NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

# CUSTOMER LIFETIME VALUE IN INSURANCE

Jorge Eduardo Carvalho Abreu

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

# CUSTOMER LIFETIME VALUE IN INSURANCE

by

Jorge Abreu

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

**Advisor:** Rui Gonçalves

**Co Advisor***:* Mauro Castelli

November 2018

# ACKNOWLEDGEMENTS

To NOVA Information Management School (IMS), for the excellence in teaching and the range of opportunities, it provides to its students, allowing them to pursue successful careers.

To professor Leonardo Vanneschi, for creating such an interesting master program, putting together a great teaching staff, and for being one of the best professors I had the pleasure to learn from.

To my supervisors, professors Rui Gonçalves and Mauro Castelli, who have guided me in this challenge with their vast knowledge and dedication.

To professor André Melo who was one of the reasons why I decided to further study the field of Data Science, leading to my enrollment in the Advanced Analytics Master, and whose help was precious at the beginning of this project.

To my work colleagues and team who have taught me so much with their experience and made me understand that I still have so much more learn as a person and as a professional.

To my family who has supported my decision of taking 2 more years of studying instead of starting my professional career back in 2016.

A special thanks to my partner and best friend Inês Tavares, who was always available to help me with my struggles and managed to make stressful times easier to deal with.

# Abstract

Throughout the years, companies from several business sectors have strived to strengthen their client portfolio by acquiring and retaining the most profitable. For this to happen, current and potential clients must be clearly classified based on their past and future interactions with a company throughout the lifetime of their relationship. This report presents how the previous scenario was implemented using Customer Lifetime Value (CLV) in one of the biggest bancassurance companies in Portugal, during a 9-month internship.

Before delving into the detailed set of this project phases, the concept of CLV was reviewed, as well as the characteristics which define its several approaches, followed by the alignment of the chosen approach to the company reality. This CLV model was limited to a 12-month future horizon, covered 7 company dimensions (one global, plus 1 per lines of business) and took into consideration as main future client interactions churn, cross-sell, upsell and risk of claiming. These previous components were modeled with the help of *SAS Enterprise Miner* or estimated using *SAS Enterprise Guide* and analyzing historical events. Besides a purely monetary CLV, it was also generated an ordinal output using a set of business rules and a ranking data discretization method. Finally, a back-test validation procedure was executed to evaluate the reliability of both types of outputs in each of the considered dimensions and its results were analyzed.



# Keywords

Customer Lifetime Value, CLV, Customer Current Value, Upsell, Cross-sell, Churn, Risk, Insurance, Bancassurance.

# Resumo

Ao longo dos anos, empresas de diversos setores têm-se esforçado para fortalecer o seu portfolio de clientes, adquirindo e retendo os mais lucrativos. Para que isto acontecer, os clientes atuais e potenciais têm de ser devidamente categorizados com base nas suas interações passadas e futuras com uma determinada empresa, ao longo do ciclo de vida da sua relação com a mesma. Este relatório vez por sua vez apresentar como o cenário anterior foi implementado durante um estágio de 9 meses numa das maiores empresas de bancassurance em Portugal, recorrendo ao Customer Lifetime Value (CLV).

Antes de aprofundar o conjunto de fases deste projeto, foi feita uma revisão do conceito de CLV, assim como das principais características que definem as diversas abordagens, seguido do alinhamento da abordagem escolhida com a realidade da companhia. Este modelo foi limitado a um horizonte futuro de 12 meses, compreendeu 7 dimensões (uma global e uma por cada linha de negócio) e integrou como principais interações futuras do cliente o *churn, cross-sell*, *upsell* e risco de sinistralidade. Estes componentes foram modelados com a ajuda da ferramenta *SAS Enterprise Miner,* ou estimados utilizando o *SAS Enterprise Guide* para analisar eventos passados. Além de um CLV puramente monetário, também foi criado um output ordinal recorrendo a um conjunto de regras de negócio e um método de *ranking data discretization.* No fim, foi executado um procedimento de validação *back-test* com o intuito de avaliar a credibilidade dos dois tipos de outputs ao longo das várias dimensões e foi feita uma análise dos resultados finais.



# Palavras-chave

Customer Lifetime Value, CLV, Valor actual do cliente, Upsell, Cross-sell, Churn, Risco, Seguros, Bancassurance.

# Index

# List of tables

# List of figures

# List of abbreviations and acronyms

**CCV** – Customer Current Value

**CLV** – Customer Lifetime Value

**LoB** – Line of Business

**MAE** – Mean Absolute Value

**RAE** – Relative Absolute Value

**RP** – Risk Premium

# 1. Introduction

The core business of insurance companies is to enable individuals and firms to protect themselves against infrequent but extreme losses at a cost which is small compared to the feared loss (Rodne, 2009). Insurance by its nature is an intangible good, involving payment in advance for an unknown quality of future service delivery and covers a wide range of risks such as natural disasters, property risks (fire, burglary, etc.), health, motor, among others.

