Final degree Project

Bachelor's degree in Industrial Technology Engineering

New customers' classifier

Assessment based on first purchases activity

REPORT

Author: Pau-Ramon Casas Cachinero
Director: Ignasi Puig de Dou

Term: July 2019

Barcelona School of Industrial Engineering
ETSEIB


ETSEIB

Summary

Algorithms exist for tracking recurring customers purchases and promote their retention. These algorithms are based on assessing whether the customer is “behaving” today as he did in the past. However, they do not work for customer acquisition assessment. New customers do not have any transaction history to base their future profitability upon.

Companies are very interested in tracking new clients. They are often the ones that consume most of their sales force energy.

The goal of the project is to develop a methodology to classify new customers with high future potential from those with poor outcome based on their first transactions with the company.

In order to fulfill the project’s objective, we have chosen the Logistic Regression method as the discriminant of choice. This is a well-known classification method. The models created using this technique classify individuals between two groups based on a series of those individuals independent variables or predictors. In our case, new clients will be classified between those who have good future potential and those who do not. We will name them good and bad clients respectively. The independent variables used will be taken from the information obtained from each customer's first purchases.

Three different logistic models have been built. Each of them contains the information of the first, the first two and the first three purchases. The models are statistically significant and are good classifiers, right at the moment each purchase is made. In addition, univariate analysis has been used as a data exploration tool. This univariate analysis has provided some interesting insights not expected beforehand.

Once the project has been completed, it can be concluded that it is possible to predict whether or not a customer will be good in the future using logistic regression models. In addition, this prediction can even be carried out right at the moment customers make their first purchase. Therefore, this is a very powerful tool for any company. Even so, it has also been found that increasing the purchasing information used by the models do not substantially improve their discriminatory capacity.


ETSEIB

Index

1. Introduction

A Catalan spare parts distributor would like to use the information they obtain from their customers first purchases to classify them according to their future potential and distribute in a proper way the resources that are assigned to each one of the new customers. The distributor wanted to know which first purchases features were important in order to decide whether the customer would be good in the future. They also wanted to know the importance of each one of them.

In order to cover company's objectives we have decided to create three different models containing the information of the first, two first and three first purchases. The aim of these models is to give an estimated probability of a customer to be a good one.

This report aims at explaining in detail the entire process of model construction and fit analysis. Our aim is to explain it in a simple way so that anyone can understand each step without having advanced statistical knowledge. We are going to explain some basic concepts while the model construction is explained.

1.1. Project objectives

The main objective of the project is to **create three statistical models that are capable of predicting whether a new customer will become a profitable one after his first, second or third purchase.**

Secondary objectives of the project are:

- To detect which purchase predictors are significative and to quantify their importance.
- To study whether relevant predictors change among models and identify those that remain in all.

1.2. Memory structure

First, it is explained what information contains the main datasets used in the model construction. These datasets have been provided by the company. It is also going to be shown the data cleaning and data transformation/creation done to the datasets with the aim to improve the subsequent model construction.

Second, it is going to be carried out an exploratory analysis of the datasets to see how the data looks before models construction.

Third, it is described the definition for a good customer. The criteria will be applied to the customers in the dataset and it is going to be shown how customers are distributed. Good customer definition will be used as the response variable of the model construction, also called Y variable.

Forth, the variables that have been considered to have some kind of effect towards our dependent variable will be explained and defined. These variables will be named predictor variables or X variables.

Fifth, it is going to be exposed which method is going to be used in the model construction. A brief and simply explanation about its main characteristics will be presented. It is also going to be explained the methodology used in the model construction step by step.

Sixth, it will be presented the results of the univariate logistic regression and the analysis of each one of the three models that has been constructed.

Finally, a review of the model fit on the data for the models and a comparison between variables and model behavior will be done.

1.3. Software

To carry out both the data wrangling and the construction of the statistical models we have chosen the computer language R. R is an open-source language for statistical analysis and graphics. It is currently widely used in both academic and professional fields.

This computer language can handle large amount of data and provides a series of statistical techniques, as for example logistic regression analysis. In addition, it allows the construction of very good quality graphics in a simple way. Another of its strengths is the large number of packages created by the community that add extra functionality to the base language, which is mainly due to the open source philosophy of the language.

The R-studio interface has been used as the R IDE (Integrated Development Environment). It is also open source and serves to improve the visualization when programming thanks to the unification in the same window of different sections necessary for programming. To write the model's construction programs we have used R-markdown, a feature present in R-studio.

1.4. Available data

One of the most interesting aspects of the project is that the models to construct and all the analysis to be done will be entirely based in real data. This means that the results will be 100% applicable to “real-word” problems.

The original datasets contain nearly 1 million of line items from purchases done by around 15,000 of the company's customers. From these initial data we will finally work with about 70,000-line items and 6,000 customers. These data correspond to customers who have started working with the company in the observation period going from May the 2nd 2016 to March the 29th 2019, i.e. those who can be categorized as new customers.

2. Datasets

The models we are going to create in the project are based in two main datasets: Tickets and masterCustomers. Both datasets have been provided by the company. This section will provide a description of each of the datasets, their cleaning process and an Exploratory Data Analysis.

2.1. Tickets

customer_id	n_container	n_delivery_note	date_deliverynote	up	family	sku	qty	price	disc	amount
66	6331689	545197	2018-12-10	1	5	360863	20	0.8220	0	16.4400
66	6331689	545197	2018-12-10	1	4	440073	10	7.6000	0	76.0000
66	6331689	545197	2018-12-10	1	2	448266	4	9.2500	0	37.0000
334	6430326	588939	2019-03-27	1	1	649015	2	6.5500	0	13.1000
374	6378578	566068	2019-02-05	8	332	60032	1	25.0000	0	25.0000
374	6378578	566068	2019-02-05	8	332	99474	1	89.0000	0	89.0000

Table 2.1: Tickets dataset structure

Table 2.1 shows the original structure of the Tickets dataset. Each line contains the information of a transaction and represents the purchase or refund of a product (or sku). The dataset fields are:

- **Customer_id**: Costumer unique identification.
- **n_container:** Purchase identifier. Each time a customer contacts the company in order to place a new order a container is created. One container can contain any combination of purchases, refunds and price adjustments, each one of them with its own date of transaction.
- **n_delivery_note**: Delivery note identifier. They are included inside containers and tend to represent actual shipments. Besides a shipment a delivery note can also be a refund or a price adjustment and it is possible to find more than one delivery note inside a unique container.
- **Date_deliverynote**: Date in which the transaction was issued.
- **Up**: First level of product identification. Inside every up group there are different families.
- **Family**: Second level of product identification. Inside every family group there are different skus.
- **Sku**: Third level of product identification. This is the last level and each product have its own sku identification number.
- **Qty**: It is the number of skus purchased in the transaction line.
- **Price**: Unit price of the line's sku in Euros.

- **Disc**: Percentage of discount applied to the transaction line.
- **Amount**: Total amount in the transaction line. It is the product of qty and price minus the discount applied.

There are three different types of delivery notes. They include one or more items. They can be classified by the amount of negative and positive lines included in each one of them.

1. Purchase delivery note: It is a "normal" delivery note linked to a customer shipment. 100% of the lines of this type of delivery note have positive values.
2. Price adjustment delivery note: They are used with the purpose to correct the price of one or more products. Wrong price is credited back and then the company invoice the correct one. They are delivery notes with the same number of negative and positive lines.
3. Refund delivery note: Delivery notes used when a customer returns one or more products to the company. All its lines are negative.

Note that these classifications are made using delivery notes and not containers. It is important to define the type of each delivery note in order to use the information we have in an optimal way when we construct the models.

In the original tickets dataset there were 901,867 lines.

2.2. MasterCustomers

customer_id	treatment	market	sector	branch	digita_cli	country	province	starting_date
334	S	MRO	28	3	N	34	GUIPUZCOA	2019-03-27
742	S	MRO	13	11	N	34	SEVILLA	2018-12-03
858	S	MRO	28	3	N	34	GUIPUZCOA	2018-12-19
920	S	MRO	28	12	S	34	BIZKAIA	2019-01-21
1491	S	MRO	25	11	N	34	SEVILLA	2019-03-11
1954	S	MRO	0	12	N	34	BIZKAIA	2018-12-20

Table 2.2: masterCustomers dataset structure

In this second dataset each line contains information about a single customer and there is only one line per customer. Table 2.2 shows the structure of masterCustomers dataset. Its field contents are:

- **Customer_id:** Customer identification number
- **Treatment**: It can take two values: Gold or Silver. These two values are given by the company depending on the customer annual revenue or its "reputation". That variable is assigned by the company in hindsight and it is not available when a customer places its first purchase.

- **Market**: It can take four levels: MRO, OEM, OTR or SI. It depends on the customer's type:
  - **MRO**: Maintenance and Repair.
  - **OEM**: Original Equipment Manufacturing.
  - **OTR**: Others.
  - **SI:** Industrial Systems.
- **Sector**: It can take 29 different levels depending on the client activity sector.
- **Branch**: Provides information about the company's branch with which the customer contacted on its first purchase.
- **Digita_cli:** It can take two values: Yes or No. Yes customers are customers who made its registration using digital methods, basically the company's webpage.
- **Country**: Country in which the customer is located.
- **Province**: Province in which the customer is located.
- **Starting_date:** Date in which the company made the customer's registration.

In the original masterCustomers dataset there are 15,157 lines which corresponds to the same number of different customers.

2.3. Cleaning the dataset

Original dataset must be prepared and cleaned before starting to create new variables and construct the model. This process is called cleaning the dataset. We have conducted the data cleaning in both tickets and masterCustomers dataset.

Before starting this process, we proceeded to translate original datasets from Spanish to English.

We have done the following cleaning processes in the tickets dataset:

1. As we mentioned before, there are three types of delivery notes. We are only interested in the purchase and refunds delivery notes. Adjustment of price is a company "internal" process that it is not induced by the customer, this means that it does not give us information about their purchase behavior. We have eliminated all the adjustment of price delivery notes from the original tickets dataset.
2. In the tickets dataset there were lines with null amount. These lines correspond to gifts made to the customers. As it is not a transaction made by the customer, we proceed to eliminate all lines with an amount's equals to zero.
3. Some line items had n_container value 0. We found it in two situations. In one case, it was clear which was the correct container identifier because customer_id, date_deliverynote and n_deliverynote of the zero-container line were equal than in

other correct lines (see Table 2.3) In this case we assigned the correct container to the wrong lines.

In the other case, we assigned n_delivery_note value to the container column.

customer_id	n_container	n_delivery_note	date_deliverynote	up	family	sku	qty	price	disc	amount
37333	0	170794	2016-05-19	1	5	145245	20	1.622	0	32.44
37333	5487206	170794	2016-05-19	1	5	361391	30	1.620	0	48.60
37333	5487206	170794	2016-05-19	1	3	669283	3	5.800	0	17.40

Table 2.3: First situation of null container example

With regards to the masterCustomers dataset, we have done the following two cleanings:

1. All customers with starting date prior to May the 2nd 2016 were removed. That was the first day of the observational period. Our model's main objective is to predict the future behavior of new customers. All customers "born" before the starting of the data collection period cannot be included as their first transaction data is missing.
2. All customers with first purchase date previous to their starting date were also removed. As seen in Figure 2.1.a, some customers had their creation date in the database after they had done some transactions with the company. This was communicated to the company and they came to the conclusion that it was due to some batch uploads of customer's that did not reflect actual creation dates.

After this cleaning we got the final selection of customers. Figure 2.1.b shows the relation between the starting and first purchase date. Most customers had placed their first purchase close to their starting date but there are some with long periods between creation and first purchase.

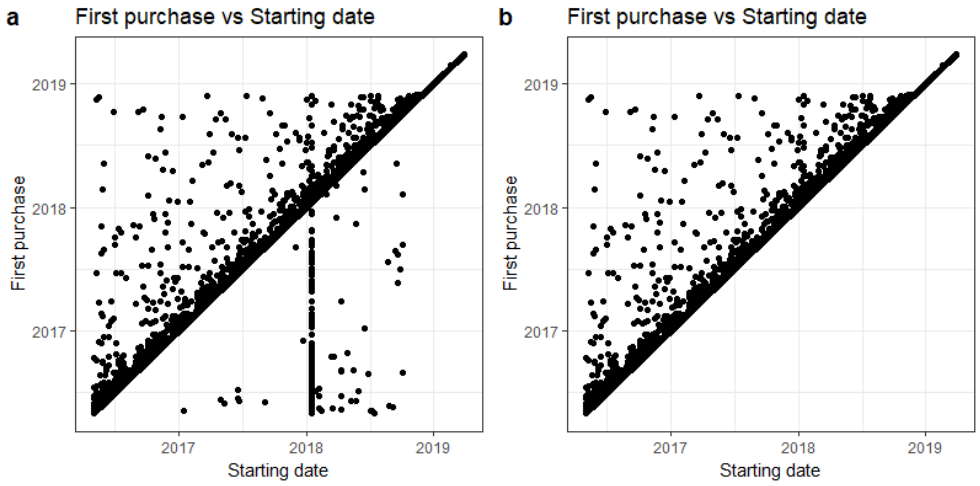


Figure 2.1: a) First purchase vs Starting date before first cleaning, b) First purchase vs Starting date after first cleaning

Finally, we removed all line items in the tickets dataset that did not belong to the new customers set.

After the cleaning process we did three modifications in the datasets to prepare them for the model construction process.

1. Apart from the main tickets and masterCustomers datasets, we created a "dictionary" with the meaning of each branch number. We used that information to translate each number to its corresponding branch location name.
2. There are two different variables that provide information about the location of the customer and the corresponding company branch. In order to standardize them to improve model construction and the analysis of the relation between both we made the transformations shown in Table 2.4.

