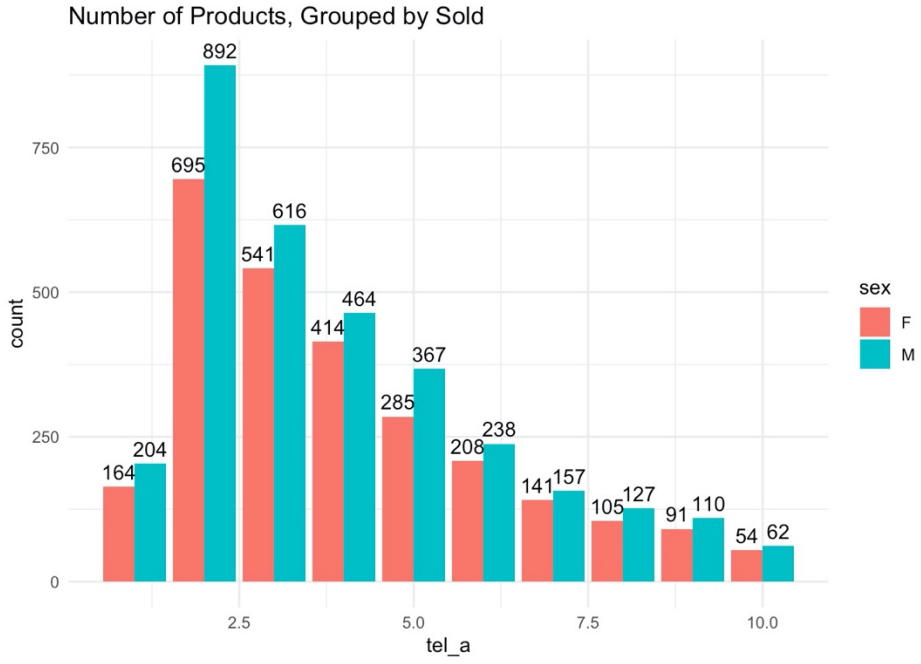
As we can see conversions are dominated by E income category with 5419 credit cards sold.

	percentage_sales
A	0.4716981
B	1.0830325
C	0.7517323
D	1.5816214
E	1.7967268

What we see from the table above is the most conversions happened for the group with the lowest income category (~1.8%). This particular information raises some ethical concerns since people with the lower-income might be financially restrained and more likely to convert as they need more financial help. It would be interesting to know what percentage of conversions from the E income category are going to default.