Traditionally, agents and brokers have been the sole distributors of insurance policies, however, developments in consumer behavior, technology, deregulations, etc. lead to the development of different ways to sell insurance, always keeping in mind the customer's preferred combination of product, pricing, and service (Chatley, 2014). One of those channels is bancassurance, known by having a bank either acting directly for an insurer or providing space for an insurer's representative in its retail outlets (Rodne, 2009). Within this distribution channel, insurance policies can be sold to bank clients in two different manners: *i)* active sale – the client acquires one or more policies within the bank channel; *ii)* associated sale – the client acquires one or more policies resulting from the subscription of a bank product (e.g., Home loan).

Throughout the years, companies from several business sectors have strived to strengthen their client portfolio by acquiring and retaining the most profitable, and Insurance companies were no exception. In order for this to happen, current and potential clients had to be clearly classified in a way that not only specified how much would a client value in the near future (e.g., next year), but also in the long run, until its relationship with the company lasted. To answer this matter, Customer Lifetime Value (CLV) has been adopted by several companies to measure clients according to their potential monetary value over various periods of time. One of the most complete definitions of CLV was presented by (Hoekstra & Huizingh, 1999) which describe it as "(…) *the total value of direct contributions and indirect contributions to overhead and profit of an individual customer during the entire customer lifecycle that is from the start of the relationship until its projected ending*". In (Statsbot, 2018), a simpler definition is presented, declaring this concept as a prediction of the amount of money that a customer will spend with a business in its lifetime, or at least, in a portion of it. Several other versions of this concept definition could be reviewed in (Abdolvand, Albadvi, & Koosha, 2014). In contrast to traditional customer classification methods (e.g., credit scoring), CLV produces a monetary value for each individual customer directly related to its expected future profitability. This simple, yet powerful measure can be used not just to determine which clients have the most potential, but also to decide how much in marketing expenditures is justified for each one (D One, 2013). Within the insurance sector, CLV has multiple applications, of which some examples are: *i*) Agent compensation; *ii*) Affinity programs; *iii*) Campaign's lead prioritization; *iv*) Expense allocations; v) New product offering/design (Towers Watson, 2015).

Considering the business potential of this measure, the main objective of this project was to estimate CLV over a 12-month horizon for all individual customers, belonging to one of the top bancassurance companies operating in Portugal, during a 9-month academic internship. Following the main objective, 3 other secondary objectives were defined: 1) Distinguish CLV estimations

between two dimensions- Global (company-wide) and by each Line of Business (LoB); 2) Consider upsell, cross-sell and churn as possible customer interactions to estimate CLV; 3) Assign the final output to the main beneficiary of each policy (policy-holder) based on the analyzed behaviors from all insured people belonging to all policies he/she holds. To accomplish these objectives, the following 5-step plan was outlined: 1) Understand the business context and align CLV to it; 2) Build the core datasets of policies and clients; 3) Estimate all necessary CLV components, based on the output of the previous step; 4) Integrate existing data mining model outputs to calculate CLV at the Global and Line of Business levels; 5) Verify and analyze the obtained results through a back-test validation process.

# 2. Literature Review – Customer Lifetime Value

At the time this document was written, it was known that several other customer classification metrics existed (e.g., Share of Wallet, Recency-Frequency-Monetary, etc.), however, because this project was done in a business environment, the development of Customer Lifetime Value was already planned at the moment it was presented to the author and no time was spent researching alternative metrics. With this in mind, no review was made regarding other metrics other than CLV.

At the time this report was written there were more than 25 different approaches, each one with the objective of predicting CLV according to distinct business environments. In annex **A 1** there is a table created by Tuomas Harju, where several methods of calculating CLV are summarized, according to specific contexts (Harju, 2015). The methods presented were divided by 4 main categories: *type of contract*, lo*st-for-good vs always-a-share*, *deterministic vs stochastic* and *aggregation level*. These categories are going to be further discussed in the following sections.

## 2.1 Type of contract

Depending on the scenario, a customer might need to sign a contract in order to acquire a given product/service. This leads to two possible contexts regarding a customer's relationship with a company: *contractual,* or *non-contractual.* A contractual setting could be defined as one where the transaction opportunities are continuous and the moment at which customers become inactive is observed (Fader, Hardie, & Ka, 2008). On the other hand, a non-contractual setting, besides the fact of not needing any type of contract to formalize a purchase, it is also characterized by the necessity to indirectly deduce the end of a customer relationship from a long-term inactive behavior (Donkers, Verhoef, & Jong, 2007).