Original branch denomination	Bilbao	Donostia	Vigo	A Coruña
Final branch denomination	Bizakaia	Guipuzcoa	Pontevedra	A Coruna

Table 2.4: Modifications in branch denominations

3. There were only four customers in the Market OTR group. As this little number could let to mathematical problems in the model construction process we assigned them to Market MRO. We assigned MRO value because as we will see below it is the major market group.

2.4. Exploratory Analysis

It is important to do an Exploratory Analysis before the variable creation and model construction processes. It provides a background knowledge about the data we are going to handle and helps identify possible errors or outliers.

The exploratory analysis that follows only takes into account the new customer subset. That is, all customers that started transacting with the company after May 2nd 2016. The exploratory analysis was done after the data cleaning of the datasets.

2.4.1. Tickets

The observational period started on the May the 2nd 2016 and ended on March the 29th 2019: We have information about all transactions done during three years approximately.

There are 60,677 line-items in the final tickets dataset from the 901,867 line-items contained

in the original one. This reduction is due to keeping purchase information from customers born after the beginning of the observation.

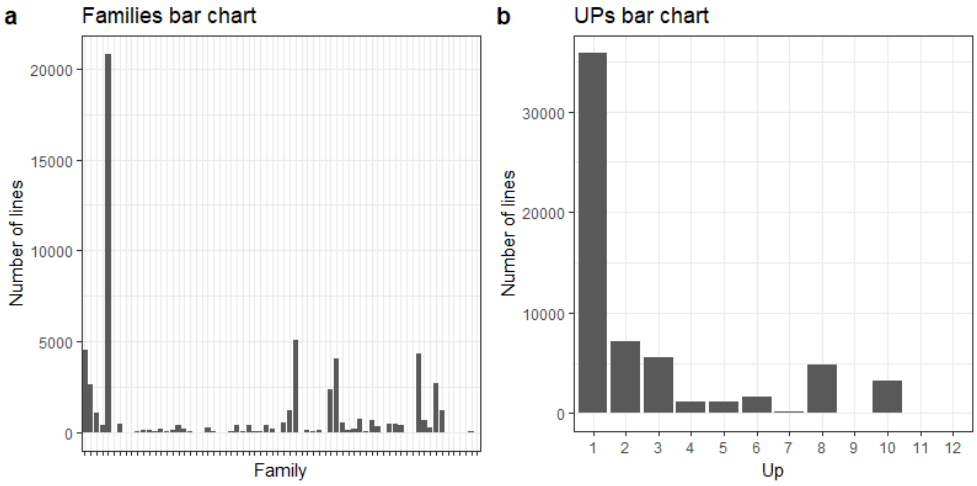


Figure 2.2: a) Families bar chart, b) UPs bar chart

As shown in Figures 2.2.a and 2.2.b most line items gather in few UP's and families groups.

Family 5 has 34.25% of all the line items. The first five families with more line items have 64 % of the total.

If we look the UPs chart we can see a similar situation. 60% of the line items belong to UP 1 and 87.73% of them belong to only 4 UPs. There are also three UP groups of which any new customer have purchased any product.

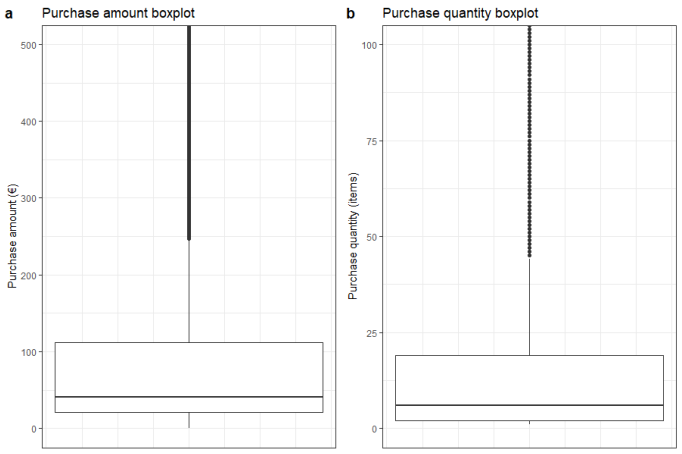


Figure 2.3: a) Amount by purchase boxplot. Limited between 0-500€, b) Quantity by purchase boxplot limited between 0-100 items

The median ticket value is 41.05 €. In the boxplot displayed in Figure 2.3.a we see that the third quartile is just above 110 €/ticket. There is a large group of purchases above 247 €/ticket,

the boxplot upper limit. 10% of purchases are larger than 291 € and the maximum amount spent in an unique purchase is 79,810€. This shows a positive skewness in the ticket value variable.

Figure 2.3.b shows the Quantity by purchase boxplot. The median quantity bought in a purchase is 6 items. Similar to the ticket value, there is a large number of purchases above the boxplot upper extreme, 44 items.

Dicounts are very rare, only the 0.30 % of purchases have any discount applied. Discounts have been only applied to 18 customers.

There are 22,216 different purchases that give a mean value of 3.84 purchases by customer in the observed period.

2.4.2. MasterCustomers

At the final masterCustomers dataset there is information about 6,098 customers, the reduction from the original number of 15,157 customers is due that as It has been previously mentioned we are only going to study customers with a starting date after the start of the observation period.

Spain hosts 98.7% of the customers. 41.5% and 34.4% of them belong to sectors 0 and 28 respectively.

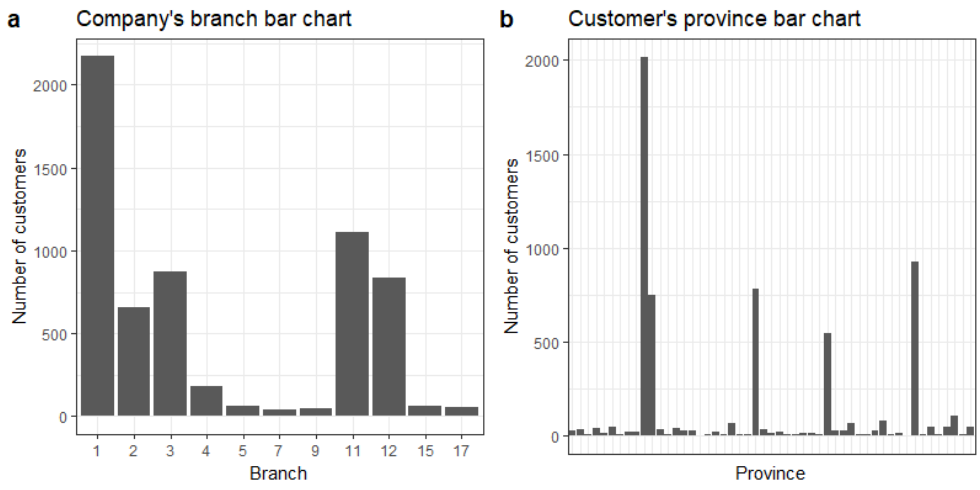


Figure 2.4: a) Company's branch bar chart, b) Customer's province bar chart

Figures 2.4.a and 2.4.b show the customers per branch they are being served and province they are located respectively. 36% of them are assigned to branch 1 located in Barcelona. There are 4 other big branches that have 57% of the customers. They are Sevilla, Donostia, Bilbao and Madrid. The six remaining ones only represents 7% of the total.

With regards to the province variable the distribution is very similar. 33% of the customers are located in Barcelona and 49% in Sevilla, Guipúzcoa, Bizkaia and Madrid. As we can see, location and branch location percentages match. It seems that most of the customers are assigned to branches in their location. We will check for this assumption in future stages.

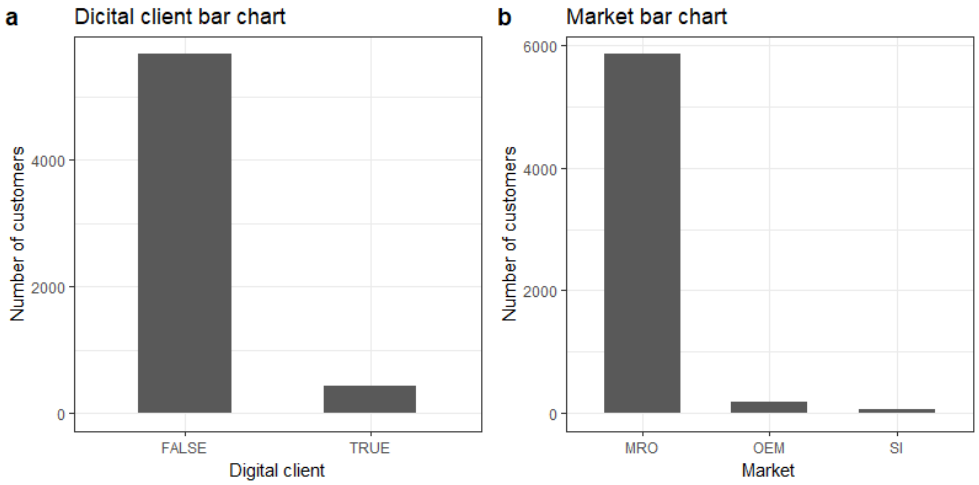


Figure 2.5: a) Digital client bar chart, b) Market bar chart

93% of customers are not "digital". MRO customers represent the 96% of the total. OEM and SI customers are 3% and 1% respectively.

After doing masterCustomers Exploratory analysis It can be said that a standard customer is a customer which is located in Spain, is not digital and belongs to market MRO. There are not a clear major group in province and branch variables.

3. Response variable

3.1. Definition of a good customer

As stated before, the main objective of this project is to determine whether a new customer will be “good” or “bad” given the features of his first purchases with the company. For this classification to work it is key to clearly classify already existing customers as good or bad from the company’s point of view.

After conversations with the marketing teams of the company we came with two parameters to define a good customer: 1) whether he is “dead” or “alive” and 2) his monetary value for to the company. A good customer is one who is “alive” and has a monetary value above average.

3.2. Alive or dead customers

To study which features determine whether a new customer is going to be good we are going to study the behavior of the ones from which we have recorded data during the observational period.

The first step is to decide when a customer is “dead” or “alive”, it has been called its activity status.

A “dead” customer is one which we think has ended his relation with the company and will not buy from us again. On the other hand, an “alive” customer is a customer who is active and is likely to place orders in the near future.

Our company’s setting in its relationship with customers is known as non-contractual. Customers do not formally communicate to the company when they plan to stop transacting with it. Contrary, settings like insurance’s, banks or phone companies are known as contractual settings as their relationship with customers is based on contracts. End of a customer relationship is known for sure. In non-contractual settings, there is never a 100 % assurance that the customer is done with transacting with the company. There is always the likelihood that he is just going through a long relapse in his activity with us. This makes classification of dead or alive customers harder

To approach the issue, we have differentiated customers in three groups depending on the amount of purchases done. One group includes customers that have made only 1 purchase, the second customers that have done 2 or 3 purchases, and finally, customers with 4 or more purchases.

We have also created two variables that will be used in the definition of a "dead" or "alive" customer:

- **Global behavior:** it can take 2 values: TRUE or FALSE. It is TRUE if the customer's recency[1] is smaller than the 80% quantile of the distribution of all customers maximum time between purchases, named $Tmax_{80}$ (312 days as will be seen below).
  We chose maximum time between purchases as it showed lower variability versus recency than average time between purchases.
  The period of time within 312 days prior to the last day of the observation period is named "Recency zone". This term is going to be used in the monetary value explanation.
- **Individual behavior:** It is a binary variable. It is TRUE if the recency of the customer is smaller than its own maximum time between purchases.

Depending to which group the customer belongs we define whether he is dead or alive using the following criteria:

1. Customers with one purchase. A one purchase customer can never be considered as an "alive" customer, they can be either "dead" or uncertain about their status, an NA customer. They will be considered dead if their "Global behavior" variable is false and NA otherwise. That is, they are dead if their first and only purchase was done before the "Recency zone" and NA if it was done within it.

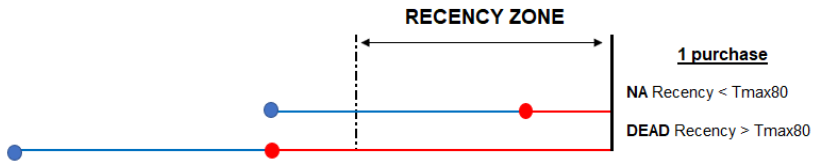


Figure 3.1: Representation of a dead and an alive One purchase customer

2. Customers with two or three purchases. They are considered "alive" if their "Global behavior" or "Individual behavior" variables are TRUE. In any other cases they are considered "dead".

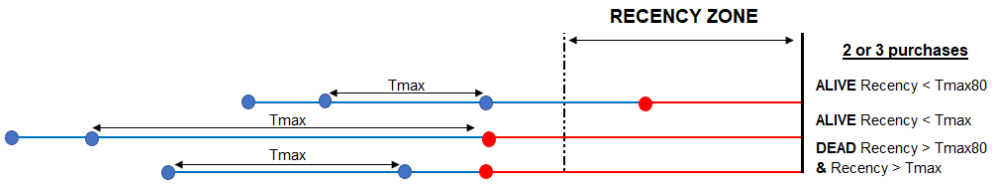


Figure 3.2: Representation of two alive and one dead Two purchases customers

[1] Recency: Time between customer's last purchase and the end of the observational period.

3. Customers with four or more purchases. These customers are considered “alive” if the “Individual behavior” variable is TRUE. Otherwise they are “dead”.