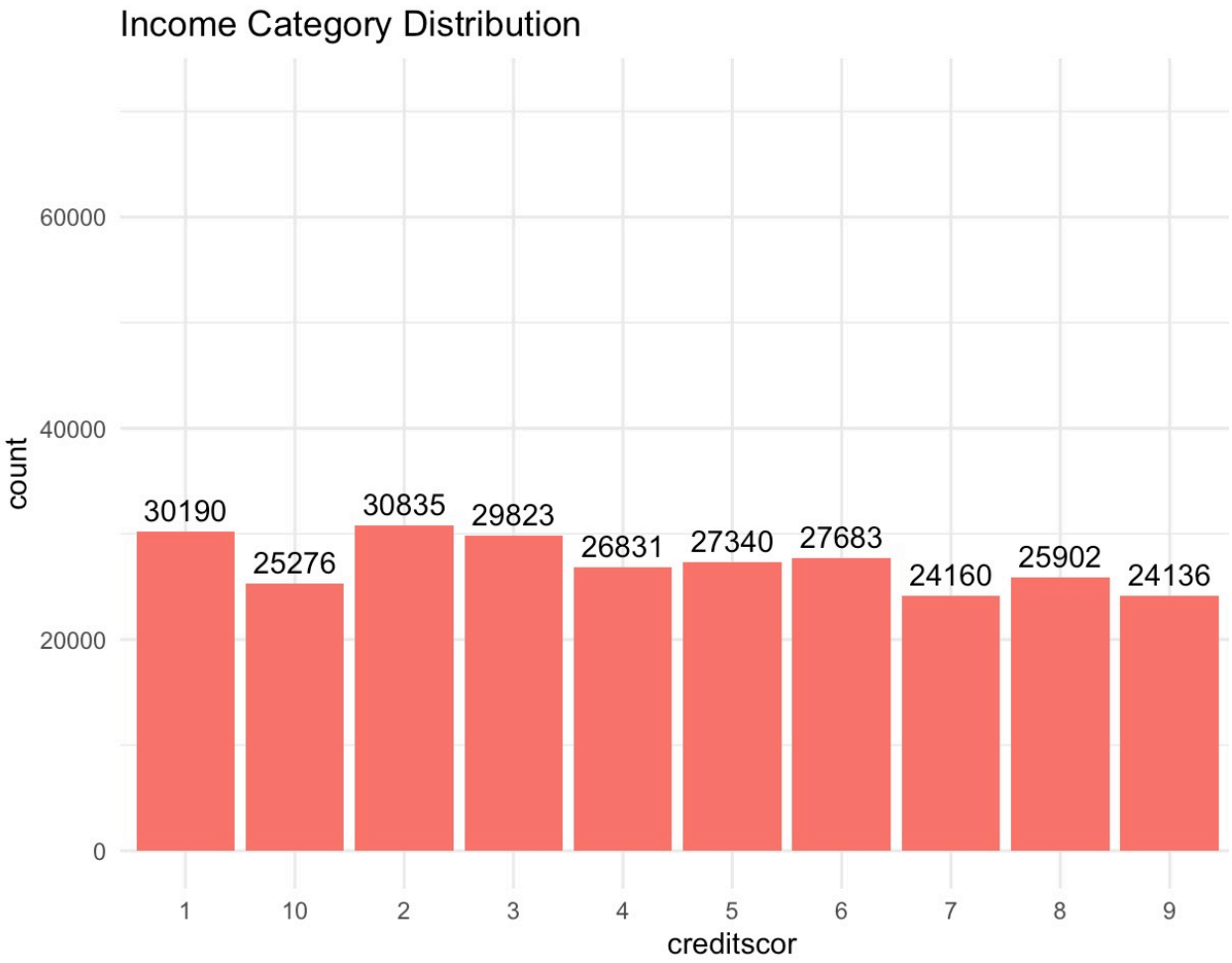


As we can observe from the graph above that the number of transactions per month has a high propensity to convert and males are more likely to convert with a low number of transactions.