The main difference between these two contractual scenarios is essentially that, in the contractual setting, there is more awareness over the duration of customer relationship duration, while in the non-contractual setting companies cannot determine how long a customer will remain active (Borle, Singh, & Jain, 2008). CLV-wise, companies in the first setting should focus into accurately predict customer retention, while in the latter the focus should rely on an accurate prediction of the customer activity and contribution margin (Venkatesan & Kumar, 2004).

Given this project was developed for an insurance company, the contractual context was the one aligned with the observed reality, since each time a customer desires to acquire any insurance policy he/she has to sign a contract with the applied terms and conditions.

## 2.2 Lost-for-good vs always-a-share

In this category, customers are classified as *lost-for-good*, or *always-a-share*. These two states are differentiated by how they handle customer retention (Gupta, Hanssens, Hardie, & Kahn, 2006).

In the lost-for-good context, a customer is assumed to always make purchases until it stops permanently, leaving the company for good and cannot be reacquired (Rust, Lemon, & Zeithaml,

2004). In this case, the probability of having another purchase is given by a value between 0 and 1, decreasing towards 0 has the duration of the customer relationship with the company increases. Additionally, lost-for-good approaches, do not consider other types of customer dynamics, other than "active" or "inactive" (Romero, van der Lans, & Wierenga, 2013). Business-to-Business and Financial services companies are examples where lost-for-good approaches are commonly adopted. On the other hand, in the always-a-share context, customers are assumed to distribute their spending across several businesses of the same sector (Rust, Lemon, & Zeithaml, 2004). A good example of an always-a-share sector is retail. Customer status can remain "active" despite a period of no purchases, therefore there is never a permanent abandonment from the company. In this scenario, instead of having a probability of retention, for each customer is predicted the possibility of repeating a purchase (Venkatesan & Kumar, 2004).

Insurance-wise, taking into consideration the characteristics of the two previous contexts, the one which appears to be more related with the observed reality of this sector is the lost-for-good, especially due to the similarities on the "active" and "inactive" classification that is given to customer dynamics.

## 2.3 Deterministic vs Stochastic

Deterministic models are ones which state variables are uniquely determined by parameters in the model and by sets of previous states of these variables. Therefore, deterministic models perform the same way for a given set of parameters and initial conditions and their solution is unique. Contrarywise, stochastic models are described by random variables or distributions rather than by a single value. Correspondingly, state variables are also described by probability distributions. In this sense, a stochastic model yields multiple equally likely solutions, which allow the modeler to evaluate the inherent uncertainty of the natural system being modelled. (Renard, Alcolea, & Ginsbourger, 2013).

Early CLV models tended to feature only deterministic inputs, i.e. the inputs regarding customer behavior were entered directly into the formulas for calculating CLV (Holm, Kumar, & Rohde, 2012). Simplicity was one of the main characteristics of these early versions, however, the introduced complexity of stochastic CLV models allowed them to grasp customers behaviors which could not be perceived by the original approaches, such as referral value (i.e., attracting new customers), influence value (i.e., the ability to influence the behavior of others) and knowledge value (i.e., how valuable is a customer feedback) (Kumar, et al., 2010).

Given this was one of the first times CLV was being implemented in this company and because tasks were limited by tight deadlines, deterministic approaches, characterized by their simplicity factor, were the ones chosen to be adopted.

## 2.4 Aggregation level

The final category used to characterize CLV approaches was related with the level of granularity/aggregation to be adopted. Two distinct levels exist: *aggregated*, or *individual*. The main differences between these two approaches are essentially based on simplicity and accuracy.

Unlike the previous categories, there was no obvious choice to be made right from the beginning regarding which of the two granularity options was to be chosen. Because of this, in the following sections is presented the performed research that supported the customer-level detail at which CLV was applied.