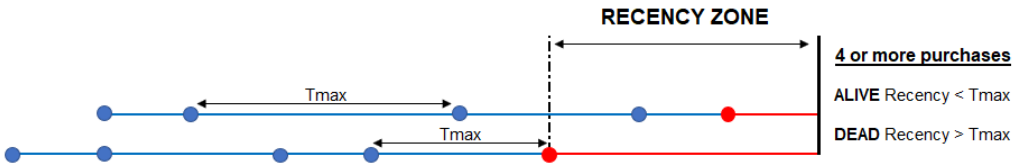


Figure 3.3: Representation of a dead and an alive Three purchases customers

The more information we have about a customer the more we can rely on their own purchase behavior the less we have to rely on global behavior data.

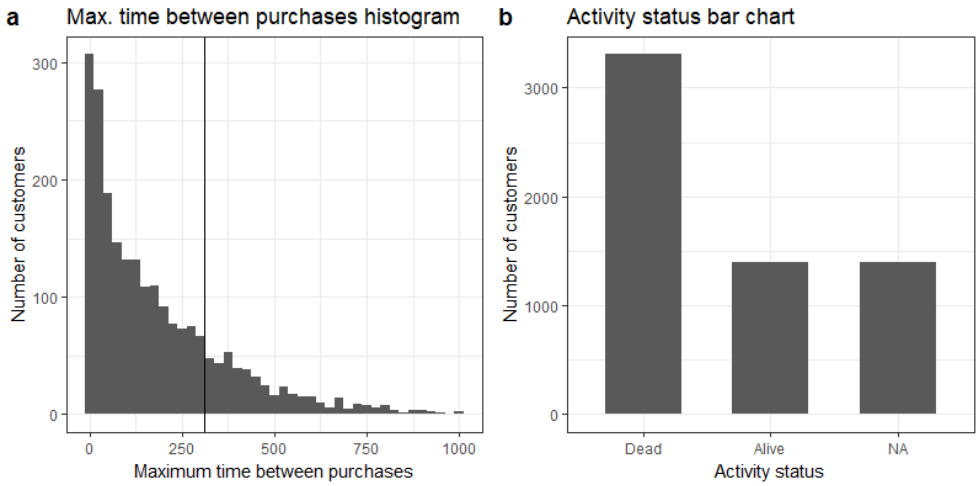


Figure 3.4: a) Maximum time between purchases histogram, b) Activity status bar chart

The recency zone is the time lapse between May 21st 2018 and March 29th 2019 as the 80% quantile of the maximum time between purchases for all customers is 312 days.

It can be seen in Figure 3.4.a the distribution of maximum time between purchases. It can be observed that that variable is small for a large number of customers and decreases exponentially.

Figure 3.4.b shows a bar chart of the distribution of customers depending on their activity status. Of the total customers born in the observation period 3,308 are assumed dead, 1,392 alive and 1,398 we cannot say anything as of 29-03-2019 (NA’s). Dead customers are 54.3% of the total number while alive and NA customers represent nearly 23% each one.

3.3. Monetary value rate

Once the dead or alive status of a customer is assessed we estimate the second key parameter: monetary value. We have defined monetary value as the total value spent by the customer by time unit. It is computed as follows:

$$Monetary\ value = \frac{totalvalue}{t.firstpurchase} \quad \text{(Eq. 3.1)}$$

Totalvalue is the amount spent by the customer since its first purchase and up to the end of the observation period minus its refunds if any. *t.fisrtpurchase* is the lapse of time between the first purchase and the last day of the observational period.

The higher the total value spent is, the higher the customer monetary value variable is. The lower the time lapse in which this amount has been spent, the higher the customer monetary value.

There were various alternatives for computing the monetary value of a customer. One was to use the total amount spent by a customer. It was discarded because It did not take into account the time lapse of the purchases. Another option was to use purchases' mean value but It was also discarded as It didn't take into account time lapse between first and last purchase.

A customer who spent 300 € in two purchases done in two months is better than another spent the same amount in two purchases done in 4 months.

Using the final chosen monetary value variable, the first customer of the previous example has twice the score than the second one. If we had used only the total value spent by customers or their purchases' mean value as the monetary value the classification would have been wrongly equal.

An issue with the monetary value proposed is that it could overestimates customers that were "born" close to the end of the observation period. In this case their purchase value would be divided by little elapsed time.

In order to solve this problem, we divided customers in two groups depending on the period they were "born" and re-defined the monetary value depending on these groups.

1. Customers that have born inside the "recency zone":

$$Monetary\ value = \frac{totalvalue}{T_{MAX}80\%} \quad \text{(Eq. 3.2)}$$

Where $T_{MAX}80\%$: It is the time that corresponds to the 80% percentile of the maximum time

between purchases distribution for all customers, in this case 312 days.

2. Customers that have born outside the recency zone.

$$Monetary\ value = \frac{totalvalue}{t.firstpurchase} \quad \text{(Eq. 3.3)}$$

For the second group we use the original variable but for customers whose first purchase was done inside the recency zone we divide their *totalvalue* by the overall population $T_{MAX}80\%$. It is shown in Figure 3.5.

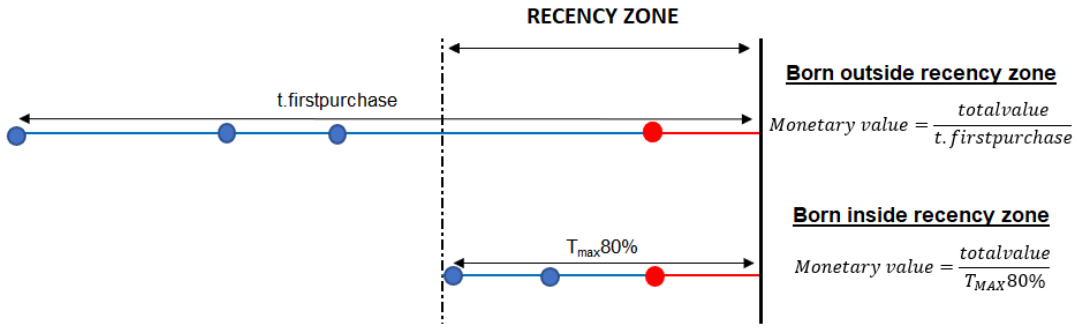


Figure 3.5: Two groups of monetary values representation

Thanks to this change we can assure that there are no customers which get an overestimation of its monetary value due a "short" active time.

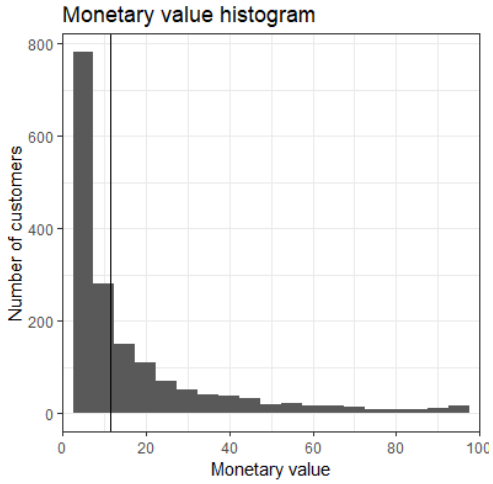


Figure 3.6: Monetary value histogram. Monetary value limited between 0 and 100 €/month

Once the monetary value variable is defined, we calculate it for each customer. Figure 3.6 shows the distribution of the variable and its 60 % percentile for customers with more than one purchase. This percentile will be the threshold for good monetary value customers. The distribution is positively skewed with a long tail for high values. Median monetary value is 7.63 €/month.

The 60% percentile of the monetary value distribution is used as a threshold for a good monetary value customer and it corresponds to 11.50 €/month or 138 €/year.

3.4. Good customers

Once both parameters are defined and calculated for each customer in the dataset there are four main possible customer's groupings. They are represented in Figure 3.7.

Based on the previous criteria a good customer is a customer which belongs to the 4th group. He is alive and his monetary value parameter is above 138 €/year, the 60% percentile of the monetary value distribution for all the more than one purchase customers in the dataset.

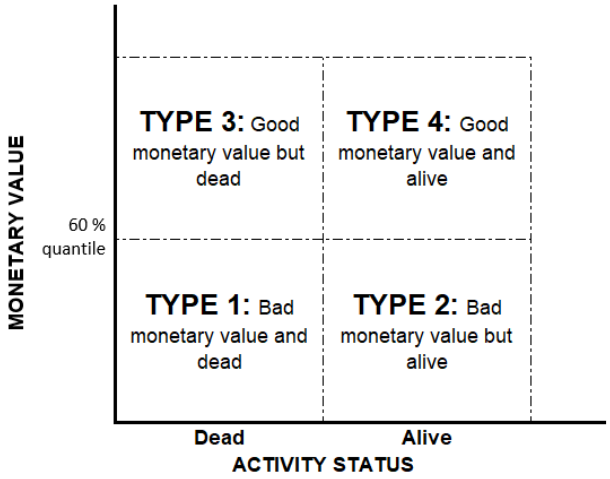


Figure 3.7: Main types of customers representation

For customers with NA active status two groups can also be defined depending on their monetary value. If their monetary value is below 138 €/year they are classified as bad customers. If their monetary value is higher than 138 €/year they cannot classified as good customers given their NA active status so they are classified as NA customers. These customers will be removed from the dataset as are neither useful model construction nor model testing.

Figure 3.8 and Table 3.1 show the distribution of customers in the above mentioned groups and their total number.

Figure 3.8: Types of customers representation

		MONEATARY VALUE	
		FALSE	TRUE
ACTIVITY STATUS	DEAD	2914	394
	ALIVE	722	**670**
	NA	1170	228

Table 3.1: Type of customers table

21.2 % of customers, 1,292 individuals, have good monetary value. 52% of them are alive, 670, and the remaining 622 are dead or not classified. The 670 alive customers with good monetary value represent 11% of all the customers in the dataset and are the good customers of the models.

Customers that are alive have significative better monetary value as can be seen in Figure 3.9. Monetary value median for alive customers is 10.53 €/month and for dead ones is 1.75 €//month.

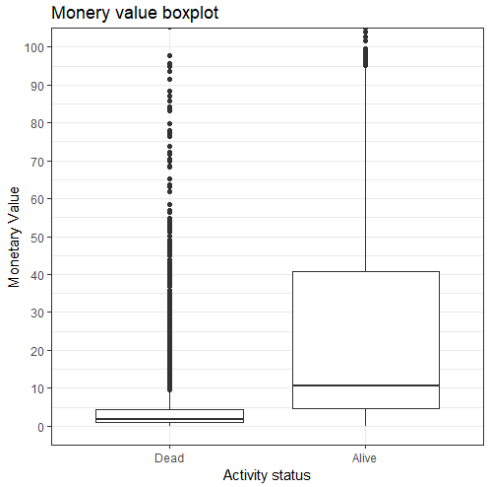


Figure 3.9: Boxplot of monetary value for alive and dead customers. Monetary value limited between 0 and 100 €/month

As we have stated before NA customers will be removed from the datasets used for the model construction. The final masterCustomers dataset has 5,780 customers from the total 6,098 new customers.

4. Predictor variables

The predictors variables are a fundamental part of the model construction process. Predictors are used to obtain the models response, in our case determine if a customer is good. Predictors are also called X variables while Response are named Y variables. In this section, we are going to explain which are the X variables that we have considered as interesting and how they are computed.

Variables included in the original client dataset can be directly used as an X variable but most come from transformations of the original ones. In this project, most of the predictors used in the model construction are obtained from transformations of the original variables in the tickets dataset. All of the new variables created were chosen before the start of the model construction process when we did not know whether they would become significant. We will check for their significance in future stages.

We have classified them in two groups: Purchasing behavior variables and Demographic variables.

4.1. Purchasing behavior variables

Purchasing behavior variables are created based on the tickets dataset information. They give information about each client purchasing behavior.

All of the purchasing behavior variables except for the month and the season the first purchase was done will be different for the three models to be constructed. Each one of them is based in a different number of transactions and therefore in different amounts of data from the tickets dataset. In the following explanation it is shown separated by // the number purchases entered in each of the three models.

These variables are:

- **Total value**: Numeric Variable. It is the total amount spent by the customer in the 1st// 1st and 2nd //1st, 2nd and 3rd purchases minus its refunds if any.
- **Total Entropy**: Numeric variable. It measures the variability of each customers' purchases. Entropy can take values between 0 and 1, been 0 null variability and 1 maximum variability.
  Our model's entropy is based on families purchased and amount spent in each family. In the model we use entropy for the 1st //1st and 2nd // 1st, 2nd and 3rd purchase. Entropy is calculated using the following formula.

$$Entropy = \sum_{i=1}^{n} p_i \cdot log_n(p_i) \quad \text{(Eq. 4.1)}$$

Where p is the total amount spent in each family divided by the total amount of the purchase. n is the number of purchase's families.

For the entropy computation we could have used quantity instead of amount spent. We discarded this option given the large differences between skus and their prices. For example, let's say a customer buys 1 hydraulic pipe that costs 300 € and 2,000 screws with an unite price of 0.1€. With the final criteria entropy is 0.97. Using quantity in the calculation, entropy would be nearly 0. We believe entropy of 0.97 is more appropriate for describing that transaction's variability.

We could also have used skus rather than families in the entropy calculation. We chose families because there are skus in the same family with very similar characteristics and the variability calculated could have been overestimated.

- **Total lines:** Numeric variable. It indicates how many lines a customer bought in its 1st// 1st and 2nd // 1st, 2nd and 3rd purchases. It is the same information as number of skus as skus are not repeated in different lines for the same purchase.
- **Total families**: Numeric variable. It is the number of different families included in the 1st// 1st and 2nd // 1st, 2nd and 3rd purchases.
- **First purchase month**: Factor variable. It shows the month in which the first purchase was placed.
- **First purchase season**: Factor variable. It shows the season in which the first purchase was placed.
- **Any disc**: Logical variable: It is TRUE if there was any discount in the 1st // 1st and 2nd// 1st, 2nd and 3rd purchases.