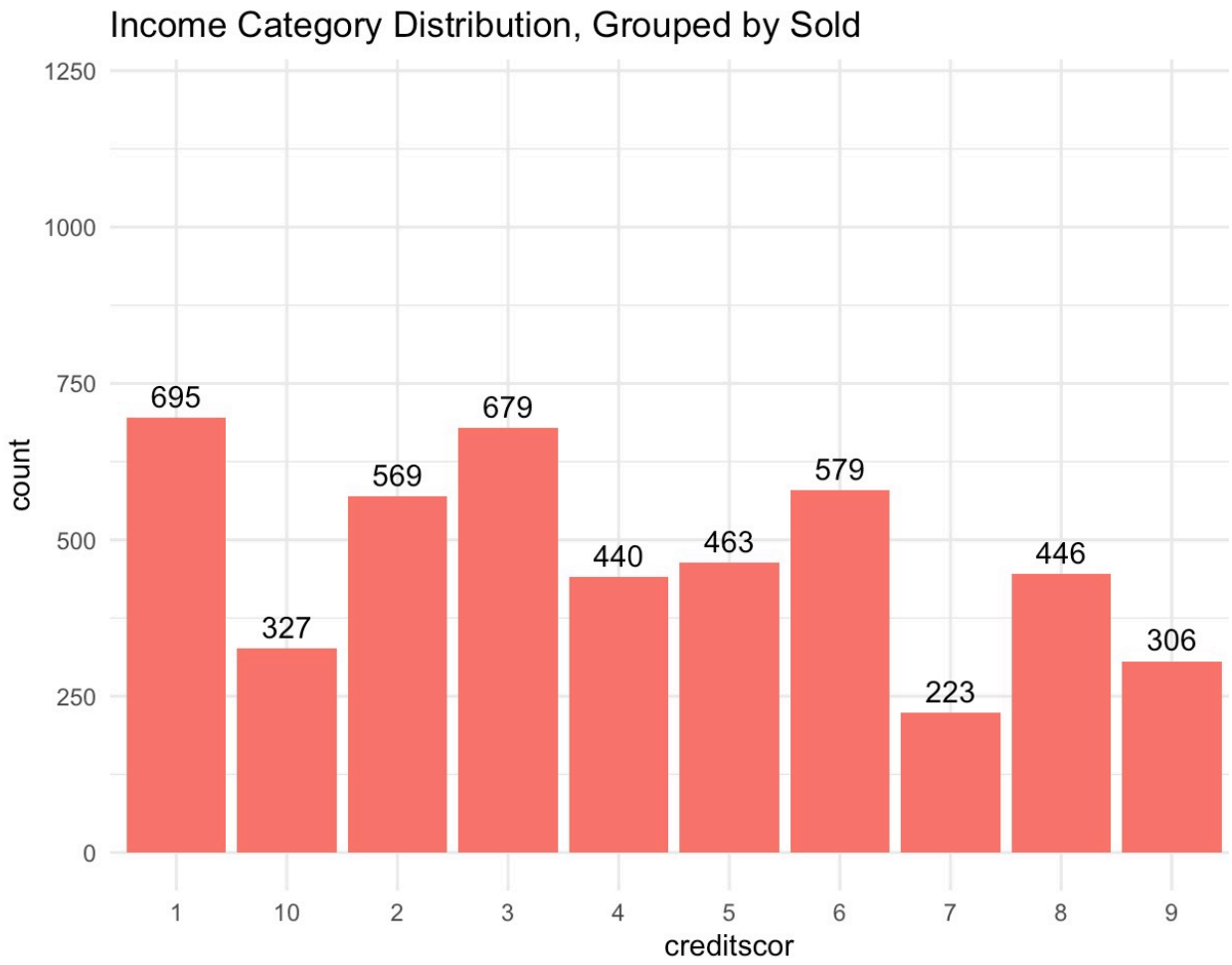


We see from the graph above that customers with 2 to 5 products are more likely to convert. We believe that this is very important information for the bank when they plan their marketing campaigns.

The plot below shows the distribution of credit score and as we can see a credit score close to the uniform distribution.