### 2.4.1 Aggregated approaches

The main assumption made in calculating an aggregated CLV is that value derives from a specific group of clients with similar characteristics, which could be related with demographics, purchasing behaviors, etc. (Alexandre, 2009). Usually, these groups of clients are created based on clustering algorithms which define several segments that make sense to the business sector they are part of. One of the first CLV aggregated approaches to be suggested was by Blattberg and Deighton in 1996 (Blattberg & Deighton, 1991) and then reinforced by Berger and Nasr in 1998 (Berger & Nasr, 1998) by formulating the approach in the following manner:

$$CLV_s = \sum_{t=0}^{T} \left[ \frac{(GC - M)}{(1+d)^t} \, r^t \right] - A \qquad (1)$$

Where,

$t$ – Period of Time

$T$ – Defined CLV time horizon

$S$ – Total number of distinct client groups

$s$ – Group $s$ of clients, with $s = 1,2,\dots,S$

$GC$ – Expected yearly average gross contribution margin of $s$

$M$ – Average costs of $s$

$d$ – Discount rate in each period of time $t$

$r$ – Retention rate of $s$

$A$ – Average cost of acquisition of clients in $s$

The formula above could be perceived as being rather complex, so other much simpler calculations of CLV were created by companies (Kiss Metrics, Sweet tooth, RJ metrics, Custora, among others)

who sell their services and tools to calculate this and other customer-centric metrics (Sweet tooth, 2015). Each of those companies has their "magic formulas", according to the business they apply it to. While some of those formulas are kept secret to protect these companies, some are publicly known. Assuming yearly periods, for a given set of customers, these formulas are defined as follows (Sweet tooth, 2015; Kiss metrics, s.d.):

$$CLV = AOC \times f \times t \quad \textbf{(2)}$$

Where,

$AOC$ – Average yearly order value

$f$ – Average yearly frequency

$t$ – Average customer lifespan (in years)

Or

$$CLV = AOC \times f \times t \times p \quad \textbf{(3)}$$

Where,

$p$ – Average yearly profit margin

Equations ***( 2 )*** and ***( 3 )*** make CLV simple to calculate, however in terms of accuracy they tend to perform worse in comparison with equation ***( 1 )***, since they do not account for some relevant components (e.g., retention) and assume customer behavior as being constant over time. To try to fight off accuracy problems, some of these companies use several simple CLV formulas, where each one generates its own output and, in the end, the final CLV is considered to be the average of all outputs. Overall, albeit the earliest or simplest approaches of CLV lean towards measuring parameters on an aggregate level, the tendency of later models was to analyze each customer individually without inferring all its interactions with the company just because he/she is part of a group characterized by similar behaviors (Harju, 2015; Holm, Kumar, & Rohde, 2012).

### 2.4.2 Individual approaches

Within the CLV individual approach, each considered parameter/component tends to be aligned with each specific customer based on its unique characteristics and past behaviors. Some group-level formulas have their parameters redefined to the individual perspective without changing its structure. An example of this is the adjustment of equation ***( 1 )*** where the parameters are retrieved for each client $i$, instead of being per client group $s$.

Among the revised formulas to calculate CLV at the individual level, one of the simplest was proposed by Jain and Singh in 2002 (Jain & Singh, 2002) and was defined as follows:

$$CLV_i = \sum_{t=1}^{T} \frac{(R_t - C_t)}{(1+d)^{t-0,5}}$$

**( 4 )**

Where,

$i$ – Client

$t$ –Period of analysis

$T$ –Total number of periods

$R_t$ – Customer revenue in period $t$

$C_t$ –Total costs of generating $R_t$ in period $t$

One of the factors that makes this formula so simple to apply is that almost no indirect costs are considered (e.g., Marketing costs). Typically, this model can indirectly support firm actions such as customer acquisition, retention, cross-sell, among others (Reinartz & Kumar, 2003). The simplicity of this formula was very appealing, and even though it did not have specific parameters to represent interactions, such as cross-sell, those could be included within the revenues of future periods. However, one important element that was not covered was the customer retention for each period of analysis. Disregarding this component is the equivalent of not accounting for an impactful event in the customer's lifetime, which makes this model incomplete in that sense.

In 2004, Gupta, Lehman and Stuart (Gupta, Lehmann, & Stuart, 2004) proposed an upgrade to the previous formula.

$$CLV_i = \sum_{t=1}^{T} \left[ \frac{(R_t - C_t) \times P(Active)_{i,t}}{(1+d)^t} \right] - AC$$

**( 5 )**

Where,

$AC$ – Acquisition costs

$P(Active)_{i,t}$ – Probability of client $i$ being active at time $t$

Regarding the formula above, the major upgrade in comparison to equation ***( 4 )*** was the integration of a retention component, making it more suitable to the scope of which it was supposed to be applied. Although the previous formula seemed to fulfil the necessary requisites to be adopted for

this project, additional arrangements were still necessary to be made in equation's ***( 4 )*** approach, enabling it to be aligned not only to the insurance sector but also to the company's reality. Nevertheless, it was clear individual approaches were the best option to choose regarding the aggregation level, due to the fact they were able to better perceive each client's past and potential behaviors.