The following purchasing variables are only entered in the two and three purchases models:

- **Value's increase:** Numeric variable. It is the delta in the value spent between the 2nd and 1st // 3rd and 1st purchase
- **Lines' increase:** Numeric variable. It is the delta in lines purchased between the 2nd and 1st // 3rd and 1st purchase
- **Families' increase:** Numeric variable. It is the delta in families purchased between 2nd and 1st // 3rd and 1st purchase
- **Entropy's increase:** Numeric variable. It is the delta in the entropy between the 2nd and 1st // 3rd and 1st purchase.
- **Any refund:** Logical variable. It is TRUE if the customers have made any refund between the 2nd and 1st// 3rd and 1st purchase.

4.2. Demographic variables

Demographic variables are completely independent from the tickets dataset. They are based on the masterCustomers dataset. They do not give information about purchase behavior but on the customer background. Two of these demographic variables have been computed on top of the masterCustomer dataset. Two are directly extracted from the dataset.

- **Same zone**: Logical variable. It is TRUE if the location of the customer is the same than the company's branch assigned to the customer. It is FALSE otherwise.
- **Dist**: Numeric variable. It is the distance between the company's branch and customer's province in kilometers.
- **Market:** Factor variable. It has 3 values "MRO", "OEM" or "SI".
- **Digital client:** Logical variable. It is TRUE if the customer made its registration using the company's webpage. It is FALSE in any other case.

Unlike most of the purchasing behavior variables, demographic variables will not vary in the different models as they are constant per customer.

As It can be seen the majority of variables have been created from the dataset and not directly extracted from them. Only 2 of the 16 variables have been obtained by this last method: market and digital client variable.

5. Models

We are going to construct three different models. One will try to identify good customers based on their first purchase. Another will do the same but with the information gathered from the two first purchases. The last one will include the relevant information from the three first purchases in its prediction. Each model is based on the tickets and masterCustomers datasets. The amount information included in each one will be different.

Purchasing information variables (except first purchase month and season) will change from model to model as tickets information included is different for each of them. Customer demographic information will remain constant in all models. Response variable Y, whether a customer is good (1) or bad (0) will not change. To be or not to be a good customer is constant for the three models.

The models are intended to classify a customer as good or bad just after they have completed their first, second or third purchase. Therefore, they only use the information that the company would have after the customer first, second or third purchase. For example, regardless whether the customer has placed more than two orders in the observation period, when using his information to build the two purchases model, only information from the first and second purchase will be used.

These are the characteristics of the three models:

1. **One purchase model**: For the first model only the first purchase's information will be used. Predictors X will be based on the first purchase. This model is intended to be used to classify new customers just after they place their first purchase. It is applicable to all the customers.
2. **Two purchases model:** Predictors, X, for this model will be based only on information from the first and second purchases information. This model is intended to be used for classification of new customers just after their second purchase, taking into account information from the first two purchases.
3. **Three purchases model**: Predictors, X, for this model will be based only on information from the first three purchases. This model is intended to be used for classification of new customers just after their third purchase, taking into account information from the first three purchases.

5.1. Logistic regression

Our model's response is a binary variable. It can only take 2 values, good (1) or bad (0) customer.

The statistical tool chosen to implement the discriminant is the Logistic Regression. It is a very successful methodology to predict Bernoulli responses based on covariates. Logistic regression allows individuals classification, in this case customers, between two groups. Logistic regression estimates the probability of a customer to be in group 1 (good customer) for a given linear combination of the customer predictor variables.

Logistic regression belongs to the group of Generalized Linear Models that assumes for a given response Y_i and a vector of predictors Xi:

1. Y_i follows an exponential family distribution (in our case binomial) depending on a parameter θ_i.
2. The distribution for all Y_i's is of the same form, only θ_i changes.
3. There is a transformation *g*, named link function, such that:

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad \text{(Eq. 5.1)}$$

Where $\mathbf{x}_i$ is the vector of individual i predictors and β a set of parameters. That is, the link function of the mean is linear to the predictors.

If Y_i follows a Bernoulli distribution, its mean is π_i. A very useful link function is the logit. It is useful as it simplifies parameter estimation and the interpretation of the results. The logit link is:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad \text{(Eq. 5.2)}$$

It implies the log odds is linear to the predictors.

Odds is the ratio between the probability of an individual being in one group vs the probability of being in the other. For instance, if in a group of 100 customers that bought product A in their first purchase, 20 eventually became good customers, the odds of that group are 0.25. That is, for each good customer that bought product A in his first purchase there are 4 bad ones.

Expressing the equation 5.2 with π_i as the dependent variable one has

$$\pi(x) = p = \frac{e^{\beta_0+\beta_1 \cdot x_1+\beta_2 \cdot x_2+\cdots}}{1+e^{\beta_0+\beta_1 \cdot x_1+\beta_2 \cdot x_2+\cdots}} \qquad \text{(Eq. 5.3)}$$

where β_i corresponds to each one of the independent variables' coefficients that are estimated in the logistic regression and $\pi(x)$ the estimated probability to be a good customer.

The logit transformation allows for responses that move between 0 and 1 as it is intended with the classifier.

Another property by which Logistic Regression was chosen is its ease of interpretation. Logistics regression results interpretation depends on whether the independent variable is continuous or dichotomous:

- Continuous variable: The odds of being a good customer for a customer with a Δx increase in the X variable is $e^{\beta_i \cdot \Delta x}$ times the odds for an equivalent customer but without that increase.
- Dichotomous variable: Switch from the reference level to the non-reference level gives an odds ratio of being a good customer of e^{β_i}.

Odds ratio is the ratio between the odds of being a good customer for clients with a given set of predictors X and the odds of other customers with another set of predictors. The relation is shown in Equation 5.4 where the upper and lower odds belong to the first and second customer group respectively.

$$OR = \frac{Odds_i}{Odds_0} = \frac{\frac{p_i}{1-p_i}}{\frac{p_0}{1-p_0}} \qquad \text{(Eq. 5.4)}$$

When using logistic regression it is important not to confuse odds with probabilities. Odds are defined as the number of good customers divided by the number of bad customers while probability is defined as the number of good customers divided by the total number of customers. Logistic regression's coefficients are related to odds and odds ratio. One need to inverse the logit link to obtain probabilities if needed.

Logistic regression parameters are obtained maximizing the likelihood function from the observed data. It is called Maximum Likelihood Estimation (MLE). MLE in logistic regression is the analogous of the *least squares'* minimization method used in linear regression models.

Logistic regression algorithm maximizes the product of each customers contribution to the likelihood function. The individual contribution is shown in Equation 5.5 where x_i is the linear combination of customer's independent variables and y_i the value of its dependent variable.

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad \text{(Eq. 5.5)}$$

It can be understood as if the customer is good, (where y is 1) the contribution to the likelihood is the probability assigned to the customer by the model. If the customer is bad (y equals 0) the contribution to the likelihood is one minus its probability to be a good customer. MLE looks for those coefficients that maximize the logarithm of the product of the individual contributions to the likelihood (see equation 5.6). In other words, the coefficient estimation algorithm tries to minimize the difference between each customer's Y variable and the assigned probability to be a good customer. Logarithm transformation of the original likelihood function is performed

for ease of use.

$$L(\beta) = ln[\prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}] \quad \text{(Eq. 5.6)}$$

We would also like to mention how new parameters significance is tested in the logistic regression model. Basically, what we do is to ask the following question: *"Does the model that includes the new variables tell us more about the outcome (or response) variable than a model that does not include them?*" **[1]** To answer this question, the following method is used:

First, we need to define what is called the Deviance of the model as shown in Equation 5.7.

$$D = 2 \cdot ln\left[\frac{(likelihood\ of\ the\ fitted\ model)}{(likelihood\ of\ the\ saturated\ model)}\right] \quad \text{(Eq. 5.7)}$$

The saturated model is the one that has as many predictors as observations. If the values of the outcome variable are either 0 or 1, the likelihood of the saturated model is 1. As it is our case, Deviance can be expressed as shows Equation 5.8

$$D = -2 \cdot \ln(likelihood\ of\ the\ fitted\ model) \quad \text{(Eq. 5.8)}$$

Then, the deviance of the old model is compared with the deviance of the newest one, which includes one or various new variables, as it is shown in equation 5.9. This difference can also be expressed as shown in equation 5.10 and 5.11. It is called statistic G

$$G = D(model\ without\ the\ variable) - D(model\ with\ the\ variable) \quad \text{(Eq. 5.9)}$$

$$G = 2(ln(likelihood\ model\ with\ the\ variables) - ln(likelihood\ model\ without\ the\ variable)) \quad \text{(Eq. 5.10)}$$

$$G = 2 \cdot \ln\left(\frac{likelihood\ with\ the\ new\ variables}{likelihood\ without\ the\ variables}\right) \quad \text{(Eq. 5.11)}$$

If the c new variables added to the model do not improve its fit, G follows a Chi-square distribution with c degrees of freedom. Its null hypothesis states that the c "slope" variables' coefficients are equal to zero.

The Chi-square test is run in order to check the significance of the variables and the test's p-value is obtained. If that p-vale is smaller than 0.05 we can reject the null hypothesis and state that the variables inclusion significantly improves the model. We will name this p-value as "p-value of inclusion" in the subsequent model construction explanations.

5.2. Model construction

This section explains step by step the construction of each one of the three models.

The first step was to divide each one of the datasets into two different ones, training and testing datasets. The training dataset with a random choice of 75% of the original's lines. The testing dataset with the remaining 25%. Both datasets have approximately the same proportion of good and bad customers.

The training dataset is used to perform the variable selection's stage. The main objective of this stage is to find a combination of variables that is capable of explaining well the set of observations but with the minimum possible number of variables. We try to minimize model's variables because, in general, the more variables entered into the model the greater the coefficient's standard errors and the more risk of obtaining an overfitted model.

An overfitted model is a model that is capable to predict well customers that have been used in the model's construction process but it is not capable to predict well new customers not seen before.

The methodology we have used in the variable selection is based in the one proposed by D. W. Hosmer and S. Lemeshow **[1]**:

1. **Univariate analysis**: The first step is to conduct an univariate analysis for each independent variable versus the dependent variable *good customer.*

   For nominal or discrete variables with few values we create a contingency table of the *good customer* variable versus each predictor variable. We also plot a mosaic plot to visualize the proportion of good customers in each of the predictor variables' levels. In addition, we perform an univariate logistic regression in order to study the coefficients, odds ratio, coefficient's significance and the "p-value of inclusion" for each of the variables.

   For continuous variables we fit a univariate logistic regression model to study the coefficients and the "p-value of inclusion". We also plot a smoothed scatterplot to visualize how the response variable varies when the predictor variable increases or decreases.

2. **First multivariate model:** Those variables from the previous univariate analysis with a p-value in their univariate logistic regression smaller than 0.25 ("*p-value of inclusion*") are going to be selected for the first multivariate model.

   Even though a p-value of 0.25 can seem a bit high, it is a recommendation seen in literature **[1].** It prevents the possibility that variables which have a weak relation with the outcome taken by themselves can become an important predictor when they are together with other

variables. Using the limit of 0.25 for the p-value and not the most commonly used 0.05 let these variables get into the multivariate analysis where we can study whether that phenomenon occurs or not.

3. **Second multivariate model:** In this multivariate model we refine the previous one until obtaining a model in which all the variables included have a "p-value of inclusion" smaller than 0.05.

When the model is obtained, we add back variables discarded at previous steps one by one. This last procedure is done in order to see if they are now significant in a 5% level. This model is called "preliminary main effects model"

4. **Interactions:** Once the "preliminary main effects model" is defined, we check for Interactions between variables. In order to check it, we have evaluated the *"p-value of inclusion"* to the "preliminary main effects model" of each one of the possible interactions between each pair of variables. Interactions that are significant in a 10% level are selected for inclusion in the next multivariate model.

5. **Final multivariate model:** Finally, we include all the interactions found significant in a 10% level to the "preliminary main effects model". When it is done, we check for its "*p-value of inclusion*" and discard all the interactions that are not significant at a 5 % level. That model is refined until obtaining a model in which all main effects and interactions variables have a "p-value of inclusion" smaller than 0.05.
That final multivariate model is called "Preliminary final model".

When the "Preliminary final model" is obtained we proceed to carry out a residual diagnostic. Its objective is observing for outliers in customers patterns referenced to the model. If the residual diagnostics identifies outliers worth removing the training dataset is cleaned and the model recalculated.

At this moment, the model is run again with the testing dataset, which have not been used in any previous step. Test dataset is used to assess the model accuracy. To assess the fit of the data and the potential model overfit we use the Area Under the Curve (AUC) of the ROC curve, the classifier sensibility and specificity and the Hosmer Lemeshow test.

In order to check if the model is overfitted we compare the tests mentioned above using the testing and training datasets. If the results are similar (whether they were good or bad) we can say that the model is not overfitted. It means that model can predict customers which has not been used in the model construction and not only customers that have been used in it.

Finally, once the overfitting and the fit of the data has been assessed we calculate the final model's coefficients. It is done by running the "Preliminary final model" using the original

dataset which includes both testing and training datasets.

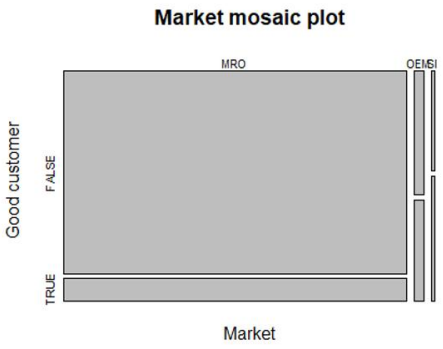
That algorithm is followed for each one of the three models that are going to be constructed. Appendix A shows the One purchase model construction. Two and three purchase model constructions are not reported as the procedures are identical.