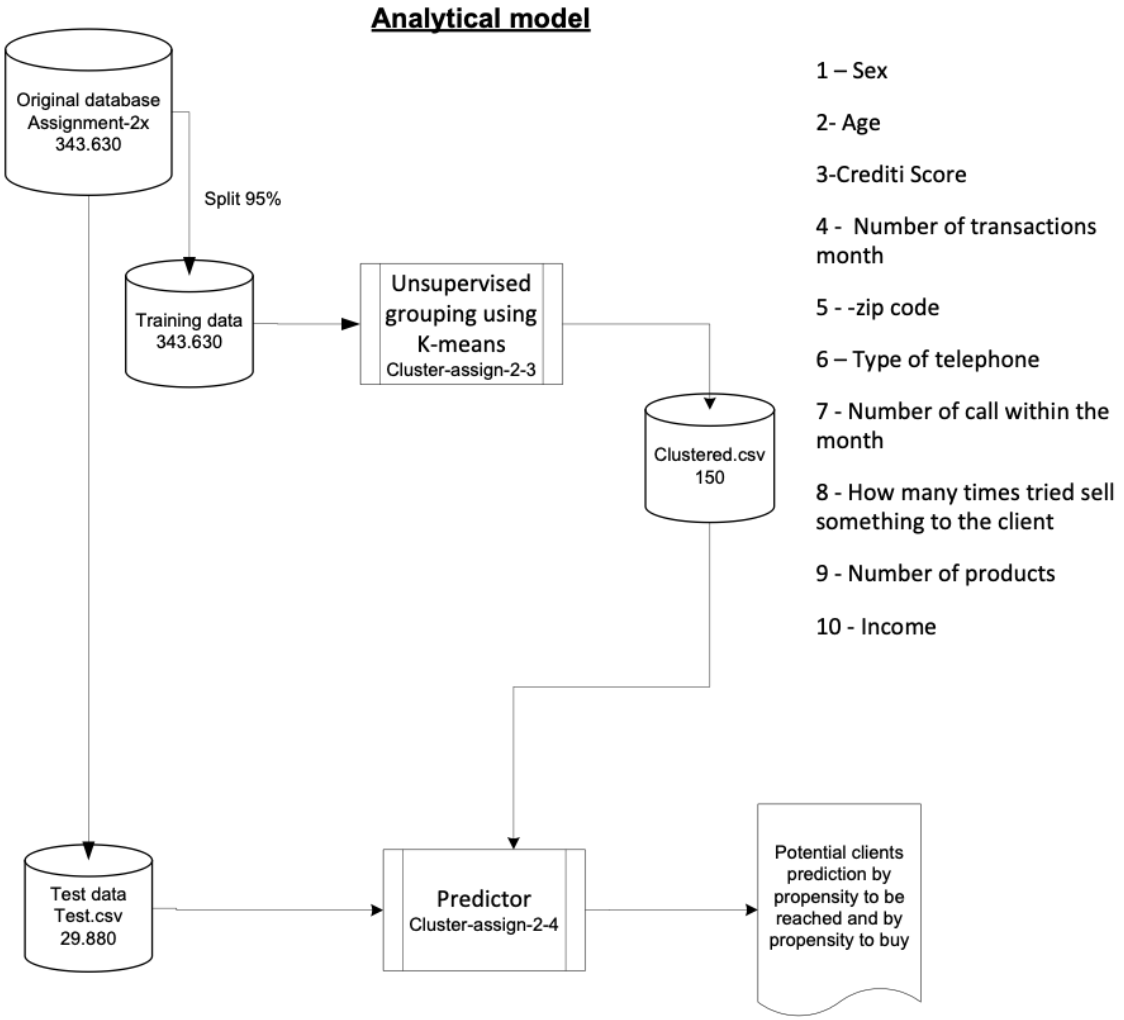
The histogram below shows that the majority of conversions contributed by lower credit score clients. It corresponds with previous findings that most of the conversions happened in the E (lowest) income category. Even though the availability of landing options for low income and low credit score might provide additional help for struggling families. We need to take into consideration that high interest and additional fees that credit cards impose can push people into more debt and lead to a high percentage of defaults. We believe this information can be valuable for decision-makers to address the issues we discovered and provide appropriate solutions.



Selecting the training data and the test data

The idea is to use 95% of the measurements as our training data and 5% as our test data. We had a database with 343.630 clients and we separated it into training data 313.750 and test data 29.880.