## 2.5 Project alignment

The developed research enabled several CLV approaches to be found and analyzed. However, there wasn't one which clearly fulfilled all project objectives, namely components related to future customer interactions with the business, such as cross-sell and upsell. As mentioned in the previous section, these components could be indirectly part of future revenues (e.g., in equation **( *5* )**, they would be part of $R_t$ when $> 0$, being $t = 0$ the current date). By making this decision, equation **( *5* )** would suit these requirements. Nevertheless, other requirements still had to be fulfilled, namely building a CLV based on the customer's behavior throughout the whole company and regarding each Line of Business.

Based on the rationale presented by Monika Seyerle (Seyerle, 2001), where an implementation of CLV in the insurance business is suggested, given each period of analysis $t$ (in years), CLV could be divided into two parts:

- Present Value (*PV*), when $t = 0$
- Future Value (*FV*), when $t > 0$

Where,

$$PV = Premiums - Claims - ABC \quad \textbf{( 6 )}$$

With,

$Premiums$ – Amount paid by the customer to acquire each policy (product)

$Claims$ – Amount paid by the company to cover a customer claim

$ABC$ – Activity based costs

And

$$FV = \frac{Future\ Premiums - Cancellation + Additional\ Revenues - Risk - (activity\ based\ costs)}{(1+d)^1}$$

(Assuming t = 1)

**( 7 )**

Where,

$Future\ Premiums$ – Premiums to be paid by the customer in period $t$

$Cancelation$ – Value of client cancelation

$Additional\ Revenues$ – Revenues from probable customer interactions (e.g., Cross-sell)

$Risk$ – Risk value assigned to a given customer in period $t$

Considering equations **( *6* )** and **( *7* )**, activity based costs (ABC) are any costs which derive from business-as-usual processes responsible by managing customers and their products. Usually each line of business has its own ABC and distribute them equally across the customer portfolio. However, because it was not possible to grasp all ABC corresponding to each LoB only commissions were considered. This way, all LoB's would be balanced by considering the same type of costs and revenues. The variable of "Cancelation" represents the value that would be lost in case the analyzed customer cancelled /churned a policy. However, the author decided not to consider this variable, since according to Seyerl this value would be calculated as:

$$Cancellation = Future\ Premium \times churn\ probability$$

**( 8 )**

In equation **( *5* )**, the implemented methodology already considered the probability of a given client to be active in period $t$, so adopting this variable would replicate the probability of churn effect. Baring this in mind, the cancelation variable was considered to be the product between future values and the probability of a client maintaining its activity in period $t$, which was already considered in the original equation.

$$Cancellation_t = FV_t \times P(Active)_t$$

**( 9 )**

Furthermore, the variable of *Additional revenues* was thought to include 2 components: cross-sell and up-sell. Following Seyerle's rationale, these components would be calculated in the following manner:

$$Cross\text{-}sell = Prob.Cross\text{-}sell \times Value\ of\ Cross\text{-}sell$$

**( 10 )**

And

$$Upsell = Prob.Upsell \times Value\ of\ Upsell$$

**( 11 )**

In equation **( 10 )** the element of "cross-sell risk" was also added, since insurance-wise, whenever a client acquires a new product there is always a new source of risk being created. Therefore, cross-sell was given by:

$$Cross\text{-}sell = Prob.Cross\text{-}sell \times (Value\ of\ Cross\text{-}sell - \ Cross\text{-}sell\ Risk)$$

**( 12 )**

Regarding the Risk component, this was given by what in insurance terms is called Risk Premium (RP). This term is defined as being the minimum amount of money necessary to be paid to cover the risk that is being taken by the company on a given policy (Anderson & Brown, 2005). Company-wise, for each analyzed period *t* (in years), the risk premium of each policy is given by:

$$RP = Claim\ Frequency_{t-n} \ \times Average\ Cost\ of\ Claim_{t-n,} \quad n=1,2,\ldots N \in \mathbb{N}$$

**( 13 )**

Where,

$$Average\ Cost\ of\ Claim\ _{t-n} = \frac{Cost\ of\ Claims_{t-n}}{No.\ Claims\ _{t-n}}$$

**( 14 )**

$$Claim\ Frequency\ _{t-n} = \frac{No.\ Claims\ _{t-n}}{Exposure_{t-n}}$$

**( 15 )**

Concerning the interest rate variable ($d$), because CLV was only being calculated over a 12-month period, it was decided not to include it, since over a year the value of money does not tend to suffer big fluctuations. The previous decision was supported by Portugal's last 5 years historical interest rate data from the European Central Bank (ECB), presented in the chart below.