5.3. Univariate analysis

In this section, we are going to explain the univariate analysis of each X variables versus the good customer response variable. Only those variables deemed significative will be shown.

After carrying out the three model's construction process, we realized that all 3 univariate analyses had very similar results. With the purpose of not repeating the same information, only one univariate analysis will be explained for each variable. Its results can be extrapolated to the remaining models.

We have chosen the one purchase model to show most of the univariate analysis' explanation because it contains the larger number of customers. Variables that can only be included in the two or three purchase model such as refunds or deltas between purchases are going to be explained using the two purchases model.



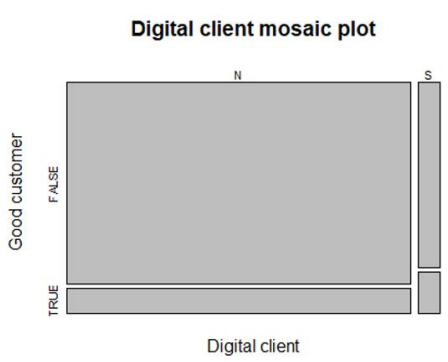

```
Call:
glm(formula = good.customer ~ market, family = "binomial",
data = masterCustomers.1pm.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2735  -0.4592  -0.4592  -0.4592   2.1456

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.19644    0.05119  -42.91  < 2e-16 ***
marketOEM    1.99577    0.19050   10.48  < 2e-16 ***
marketSI     2.41958    0.30434    7.95 1.86e-15 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3128.4  on 4401  degrees of freedom
Residual deviance: 2983.1  on 4399  degrees of freedom
AIC: 2989.1

Number of Fisher Scoring iterations: 4
```

Figure 5.1: Market univariate analysis

First, It has been conducted the ***market*** vs *good.customer* univariate analysis. As It can be seen in the mosaic plot[2] shown in Figure 5.1 both SI and OEM markets have a much larger proportion of good customers than the MRO market group. Only 10 % of MRO customers are good while 45 % and 56 % of OEM and SI customers respectively are good customers.

[2] Mosaic plot: It gives information about the proportion of good customers in each variable's level (vertical blocs) and the proportion of customers in each group (horizontal blocs).

OEM and SI customers have an odds ratio of being good of 7.36 (5.06, 10.69) and 11.24 (6.19, 20.41) respectively, being MRO the reference group. These odds ratio are very large so it is possible that market variable could enter to the multivariate model although the number of SI and OEM customers is very low compared with the number of MRO's customers.




```
Call:
glm(formula = good.customer ~ digita_cli, family = "binomial",
    data = masterCustomers.1pm.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6296  -0.4829  -0.4829  -0.4829   2.1009

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.09037    0.04969 -42.066  < 2e-16 ***
digita_cliYES  0.57250    0.16694   3.429 0.000605 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3128.4  on 4401  degrees of freedom
Residual deviance: 3117.9  on 4400  degrees of freedom
AIC: 3121.9

Number of Fisher Scoring iterations: 4
```

Figure 5.2: Digital client univariate analysis

Similar results can be seen in the **digital client** univariate analysis. The number of digital customers is only the 6.1 % of the total. 18 % of the digital clients are good while only 11 % of non-digital customers are good. The odds ratio of being good customer is 1.77 (1.27, 2.46), taking non-digital customers as the reference group.

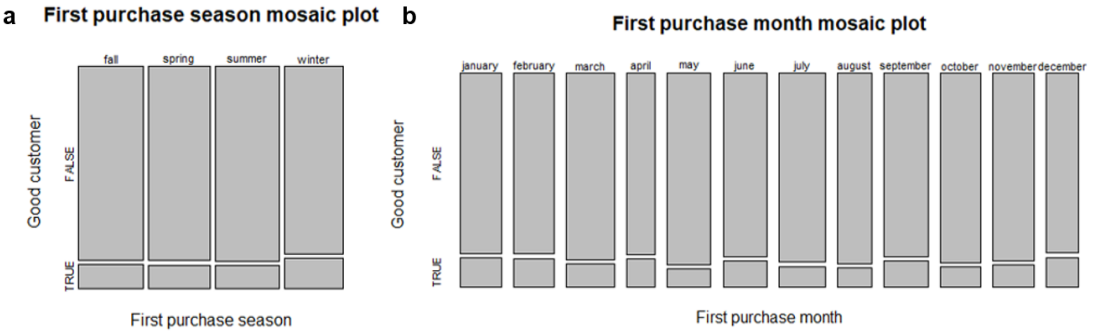


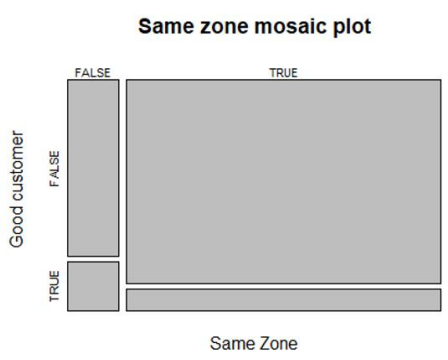
Figure 5.3: a) First purchase season mosaic plot, b) First purchase month mosaic plot.

Variables *fp.month* and *fp.season* have more homogeneous distribution in the number of customers in each group than in the variables studied above

The most significant **season** is winter with an odds ratio of 1.33 (1.03, 1.72) using fall as reference group. Winter is the season which a greater proportion of good customers: 14 %.

With regards to the first purchase **month** variable, August and May are the most significant ones. In this case customers which have placed its first purchase during these months are less likely to be good customers. Odds ratio are 0.61 (0.38, 0.99) and 0.58 (0.36, 0.91) for August and May respectively. We think it can be related with holyday periods. Customers who place

their first purchase in summer vacations or during the Easter Holiday are less likely to be good customers than those who place their first purchase in the remaining months.




```
Call:
glm(formula = good.customer ~ same.zone, family = "binomial",
    data = masterCustomers.1pm.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7002  -0.4525  -0.4525  -0.4525   2.1585

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.28093    0.09729 -13.166   <2e-16 ***
same.zoneTRUE -0.94633    0.11169  -8.473   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3128.4  on 4401  degrees of freedom
Residual deviance: 3063.9  on 4400  degrees of freedom
AIC: 3067.9

Number of Fisher Scoring iterations: 4
```

Figure 5.4: Same.zone Univariate Analysis

Figure 5.4 shows the univariate analysis for ***same zone*** variable. Customers who do not belong to the same province than the company branch they contacted are more likely to be good customers than the ones located in the same province. Odds ratio for same zone's customers is 0.39 (0.31, 0.48) using same zone variable FALSE as the reference group. It can also be stated that the assumption made in the Exploratory Analysis about that most of the customers are assigned to branches in their location is correct: they represent the 86% of the customers.

Our hypothesis is that a customer which is far from the company branch has more interest in purchasing from that company than customers from the same zone. It is possible that customers from the same zone contact the company only because it is the distributor with whom they had easiest access. On the other hand, customers from another province are likely to have done previous research in distributors and chosen this one because they were really interested in it.

Univariate analysis of **distance** variable returns a consistent result with those obtained at the same zone variable analysis: When the distance increases the probability of being a good customer also increases. Just to show this relation see the following cases: for a 100 km distance the odds ratio of being a good customer takes a value of 1.05 (1.03, 1.07), for 250 km of 1.13 (1.07, 1.18) and for 500km of 1.27 (1.15, 1.40).

With regards to the **total value** variable (in this case the value of the first purchase as it is the one purchase model) we observe in the smoothed scatterplot shown in Figure 5.5.a a possible logarithmic relation with the number of good customers in each group.

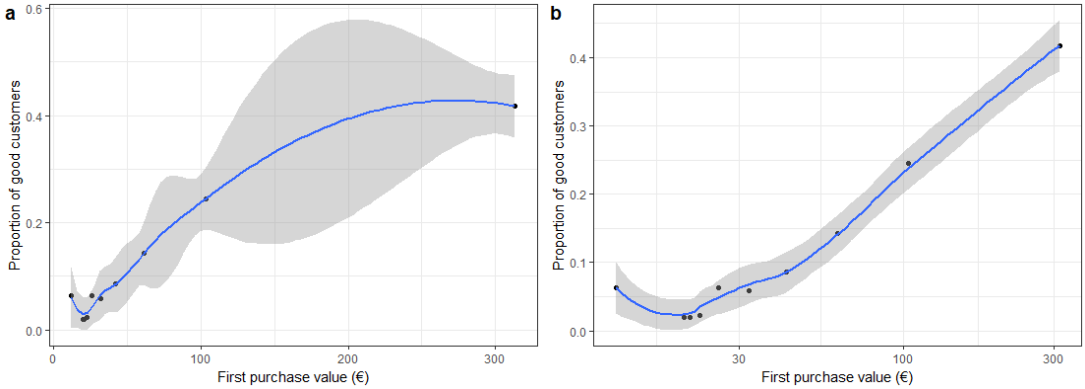
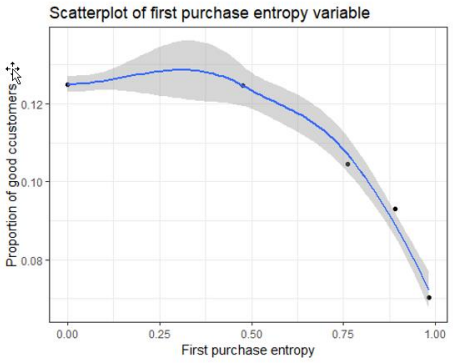


Figure 5.5: a) Scatterplot of the proportion of good customers vs first purchase value, b) Scatterplot of the proportion of good customers vs first purchase value, using log10 scale

Figure 5.5.b shows the same scatterplot using $\log_{10}$ scale at x axis. Except the first point we can see that the hypothesis of logarithmic behavior seems correct. Because of that, we are going to use $\log_{10}$(totalvalue) instead of totalvalue variable in the multivariate logistic analyses.

Log_{10}(totalvalue) versus *good.customer* univariate analysis result shows a significant relation between both variables. Odds ratio is 6.59 (5.45, 7.98) when the amount spent in the first purchase is multiplied by 10.




```
Call:
glm(formula = good.customer ~ firstp.entropy, family = "binomial",
    data = masterCustomers.1pm.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5224  -0.5224  -0.5224  -0.4320   2.2190

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.92291    0.05687 -33.811  < 2e-16 ***
firstp.entropy -0.44995    0.12491  -3.602 0.000315 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3128.4  on 4401  degrees of freedom
Residual deviance: 3114.9  on 4400  degrees of freedom
AIC: 3118.9

Number of Fisher Scoring iterations: 5
```

Figure 5.6: First purchase entropy univariate analysis

In Figure 5.6 it can be seen the univariate analysis for the **first purchase's entropy** variable. There is a decrease in the proportion of good customers when entropy increases. It is important to note that 60% of the customers have a value of 0 in this variable which means that they have only purchased one family in its first transaction.

The negative coefficients of the logistic regresion analys means that when entropy increases the odds of being a good customer decreases. For a 0.1 increase in entroy the odds ratio is 0.96 (0.93, 0.98) and for a 1 point increase (which is the maximum possible) the odds ratio decreases to 0.64 (0.50, 0.81).

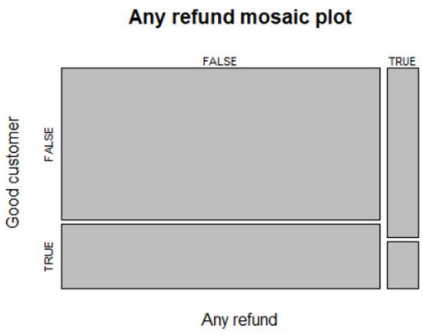
```
Call:
glm(formula = good.customer ~ firstp.lines, family = "binomial",
    data = masterCustomers.1pm.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0569  -0.5022  -0.4801  -0.4589   2.1462

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.29320    0.06892 -33.271  < 2e-16 ***
firstp.lines  0.09538    0.01835   5.197 2.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3128.4  on 4401  degrees of freedom
Residual deviance: 3101.7  on 4400  degrees of freedom
AIC: 3105.7

Number of Fisher Scoring iterations: 4
```

Figure 5.7: First purchase lines univariate analysis

Finally, univariate analysis for the **lines of the first purchase** states that when a customer purchases a larger number of skus its odds of being a good customer increase. For a one line increase the odds ratio is only 1.1 (1.06, 1.14) but when the lines number's increase is of 10, the odds ratio rise to 2.59 (1.81, 3.72).

The variables *any.disc* and *total.families* are not significant for a 5% level in the One purchase model univariate analysis. They are also not significant in the second and third models.

When variables that are only present in the Two purchases and Three purchases models are studied It is found that only *any.refund* variable is significative. All the variables of increase have a "p-value of iclusion" greater than 0.05.




```
Call:
glm(formula = good.customer ~ any.refund, family = "binomial",
    data = masterCustomers.2pm.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8384  -0.8384  -0.8384   1.5597   1.7564

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.8649     0.0547 -15.812   <2e-16 ***
any.refundTRUE  -0.4371     0.2010  -2.174   0.0297 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2118.4  on 1761  degrees of freedom
Residual deviance: 2113.4  on 1760  degrees of freedom
AIC: 2117.4

Number of Fisher Scoring iterations: 4
```

Figure 5.8: Any refund univariate analysis

The analysis for **Any.refund** variable states that a customer who has made any refund between the first and second purchase has less probability to be a good customer than those who have not. Odds ratio of being a good customer if any refund variable is TRUE takes a value of 0.65 (0.44, 0.96).