Model/Analytical model



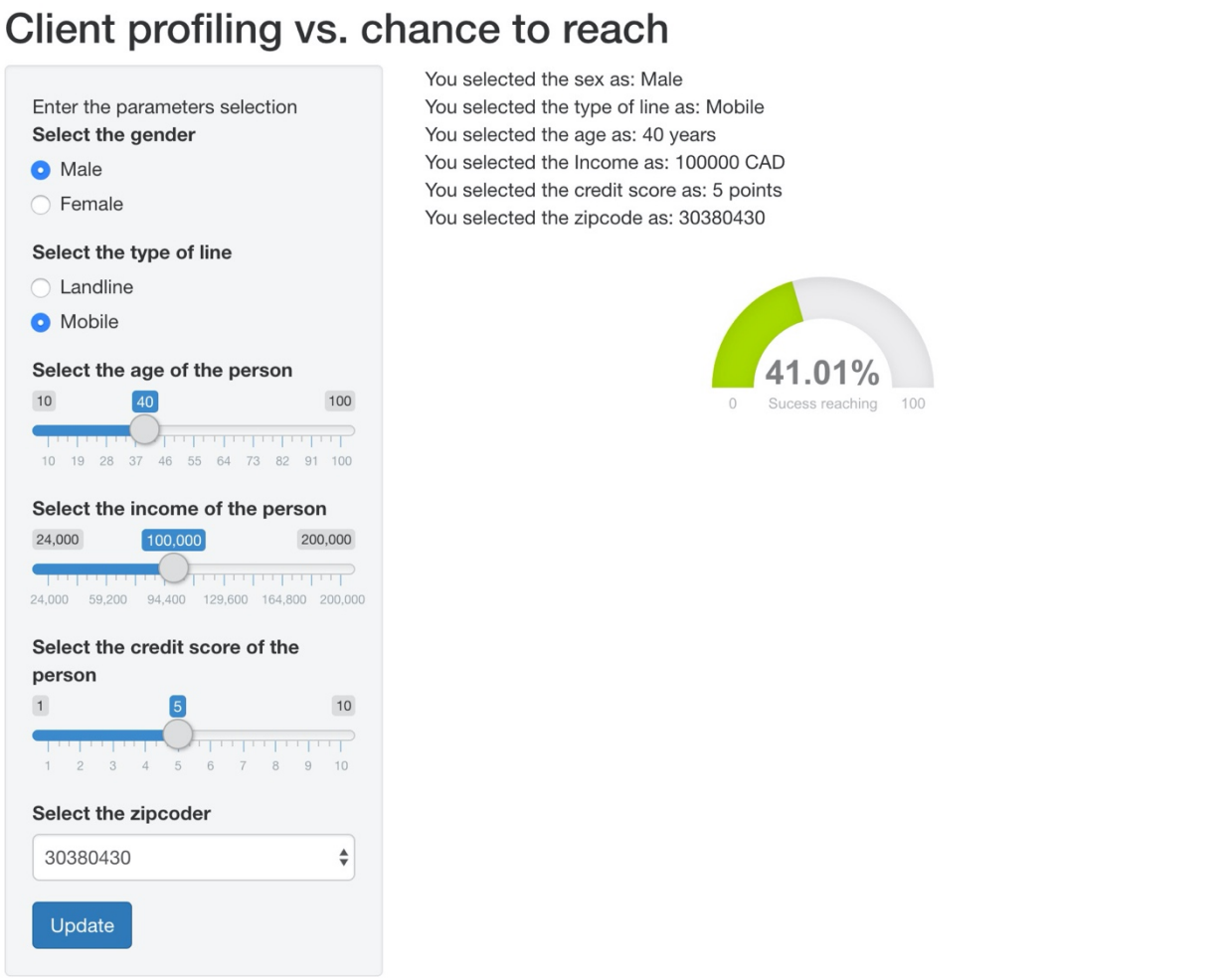
Normalization criteria

- SEX – 0 M – 1 F
- Age – age/ bigger age (80)
- Income A – E -> 1 to 5 divided by 5
- Credit score – 1 to 10 -> credit score/10 (10)
- Recencia -> Number of transactions within the month 0 to 13 – Divided by 13 (13)
- Type Phone -> 0 Mobile and 1 Landline
- Tel_o -> Number of call within the month -> 0 – 10 Divided by 10
- Tel_a ->Number of products Tel_a 0 -10 Divided by 10 Tel_a

- Enrri -> How many times tried sell something to the client 0-8 Divided by 8
- Zipcode divided by 100.000.000

Shiny App Implementation

We created an app that would allow users to utilize information about a client to predict a probability to reach a client by phone. We believe it is useful information for the business to increase conversion rates and minimize expenses. The app allows users to pick gender, type of phone line, age, income level, credit score and zip code. The output of the app shows the probability of reaching the client by phone. The statement output depends on the predictions made by our unsupervised clustering model.



Summary of findings /Evaluating the results

We used a training dataset with 313.750 rows and a test set with 29.880 rows.