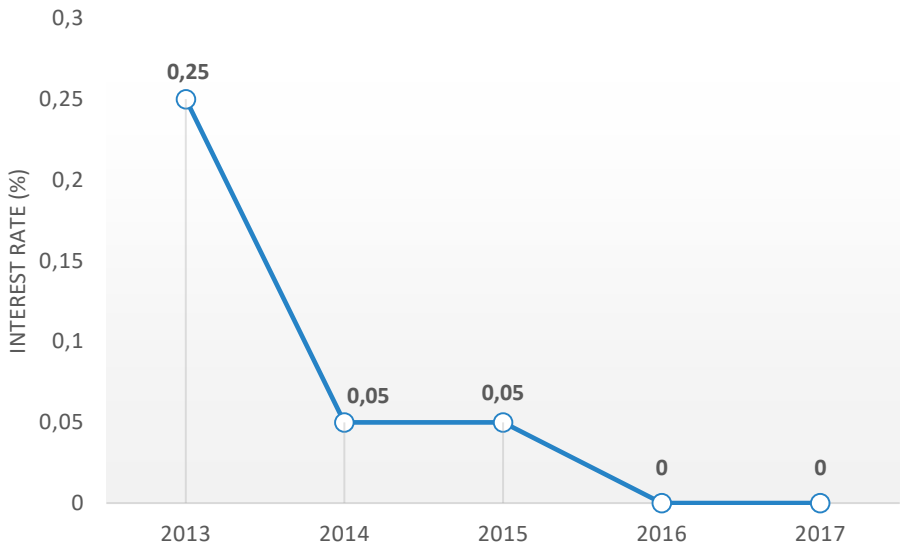



*Chart 1 – Portugal's interest rate over the last 5 years according to the European Central Bank*[1]

Having almost every component examined, it was defined that for each customer $i$, line of business $l$ and analyzed period $t$, CLV was given by:

$$CLV_{\,i,l} = Curent\ Customer\ Value_{\,i,l\ t_0} + \ (Revenues_{\,i,l\ \ t_1} - Costs_{\,i,l\ t_1}) * (1 - \ Churn_{i,l\,t_1})$$

**( 16 )**

Where,

$t_0$ – present period of analysis

$Churn_{i,l\,t_1}$ – probability of client $i$ to leave LoB $l$ in the next 12 months

And

$$Curent\ Customer\ Value_{\,i,l\ t_0} = \sum_{n=1}^{N} Premiums\ Paid_{\,t_0-n} - \ Claims\ Charged_{\,t_0-n} - \ Comissions\ Charged_{\,t_0-n}$$

**( 17 )**

$$Costs_{\,i,l} = \begin{cases} Risk_{\,i,l} + Comissions_{\,i,l} \ \ if \ \ Y(i,l) = 1 \\ 0 \ \ \ if\ Y(i,l) = 0 \end{cases}$$

**( 18 )**

[1] Retrieved from: https://tradingeconomics.com/portugal/interest-rate in the 6th of August 2018.

And

$$Revenues_{\ i,l,\ t=1} = \begin{cases} Upsell\ Prob._{i,l} * (\ Upsell\ Value_{i,l} + Expected\ Premium_{i,l}\ ) & if\ Y(i,l) = 1 \\ Cross\text{-}Sell\ Prob._{i,l} * (\ Cross\text{-}Sell\ Value_{i,l} - Cross\text{-}Sell\ Risk_{i,l}\ ) & if\ Y(i,l) = 0 \end{cases}$$

**( 19 )**

With,

$$n = 1, 2, \dots, N \in \mathbb{N}$$

*Y (i, l) = Function indicating client $i$ is present (**1**), or not (**0**), in Line of Business $l$*

In the end, the global CLV of each client was simply given by summing its respective CLV's regarding each line of business.

$$CLV_{\ i} = \sum CLV_{i,l}$$

**( 20 )**

With all formulas properly defined and aligned to the company's reality, it was possible to understand which components were going to take part on CLV, how would they interact with each other and in which periods of time would they be relevant.

# 3. Methodology

This section had as main focus further detailing all tasks required to generate the final output. The set of topics composing this project's methodology could be divided into two different parts:

1. A project planning part, where the emphasis was directed towards defining project requirements and aligning CLV to them, determining project phases and their timeframes, among other planning elements.
2. A more practical part focused on explaining the series of steps executed to build all the components defined in the previous group. This part was mostly developed using *SAS Guide 7.1*, but also comprehended some analysis procedures, produced with the help of *SAS Enterprise Miner 14.1* and *Excel 2013*.