This results can also be observed in the mossaic plot: Customers who have made any refund

are the 9% of the total and its percentatge of good customers is of 21.38%. Conversely, 29% of customers who have not made any refund are classified as good customer.

To summarize, It can be seen in Table 5.1 which are the variables that increase and decrease odds of being a good customers according the univariate analysis. It is also shown the odds ratio of being a good customer for each variable

Increase odds		Decrease odds	
Variable	**OR**	**Variable**	**OR**
marketOEM	7.36	same.zone	0.39
marketSI	11.24	fp.monthmay	0.58
digital client	1.77	fp.monthaugust	0.61
fp.season.winter	1.33	firstp.entorpy	0.64***
firstp.lines	2.59 *	any.refund	0.65
$\log_{10}$(totalvalue)	6.59**		

Table 5.1: Summary of effect on the odds for each variable.
** For 10 lines increase. ** For x10 in the total amount spent. *** For a 1-point increase in the entropy.*

5.4. One purchase model

Model construction procedure for this model can be seen in the R-markdown located in Appendix A. Even so, we want to show first the residual diagnostics as we found an extreme value that was removed previous to run the model again.

Figure 5.9.a shows how much model's deviance varies when each of the customers are excluded from the training dataset. There is a customer with a very extreme sΔD value of 24.38.

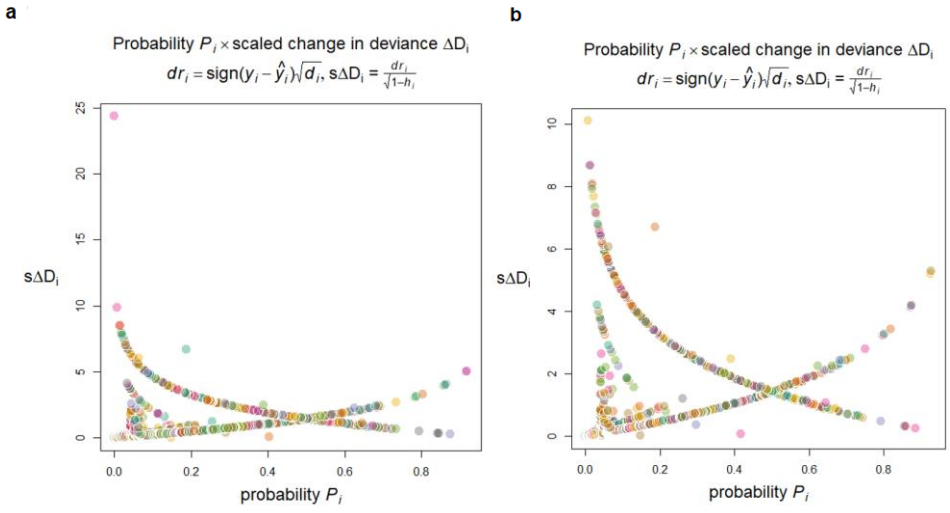


Figure 5.9: a) sΔD vs pi before diagnostics, b) sΔD vs pi after diagnostics

This customer has a very low value in its first purchase but subsequent purchases led to the

customer classification as a good one. In this case, a customer with very low first purchase value, and hence very low probability of being a good customer under the One purchase model, was classified as a good customer. It leads to that very high decrease in deviance when it is removed from the dataset. Tickets dataset's lines of this customer can be seen in Table 5.2.

customer_id	n_container	n_delivery_note	date_deliverynote	sku	family	qty	price	disc	amount	up
732066	5678346	255386	2016-12-29	476514	502	4	0.0001	0	0.0004	1
732066	6175635	475017	2018-06-11	551891	504	50	6.5000	0	325.0000	1
732066	6277449	521485	2018-10-18	570551	502	20	7.3200	0	146.4000	1
732066	6330003	544803	2018-12-18	570551	502	20	7.3200	0	146.4000	1

Table 5.2: Tickets dataset's lines of the extreme customer

Review with the company sales group reflected that the transaction made on the 2016-12-29 was a gift and not a purchase. His first purchase should have been the one on 2018-06-11 instead. We decided to remove the customer from the dataset and recalculate the model excluding him.

Figure 5.9.b shows the new diagnostics plot once the customer was removed. The diagnostic is now acceptable as there are no extreme values and only the 3.45% of the customers have an sΔD greater than 4, which is close to the upper ninety-fifth percentile of the distribution as expected **[1]**

5.4.1. Model analysis

The model obtained for the one purchase analysis is presented in Table 5.3. One purchase model includes two dichotomous variables (*digita_cli* and *fp.seasonWINTER)*, a categorical variable (*market: reference MRO, OEM and SI)*, a continuous variable (log_{10}*(totalvalue)*) and their interaction.

Variable	Coefficient	Std. Error	z-value	P > |z|	95% CI
Intercept	-5.939	0.188	-31.561	<0.001	(-6.308, -5.570)
marketOEM	4.813	0.486	9.914	<0.001	(3.862, 5.765)
marketSI	5.385	0.878	6.131	<0.001	(3.663, 7.106)
digita_cli	0.331	0.165	2.011	0.044	(0.008, 0.654)
fp.seasonWINTER	0.254	0.100	2.525	0.012	(0.057, 0.450)
log10(totalvalue)	2.105	0.096	21.839	<0.001	(1.916, 2.294)
marketOEM x log10(totalvalue)	-1.699	0.222	-7.669	<0.001	(-2.133, -1.265)
marketSI x log10(totalvalue)	-1.814	0.435	-4.165	<0.001	(-2.667, -0.960)

Table 5.3: Estimated Coefficients, Standard Errors, z value, Wald test p-value and 95% Confidence Intervals for the coefficients of the One purchase model Logistic Regression

Table 5.4 shows the odds ratio for the two dichotomous variables that do not interact. The odds of being a good customer being a digital client is 1.39 (1.01, 1.92) times larger than when

being non-digital. The same reasoning can be done for the *fp.seasonWINTER* variable that also increases odds of being a good customer.

The readings is: a digital client or a client who has placed its first purchase during winter is more likely to be a good customer than a non-digital one or one who placed the first purchase any other season.

Variable	Value	Odds Ratio	95% CI
Digital client			
	False	1.00	
	True	1.39	1.01, 1.92
Fp.seasonwinter			
	False	1.00	
	True	1.29	1.06, 1.57

Table 5.4: Estimated odds ratio and 95% Confidence Intervals for digita_cli and fp.seasonWINTER

Figure 5.10 represents the change in the odds ratio when the first purchase value is multiplied by two, five or ten for three customers, each one of them belonging to a different market group. The odds ratio for MRO market customers increase with a much higher rate than the odds ratio of OEM and SI customers when the first purchase value increases.

For example, an MRO customer whose first purchase is 5 times larger than another one is 4.34 (3.94, 4.77) times more likely to be a good customer. This value happens to be 1.32 (1.09, 1.60) for OEM customers and 1.22 (0.81, 1.84) for customers which belongs to SI market.

Confidence intervals for the odds ratio covers the value 1.00 when the first purchase is five or ten times greater for SI and OEM customers respectively. For these factor's intervals we cannot say that the interaction is significative as there is a point in the confidence interval in which the odds of being a good customer does not increase nor decrease.

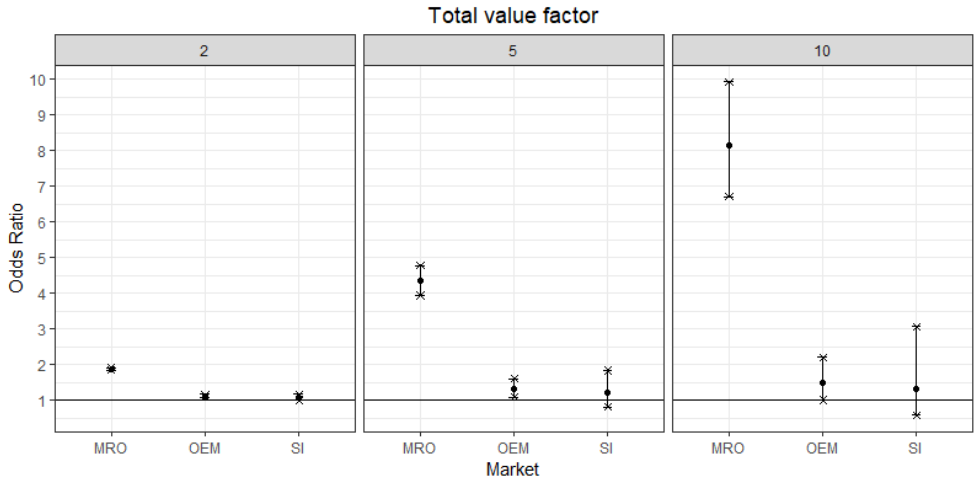


Figure 5.10: Odds ratio and 95% confidence limits when first purchase is multiplied by 2, 5 or 10 for different types of markets

Conversely, it can be seen in Figure 5.11 that although odds ratio increase with a higher rate for MRO customers when the amount spent increases, its estimated probabilities of being a good customer are lower than for the OEM or SI market customers for small to medium amounts of first purchase value. This is caused by marketOEM and marketSI independent coefficients, that increase base probabilities of being a good customer.

Figure 5.11 represents a comparison between three similar clients in which only the type of market assigned varies. All customers are non-digital and have not placed its first purchase in winter. Graphs show probability of being a good customer when to total amount increases and their 95% confidence interval.

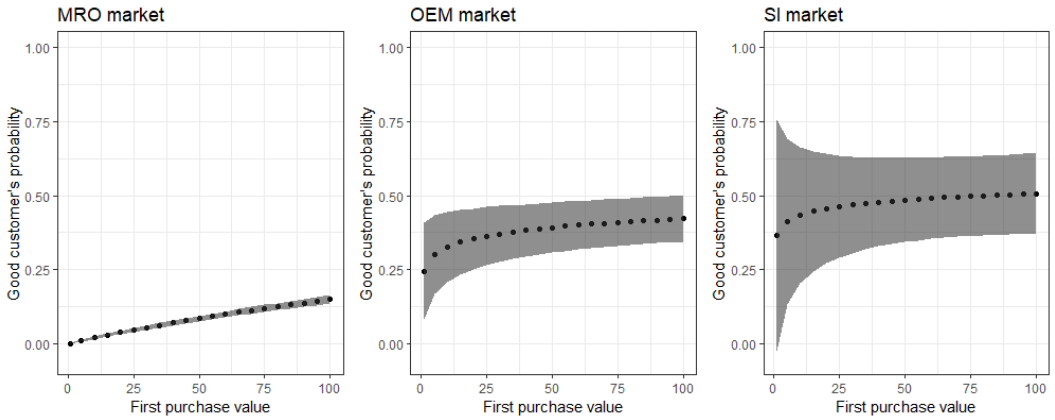


Figure 5.11: Probability of good customer vs first purchase value for non-digital and not "born" in winter customers for the three different types of markets

Figure 5.12 represents a graphical overview for the one purchase model behavior. The variables that most affect the probability of being a good customer is the total value spent in the first purchase and the customer's market. The probability is impacted by two more variables: *digita_cli* and *fp.seasonWINTER*.

It is easy to visualize the effect of the market x $\log_{10}$(totalvalue) interaction mentioned above. Customers which do not belong to MRO market have larger probabilities of being good customers when the amount spent is medium or small. As the total value spent increases their probabilities increase with a smaller slope. Therefore, when a client spends more than 1000€[3] in his first purchase his probability of being a good customer is larger for MRO customers than for SI or OEM customers.

It can also be seen the effect of *digita_cli* and *fp.seasonWINTER*. For customers with the same amount spent in the first purchase and same market the probability of being a good customer are higher for those who are digital or placed their first order in the winter period.

[3] 1000€ are equivalent to $\log_{10}$(totalvalue) equals 2.

Figure 5.12: Graphical overview for the One purchase's model

Finally, Table 5.5 shows the Deviance analysis for this model. For each variable it shows: 1) how many degrees of freedom it removes from the model, 2) what is the decrease in deviance due to the introduction of the variable, 3) and 4) total number of degrees of freedom and deviance once the variable has been added and 5) the p-value for the null hypothesis including the variable does not improve the model. The table must be read from top to bottom, as it indicates how the model varies when including each one of the variables with respect to the model including all the variables above.

Variable added	Degrees of freedom	Deviance decrease	Resid. Df	Resid. Dev	"p-value of inclusion"
NULL model	-	-	5868	4164	-
log10(totalvalue)	1	590	5867	3574	<0.001
market	2	81	5865	3493	<0.001
fp.seasonwinter	1	7	5864	3486	0.009
digita_cli	1	5	5863	3481	0.025
log10(totalvalue):market	2	57	5861	3424	<0.001

Table 5.5: One purchase model Analysis of Deviance

As we can see, all the variables included are significant as they have a p-value smaller than 0.05. *Log_{10}(totalvalue)* is by far the variable which a larger decrease in deviance, i.e. the variable that better explains it. It is followed by *market* and by *fp.seasonwinter* and *digita_cli* with a much lower deviance decrease.