The idea was to make the program predict the chance to reach and the chance to sell each one of the 29.880 clients and after doing the prediction check if the reach and the sales really happened. The results were as follows:

Reaching

Number of items range		test set	Real quant sucess reaching	Real % sucess reaching
10,00%	20,00%	1.067	160	15,00%
20,00%	30,00%	6.621	1.731	26,14%
30,00%	40,00%	9.768	3.331	34,10%
40,00%	50,00%	8.637	3.817	44,19%
50,00%	60,00%	3.678	1.923	52,28%
60,00%	70,00%	109	76	69,72%
Total		**29.880**	**11.038**	**36,94%**

Table 1

Note that the code classified the clients very precisely. The table shows the number of clients classified in each range for reaching (by the code) and the two columns in the right show what effectively happened. We can note that real success was within the forecast in every range. That demonstrates that the code was effectively separating the clients by their potential in being reached.

Selling

Number of items range		test set	Real quant sucess selling	Real % sucess selling
0,00%	1,00%	9.540	72	0,75%
1,00%	2,00%	10.410	154	1,48%
2,00%	3,00%	5.559	123	2,21%
3,00%	4,00%	2.687	86	3,20%
4,00%	5,00%	1.684	74	4,39%
Total		**29.880**	**509**	**1,70%**

Table 2

Again, the code classified the clients very precisely. Note the reality matches exactly what was forecasted. That means we are able to forecast the chances a client has to buy the service based on his profile.

In practical terms, the code allows the marketing operator to focus on those clients easer to be reached and the ones more prone to actually buy the services. The spreadsheet below gives a better view of the correlation between sales effort and success:

Group	Percentage of the clients	Percentage of the actual sales	Ratio % sales vs % clients
Group 1	31,93%	14,15%	0,44
Group 2	34,84%	30,26%	0,87
Group 3	18,60%	24,17%	1,30
Group 4	8,99%	16,90%	1,88
Group 5	5,64%	14,54%	2,58

Table 3

Note that with 33,63% of the potential clients (Group 3, 4 and 5) the company would be able to sell 55,61% of all sales. This code gives the organization the chance to for example: remove 66.37% of the clients from the calling list (saving the correspondent sales effort) but loosing just 44.39% of the sales. From the point of view of the third part company executing the sales process we have a situation with the following costs:

1) Without classifying the clients and operating 100% of the mailing, the third part company would need 100 sales agents (call-center operators) – Each earns 3.000 month -> total of 300.000 month with personnel (Plus around 80.000 with telecom and infra). This number of agents operates 300.000 sales tries per month with a success rate around 2%-> 6.000 actual sales month.

For each sales the company would get 70 as commission. Total revenue generated by the operation 420.000

Expenses: 380.000 Revenue: 420.000 Profit : 40.000

2) Classifying the clients the third part company becomes able to make 55% of the sales with 33% of the workforce (reducing the 300.000 to 100.000 the potential clients). That would bring the personnel cost down from 300.000 to 100.000. and the cost of infra from 80.000 to 23.000. 55% of the sales would bring the revenue down from 420.000 to 230.000 Expenses: 123.000 Revenue: 230.000 Profit : 117.000

Through the classification process the third part company managed to almost triple the profit in this operation (from 40 to 117).

Confusion Matrix and Statistics

```
          Reference
Prediction    0    1
         0 9459 4734
         1 1324 1664

               Accuracy : 0.6474
                 95% CI : (0.6402, 0.6545)
    No Information Rate : 0.6276
    P-Value [Acc > NIR] : 3.784e-08

                  Kappa : 0.154

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8772
            Specificity : 0.2601
         Pos Pred Value : 0.6665
         Neg Pred Value : 0.5569
             Prevalence : 0.6276
         Detection Rate : 0.5506
   Detection Prevalence : 0.8261
      Balanced Accuracy : 0.5686

       'Positive' Class : 0
```

Our model shows better level of accuracy (65%) compare to the current model used by the client that gives only 50% of accuracy.