## 3.1 Project requirements

To build the CLV metric two main data sources were used: Client and Policy data marts. These tables provided daily information regarding the status of all clients and policies (active or inactive). In order to generate the final set of clients and policies a defined set of filters was applied aligned with a predefined set of project requirements. The table below presents a high-level view of the adopted criteria.

| CLIENT | POLICY |
|---|---|
| ▪ *non-missing* Fiscal Number (NIF)<br>▪ Individual (non-corporate)<br>▪ Age ≥ 18 and *non-missing*<br>▪ Not employee | ▪ Individual policy (non-corporate)<br>▪ Error-free (***e.g.,*** entry date equal or older than departure date)<br>▪ Non-Financial policy |

*Table 1 – Adopted Client and Policy-level filters*

Based on the table above, there are some notes worth taking into consideration, those being:

1. "Non-corporate" filters were necessary at both the client and policy level because it was possible to observe individual customers with corporate and individual policies, simultaneously. In these cases, the client was not rejected, but only its individual policies were considered.

2. Financial policies were filtered out because of their distinct behavior in comparison to the remaining products, which the adopted formula of CLV could not handle. (e.g., in specific scenarios, if a client decided to churn a financial policy, that would be positive for the company because penalties would be applied, and the value generated by that policy could be higher than the one generated in case the policy completed its full period).

3. Though only active policies were taken into account to calculate CLV, the "activity" filter was not applied right from the beginning, since it was necessary to analyze historical data, significantly composed by inactive policies, in order to understand past customer behavior needed to produce estimations for several CLV components.

In the end, Customer Lifetime Value was assigned to **713 125 active clients**, among **1 433 388** of their respective **active policies**.

With the purpose of better understand the previous client population several charts were created.

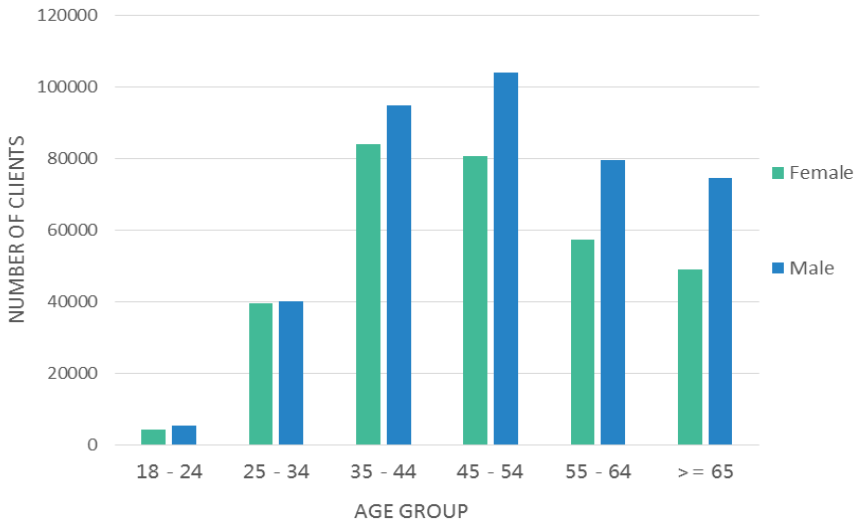



*Chart 2 - Number of Clients by Gender and Age Group*

Based on the chart above regarding age group and gender, it is possible to understand:

- Men represent **56%** of the client universe;
- Clients with **age of 45 +** represent **62%** of the client base;

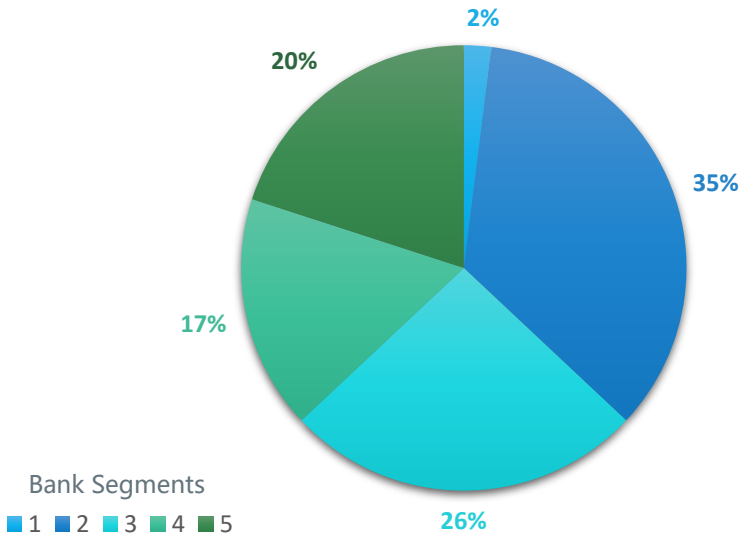



*Chart 3 - Percentage of clients per Bank segment*

Based on the chart above regarding bank segment distribution, it is possible to notice that segments 2 and 3 represent **61%** of the whole client base. It is relevant to mention that not all clients had a bank segment assigned, due to some business rules. Those clients were grouped up separately.