5.5. Two purchases model

5.5.1. Model analysis

Table 5.6 shows a summary of the Two purchases model. It shows the following changes compared to the first one: it includes the continuous variable *total.lines* and the interactions *digital_cli* x *any.refund* and *digital_cli* x log_{10}*(totalvalue).*

Variable	Coefficient	Std. Error	z-value	P > |z|	95% CI
Intercept	-6.420	0.311	-20.633	<0.001	(-7.030, -5.810)
marketOEM	5.464	0.752	7.269	<0.001	(3.991, 6.938)
marketSI	4.108	2.040	2.013	0.044	(0.109, 8.107)
digita_cli	-3.762	1.750	-2.150	0.032	(-7.192, -0.333)
fp.seasonWINTER	0.269	0.121	2.222	0.026	(0.032, 0.506)
log10(totalvalue)	2.445	0.143	17.054	<0.001	(2.164, 2.726)
any.refund	-0.367	0.208	-1.767	0.077	(-0.774, 0.040)
total.lines	0.036	0.014	2.556	0.011	(0.008, 0.063)
marketOEM x log10(totalvalue)	-2.017	0.295	-6.826	<0.001	(-2.596, -1.438)
marketSI x log10(totalvalue)	-1.015	0.940	-1.079	0.280	(-2.858, 0.828)
digita_cli x log10(totalvalue)	2.100	0.819	2.563	0.010	(0.494, 3.706)
digita_cli x any.refund	-2.729	1.378	-1.980	0.048	(-5.430, -0.027)

Table 5.6: Estimated Coefficients, Standard Errors, z value, Wald test p-value and 95% Confidence Intervals for the coefficients of the Two purchases model Logistic Regression

Fp.seasonwinter, *market* and the interaction between *market* and log_{10}*(totalvalue)* remain.

Only the effects of the new variables are going to be explained in this chapter. The effect of the repeated variables is the same than in the previous model.

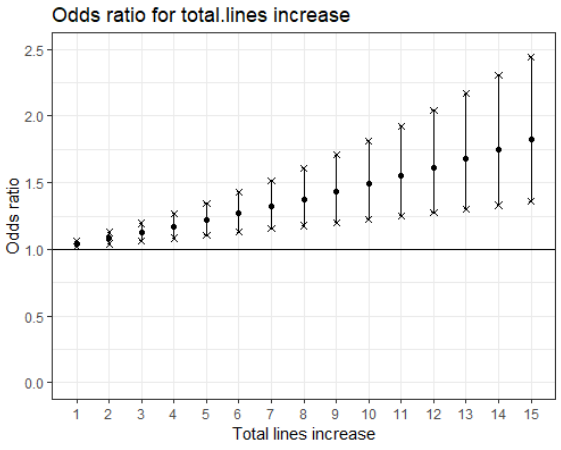


Figure 5.13: Estimated odds ratio and 95% confidence limits for total lines increase

The *total.lines* coefficient gives the change in the log odds for an increase of one line in the two first purchases. For a Δc increase in the number of lines purchased the odds ratio is

$e^{\Delta c \cdot 0.04}$. Figure 5.13 shows how the odds ratio changes as a function of the number of lines increase. For example, a customer who has purchased 10 lines more than another similar one has 1.5 (1.25,1.80) larger odds of being a good customer than the second one.

Table 5.7 shows the effect of the interaction between *digita_cli* and *any.refund* variables. The interpretations of the results are that a non-digital client which has made any refund between the first and second purchase is 0.69 (0.46, 1.04) times likely to be a good customer than another who has not made any refund. If the customer who has made the refund is digital the odds ratio decreases to 0.05 (0.00, 0.65), which is a very important drop.

However, the confidence interval of odds ratio for the first case includes value 1.0 so the interaction cannot be interpreted as significative.

Digital client	Any refund	Odds Ratio	95% CI
False			
	False	1.00	
	True	0.69	0.46, 1.04
True			
	False	1.00	
	True	0.05	0.00, 0.65

Table 5.7: Estimated odds ratios and 95% Confidence Interval for any.refund within digita_cli

Digita_cli variable also interacts with *log_{10}(totalvalue).* The coefficient of the interaction can be interpreted as an increase in the slope of *log_{10}(totalvalue)* when the customer is digital. This is opposite to the one purchase model interaction between market and l*og_{10}(totalvalue.)* In this case the log odds ratio increases, but for low values of *log_{10}(totalvalue)* the odds of being a good customer are very similar. This is due to the fact that the interaction coefficient is positive while the main effect *digita_cli* coefficient is negative.

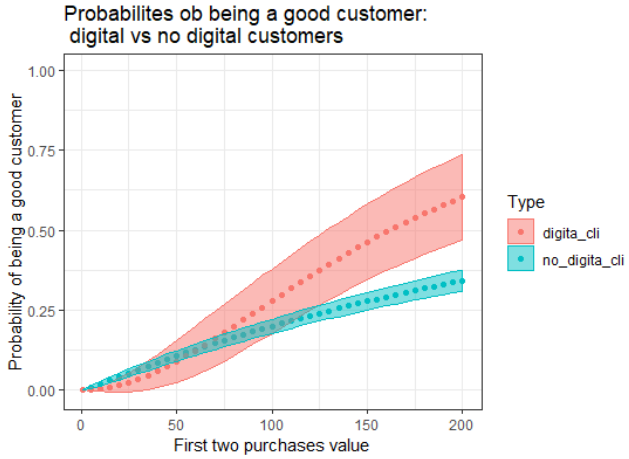


Figure 5.14: Probability of being a good customer vs Purchase value for two MRO customers who have not placed their first purchase in winter, have 4 lines in their two first purchases and have not made any.refund.

The interaction effect is shown in Figure 5.14. It displays the probability of being a good customer for a digital and non-digital customer as a function of the customer first two purchases total value. It can be observed that for small values the probability of being a good customer for non-digital customers is very similar than for digital ones. But when total amount reaches a value of 125 € digital customers are rated higher in their probability of being good.

Table 5.8 shows the Analysis of Deviance for the second model. As in the previous model, all the variables are statistically significant.

Variable added	Degrees of freedom	Deviance decrease	Resid. Df	Resid. Dev	"p-value of inclusion"
NULL model	-	-	2349	2809	-
log10(totalvalue)	1	527	2348	2282	<0.001
market	2	28	2346	2254	<0.001
digita_cli	1	9	2345	2245	0.002
total.lines	1	8	2344	2237	0.005
any.refund	1	5	2343	2232	0.020
fp.seasonwinter	1	5	2342	2226	0.023
log10(totalvalue):market	2	41	2340	2186	<0.001
log10(totalvalue):digita_cli	1	7	2339	2178	0.007
digita_cli:any.refund	1	5	2338	2173	0.027

Table 5.8: Two purchases model Analysis of Deviance

5.6. Three purchases model

5.6.1. Model analysis

Table 5.9 shows the model for three purchases. Variables included in that model are very similar to the two purchases one. Contrary to the two purchase model, *valueincrease* is now significant while *any.refund* and the interaction between *any.refund* and *digital_cli* are not. All other variables and interactions remain in the three models.

Variable	Coefficient	Std. Error	z-value	P > |z|	95% CI
Intercept	-6.369	0.431	-14.790	<0.001	(-7.213, -5.525)
marketOEM	4.425	1.189	3.723	<0.001	(2.095, 6.754)
marketSI	3.238	2.169	1.493	0.136	(-1.014, 7.490)
digita_cli	-3.942	2.303	-1.712	0.087	(-.8.456, 0.572)
log10(totalvalue)	2.450	0.184	13.312	<0.001	(2.090, 2.811)
total.lines	0.036	0.014	2.511	0.012	(0.008, 0.064)
valueincrease	4.7e-4	1.3e-4	3.592	<0.001	(2.1e-4, 7.3e-4)
marketOEM x log10(totalvalue)	-1.549	0.442	-3.500	<0.001	(-2.416, -0.681)
marketSI x log10(totalvalue)	-0.660	0.916	-0.720	0.471	(-2.455, 1.135)
digita_cli x log10(totalvalue)	2.006	0.995	2.016	0.044	(0.056, 3.956)

Table 5.9: Estimated Coefficients, Standard Errors, z value, Wald test p-value and 95% Confidence Intervals for the coefficients of the Three purchases model Logistic Regression

Figure 5.15 shows how the odds ratio changes as a function of delta between third and first purchase. Value increase is a continuous variable that do not interact with any other one. *valueincrease* coefficient can be interpreted as a Δv increase in the amount spent by the customer between the first and third purchase gives an odds ratio of being a good customer of $e^{\Delta x \cdot 0.000471}$.

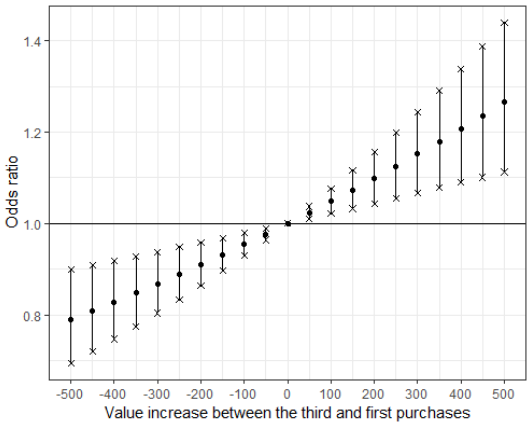


Figure 5.15: Odds ratio with 95% confidence limits for difference of purchase value between third and first purchase

When the value of the third purchase is smaller than the first one, (*valueincrease* is negative) the odds ratio take values smaller than 1.00. Customers in that situation are less likely to be good ones. Conversely, when customers spend more in the third purchase, their odds ratio are positive so they are more likely to be good customers. For example, a drop of -300 € in the difference between third and first purchase increases the customer odds ratio by 0.87 (0.80, 0.94). If there is an increase of 300 € the odds between these two types of customer also increases by a factor 1.15 (1.07, 1.24).

Variable added	Degrees of freedom	Deviance decrease	Resid. Df	Resid. Dev	"p-value of inclusion"
NULL model	-	-	1328	1805	-
log10(totalvalue)	1	291	1327	1514	<0.001
valueincrease	1	16	1326	1499	<0.001
market	2	16	1324	1483	<0.001
total.lines	1	7	1323	1476	0.010
digita_cli	1	6	1322	1470	0.011
log10(totalvalue):market	2	10	1320	1459	0.006
log10(totalvalue):digita_cli	1	6	1319	1454	0.018

Table 5.10: Three purchases model Analysis of Deviance

Table 5.10 shows the last model Analysis of Deviance. As in all previous cases, all the variables entered are significant. In this case the variable $\log_{10}$(totalvalue) is again the one that reduces the deviance of the model the most, but with less difference with respect to the others than in the model of two purchases.

5.7. Model comparison

One purchase model		Two purhcases model		Three purchases model	
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
marketOEM	4.81	marketOEM	5.46	marketOEM	4.42
marketSI	5.38	marketSI	4.11	marketSI	3.24
digita_cliYES	0.33	digita_cliYES	-3.76	digita_cliYES	-3.94
log10(totalvalue)	2.10	log10(totalvalue)	2.45	log10(totalvalue)	2.45
marketOEM x log10(totalvalue)	-1.70	marketOEM x log10(totalvalue)	-2.02	marketOEM x log10(totalvalue)	-1.55
marketSI x log10(totalvalue)	-1.81	marketSI x log10(totalvalue)	-1.01	marketSI x log10(totalvalue)	-0.66
fp.seasonwinterTRUE	0.25	fp.seasonwinterTRUE	0.27		
		total.lines	0.04	total.lines	0.04
		digita_cliYES x log10(totalvalue)	2.1	digita_cliYES x log10(totalvalue)	2.01
		digita_cliYES x any.refund	-2.73		
		any.refundTRUE	-0.37		
				valueincrease	4.71 e-04
Main effects	3	Main effects	6	Main effects	5
Interactions	1	Interactions	3	Interactions	2

Table 5.11: Variables present at the three models and its coefficients

Table 5.11 shows how the variables and interactions included in each of the three models vary. There are 3 main variables and one interaction that are common to all: *market*, *digita_cli,* log_{10}*(totalvalue)* and the interaction between *market* and log_{10}*(totalvalue).* They seem to be the better predictors in any case.

fp.seasonWINTER stops being significative in the three purchase model. This seems logical as the variable takes into account the moment when the client was "born", so after three purchases this fact does not seem to be relevant anymore.

any.refund variable seems not to be a robust one. It enters the Two purchases model interacting with log_{10}(totalvalue) but it is no longer maintained in the third model. Moreover, in the third model it does not pass the univariate analysis.

Total.lines enters in the two purchases model and is maintained with the same coefficient in the three purchase one. In both models the coefficient has a low p-value, which makes us think that it is quite significative. In order to verify this, we should check whether models with more purchases keep the variable in.

Finally, *valueincrease* variable appears only in the three purchase model. Its p-value is also low. In order to test its stability, we should check whether it remains in more purchases models.

For coefficients that appear in all models they seem to follow a similar behavior. An exception seems to be the variable *digita_cli*. It goes from having a positive value in the first model to being negative in the other two.

Actually, it is a correct behavior. If we look at the last two models we can see the appearance of the interaction *digita_cli* x log_{10}*(totalvalue)* that increases the slope of log_{10}*(totalvalue)* for

digital customers. As shown in Figure 5.14, for low expenditure values the probabilities of being a good customer when being digital are very similar, but as the value of the total purchases increase they also become bigger. Therefore, we can say that the changes in *digita_cli* coefficients are also consistent.

In Figure 5.16 and Figure 5.17 It can be seen the model behavior's evolution from the first one to the third measured as probability of being good customer vs. value of purchases grouped by sector (first graphs) and by actual good/bad status.

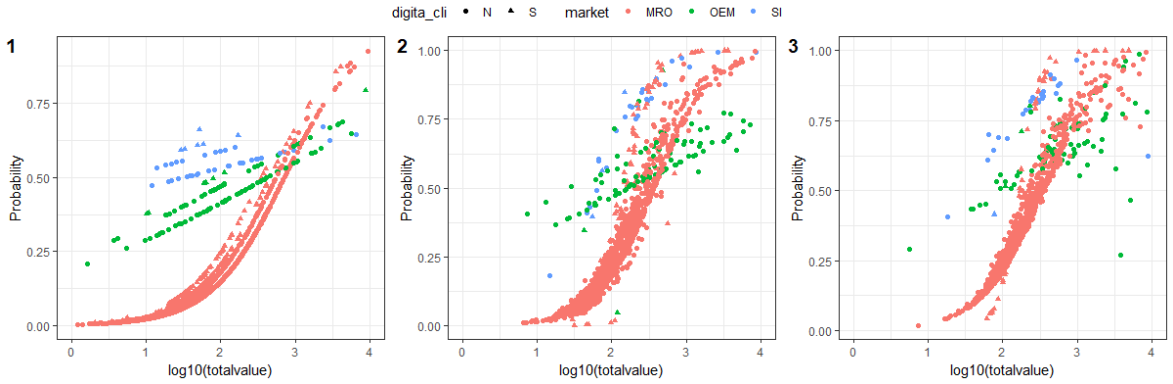


Figure 5.16: Models behavior's evolution by market and digita_cli

Figure 5.17: Model behavior's evolution by good.customer

It can be observed that the $log_{10}(totalvalue)$ is the main covariate in each of the models. It is what most affect the model's probability of being a good customer. It can also be seen that when the number of purchases information increases the effect of market variable decreases. OEM and SI customers' probability variability is higher for the third and second model than for the first one. The market variables relevance decrease is due the inclusion of new variables in those models, as for example *total.lines* or *valueincrease*, that provides more complex information to the models.

It can also be seen that good customers tend to concentrate in the right upper corner while bad customers stay in the left lower corner as models includes more purchases information.

This is a desirable behavior as the aim of the models is to give higher probabilities to the good customers.

In the three purchases model there are four ill placed customers in the right-bottom part of the graphic. The customer with a probability of being good of 0.27 and an expenditure value of 3,792.85 €[4] in its three first purchases is a bad customer, so it has been well classified. His low-probability regardless his very high totalvalue, is due to a *valueincrease* value of -3,742.64€. Even he has a good monetary value, he is a bad customer because he is "dead". He placed its last purchase on the 2019-01-10 and has a recency of 78 days while his maximum time between his 4 purchases was 51 days.

5.8. Model check

Figure 5.18 shows the ROC Curve and the Area Under the Curve for each one of the three models. The ROC Curve describes the relation between Specificity and Sensibility for every possible cutoff[5]. Sensibility measures how good is the model to identify good customers while specificity measures how good it is to detect bad ones. Sensibility and Specifity are calculated using Equation 5.12 and 5.13 respectively.

$$Sensibility = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \text{(Eq. 5.12)}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad \text{(Eq. 5.13)}$$

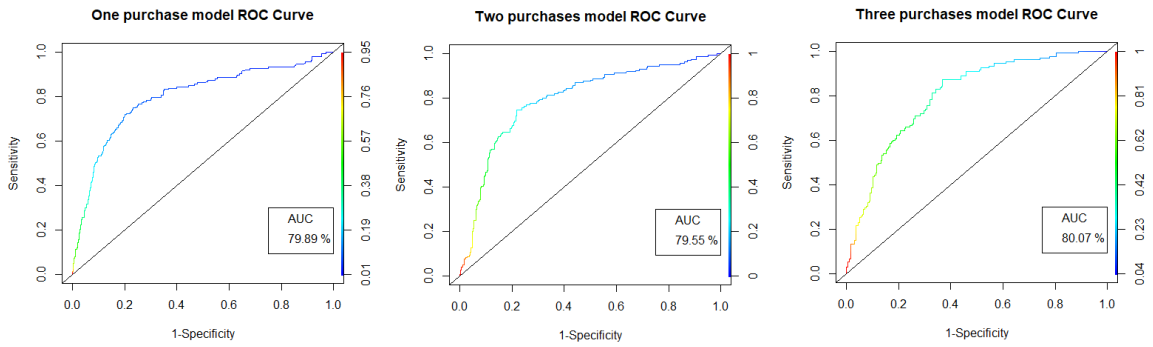


Figure 5.18: ROC Curve and Area Under the Curve for the three models

[4] 3,792.85 € is equivalent to $\log_{10}$(totalvalue) = 3.58

[5] Cutoff: Probability's value chosen by to discriminate whether a customer is good or not. If the probability assigned by the model is greater than the cutoff the customer is classified as good. In the other case, it is classified as bad.

If the model were perfect ROC curve would be close the upper left corner from the very beginning where both sensibility and specificity take 100 % value. The black diagonal line represents how would perform a very simple model that gives 50% of the times True values and the other 50% False values for good customers, as in a coin flip. When the ROC Curve is on top of that line, it can be stated that the model improves that 50/50 simple model.

AUC goes from 79.89% in the first model to 79.55% in the second one and to 80.07 % in the third and last one. Model performance is good but does not seem to improve much when more purchase information is added.

Hosmer and Lemeshow states in their book that models with an AUC going from 70% to 80% are considered as models with an acceptable discrimination and those whose AUC is between 80% and 90% are considered as excellent discrimination ones **[1]**.

The "Area Under the Curve" can be interpreted as the probability that the model classifies well two randomly chosen good and bad customers. Because of that, we can say that model three is the best discriminator one and model One is best than the second model. However, the improvement of the third model is not very significative: it is less than 0.5 percentage points.

Table 5.12 shows the sensibility, specificity, optimal cutoff and AUC for each of the models. Sensibility and specificity depend on the cutoff chosen. We have chosen the cutoff which maximizes both Sensibility and Specificity parameters.

	Sensibility	Specificity	Cutoff	AUC
One purchase's model	75%	76%	0.10	79.89%
Two purchases' model	75%	76%	0.24	79.55%
Three purchases' model	72%	70%	0.40	80.07%

Table 5.12: Sensibility, specificity, optimal cutoff an AUC for the three models.

It can be seen that there is not a clear trend in both Sensibility and Specificity parameters as we go from a model to another. Both parameters maintain the same value when we go from the first to the second model but they decrease when we move to the third model. Sensibility and specificity parameters are very good for the three models: the three correctly classifies more than 70% good and bad customers.

As we have stated before cutoff is a choice made by the model's user. For example, if the company believes it is more important to detect good customers, even at the cost of classifying more bad ones as good, one could lower the cut-off point in order to increase sensitivity while reducing specificity.

Although the percentage of good and bad customers that are detected remains practically the same for three models, the cut-off point does increase significantly as the number of purchases

introduced in the models increases. This indicates that the goodness of fit of the model increases with the inclusion of more information.

Finally, we compute the Hosmer test for the models. The Hosmer test is a so-called goodness-of-fit test, while the AUC is a discrimination test.

This test measures how similar is the number of individuals observed per group versus those expected according to the logistic regression model. To perform this measurement, we calculate the H-value using the Equation 5.14. This statistic follows a Chi-squared distribution with n-2 degrees of freedom, being n the number of groups customers were divided by, often ten. The null hypothesis states that the expected and observed probabilities are equal and, therefore, the predicted and observed values per group are also similar.

$$H = \sum_{q=1}^{n} \left(\frac{(Observed.1 - Expected.1)^2}{Expected.1} + \frac{(Observed.0 - Expected.0)^2}{Expected.0} \right) \qquad \text{(Eq. 5.14)}$$

Table 5.13 shows the result of the Hosmer test for the one purchase model. The expected and observed number of customers are similar for each of the ten groups. Where there is a larger difference between expected and observed values is in the ninth group. This group alone adds 7.38 to the total H parameter. This discrepancy alone makes highly unfeasible to obtain a p-value greater than 0.05 in the Chi-square test with 8 degrees of freedom and not reject the null hypothesis.

Probability	Expexted BAD	Observed BAD	Expected GOOD	Observed GOOD	Contribution to H
[0.00568,0.0404]	144	141	4	7	2.01
[0.0404,0.0421]	143	145	6	4	0.77
[0.0421,0.0468]	138	142	6	2	3.10
[0.0468,0.0525]	139	140	7	6	0.24
[0.0525,0.0608]	139	141	8	6	0.67
[0.0608,0.075]	137	141	10	6	1.61
[0.075,0.101]	133	136	13	10	0.59
[0.101,0.152]	129	125	18	22	0.88
[0.152,0.273]	118	105	29	42	7.28
[0.273,0.947]	83	85	64	62	0.13
p-value	0.03			**total H**	17.27

Table 5.13: One purchase model Hosmer test

As expected, the p-value provided by the test is 0.03, which is lower than the limit of 0.05 proposed by Hosmer-Lemeshow.

The test has several drawbacks. The first is that if the number of expected subjects is small in some interval it has a large impact in the H value as it appears in the denominator of the test value computation. This is the case of our model, where the number of good clients is 11% of the total.

Another big drawback of the test is that the result depends very much on the number of groups into which individuals are divided **[2].** For example, we have divided them into ten deciles but another number could have been chosen and the result would have been different.

For these reasons this type of test is not widely used nowadays, specially when the objective is to obtain good discriminators that perform good enough at classifying customers. Therefore, we are not going to reject the model and we are going to use the test simply to see if the predicted and observed values are similar. Actually, except for the ninth group, all the others match quite well.

The Hosmer test for the two remaining models provides similar results, p-value for the second model is 0.001 and for third is 0.104. The third model obtains a p-value greater than 0.05 which according to Hosmer-Lemeshow test would mean that the model has a significant goodness of fit. Even so, since we are not going to use the test as definitive to reject or accept the models, we have decided to only show the complete test results for the one purchase model.

Conclusions

After building and analyzing the three models, we can conclude that it is possible to predict the future behavior of new customers using only the information obtained in their first transactions with the company. This predictive capacity is very interesting for any company as it helps to focus its sales efforts into future good customers.

Even so, the ability to discriminate between good and bad customers does not increase significantly when going from the one purchase model to the two or three purchase ones. The increase in the AUC or sensitivity and specificity is almost null from the first to the third model. On the other hand, the calibration of the models obtained through the Hosmer test does have a notable increase when increasing the number of purchases. It seems that the model fits better but does not improve its discrimination capabilities.

The main goal for the company is to know whether a customer will be good or bad in the future, i.e. the discriminating power of the models. Because of this, our recommendation is that the one purchase model is quite reliable in valuating new customers. One can give quite good assessment of customer value with its first purchase. Subsequent models can be used to complements the first one and track whether customer behavior follows the prediction given in the first regression model.

Apart from the multivariate models, we also find the univariate analyses obtained in the exploratory analysis very interesting. These models serve to confirm some hypotheses that may seem logical beforehand, such as that a customer who spends more or buys more products has higher probability of being good one or, conversely, that a customer who makes a refund is less likely to be eventually good.

In addition, the univariate analysis can also be used to obtain conclusions that may not be as obvious beforehand. For example, that a customer not located in the same province that the companies distributor is more likely to be a good one, or that non-digital or winter-born customers are more likely to be bad customers. This information can focus the company to further investigate in these types of customer.

There is one aspect of the models that we believe could be improved. It is the improvement of the predictive power of the models as the customer number of purchases increases. We believe we are missing some information provided by the company. This additional information would have allowed us to improve the second and third model.

Regarding the project objectives, we can say that they have been fulfilled.

1. The process of defining dependent and independent variables has been carried out in a proper way and validated by the company.

2. The three purchasing models have been built. They are statistically significative and are capable of classifying new customer with an acceptable error level using the information obtained from their first purchases.
3. We have studied which are the variables that are related to whether a customer is good or bat and their importance and significance.
4. We have been able to compare the variables used in each of the models and observe which ones are present in the three and which ones appear or are removed from model to model.
5. It has been concluded that the addition of new purchases to the models is not as relevant as initially thought. The predictive power provided by the three purchases model is very similar to the one-purchase one. This result seems counterintuitive sot it is worth further exploration before giving it for good.

Next steps

There are still some aspects in the project that deserve further investigation. They are:

1. We have used the link function logit which is the one used by the Logistic regression model to map the linear relationship between predictors and response mean. There are other link functions that map other latent behaviors. Other link functions such as *log-log* or *probit* could be tested to see if they generate models that fit the observed data better.
2. When we analyzed the models we found some variables or interactions which had very high variance and lead to not being able to define them as significant. We should study the existence of over or underdispersion in the model. In such case the actual data variance could be larger or smaller than the modelled one, respectively. The presence of this situation increases coefficient variances and reduces the power of the significance test for coefficients. Modeling for over or under dispersion would not change coefficient values but would reduce their standard errors.
3. It would be also interesting to redo the entire project using another type of classification method. One could use Decision trees or K-Nearest Neighbors and carry out a comparison between them and the models obtained in this project in order to see which is the best classification method for the purpose of the project.
4. Another area of research would be to assess whether a new customer classification changes as he is doing more purchases. A customer that does not change status as doing more purchases could be a further indicator of future potential, while one that goes back and forth could be identified as riskier regardless its actual classification.
5. As It has been mentioned in the conclusion, we think that the second and third model can be improved by using new data or improving the data available. It would be needed to identify changes in the dataset, in conjunction with the company, in order to improve its quality and, therefore, improve the models.

Bibliography

Bibliographic References

[1] D**.** W. Hosmer y S. Lemeshow, Applied Logistic Regression, John Wiley & Sons, 2000, p. 11

[2] P. Allison, «Statistical Horizons,» 2013. [En línea]. Available: https://statisticalhorizons.com/hosmer-lemeshow

Additional bibliography

G. James, D. Witten, T. Hastie y R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.

A. J. Dobson y A. G. Barbett, An Introduction to Generalized Linear Models, Chapman & Hall/CRC, 2008

RStudio, "rstudio," 2018. [Online]. Available: https://www.rstudio.com/resources/cheatsheets/.