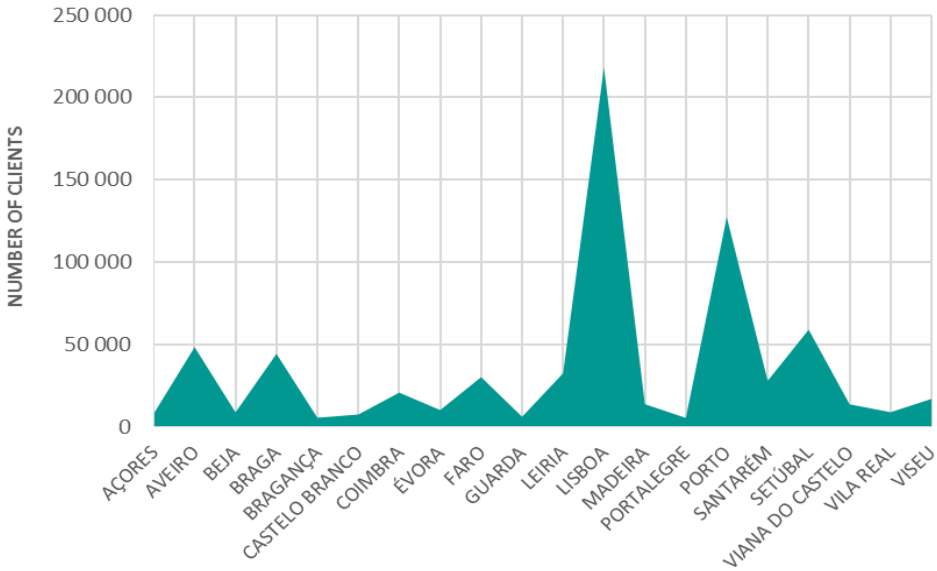



*Chart 4 - Number of Clients per District/Region of Portugal*

Based on the chart above regarding client distribution across Portuguese districts/regions, it is possible to conclude:

- **Lisboa** and **Porto** have the bigger concentration of clients with 3**0%** and **18%** of the whole client base, respectively;
- Aside from the previous districts, **Setúbal**, **Aveiro** and **Braga** are the regions with bigger concentration of clients with **8%, 7%** and **6%**, respectively;

## 3.2 Customer Lifetime Value project alignment

Before applying any version of CLV, it was necessary to align this concept with the initially defined project requirements. Baring this in mind, two sub-topics were considered relevant to be further detailed: components and lines of business. The first one covers the set of identified elements from which CLV would be calculated from, whereas the second, comprehends which business areas were taken into consideration and from which the computation of the defined components was based on.

### 3.2.1 Components

To produce the metric of Customer Lifetime Value there was the necessity to integrate a set of unique components, in order to align this concept with the insurance business context and the project requirements. These could either represent past or future customer interactions which together defined how valuable each given customer could be in the next 12-month period. The following table presents the set of components considered to build CLV.

| PRESENT VALUE COMPONENTS | FUTURE VALUE COMPONENTS |
|---|---|
| ▪ Premiums Paid<br>▪ Claims charged<br>▪ Commissions charged | ▪ Upsell<br>▪ Cross-sell<br>▪ Churn<br>▪ Risk<br>▪ Expected Future Premiums |

*Table 2 – Customer Lifetime Value adopted Present and Future value components*

Throughout this document, future value components will be further scrutinized with the exception of future premiums since these were simply retrieved from a table column, without requiring any future calculations and were no more than a replication of the total current customer's premiums. Time constraints did not allow a more detailed analysis of how future premiums would vary on a new policy annuity. Because of this, and after speaking with business stakeholders, the best workaround to solve the problem in question was the replication of current premiums, since most of the times the variation of a policy's premium is not that significant. Regarding the 3 present value components (Premiums, Claims and Commissions) from the table above, they were gathered from reliable sources of historical data that went up to 3 years ago. Because these 3 components were used to calculate the Current Customer Value (see equation **( 17 )**) they also ended up restraining the time horizon considered to compute it. Ideally, CCV would consider premiums, claims and commissions associated to each client since the start of its relationship with the company, but

because that wasn't possible due to data limitations, the 3-year interval was found to be the best workaround, given the circumstances.

### 3.2.2 Lines of Business

The set of Lines of Business had to be defined in order to apply the previously presented components. To comply with confidentiality requirements, it could only be mentioned that 6 different Lines of Business were identified and labelled from "A" to "F". The chosen LoB's were aligned with other choices made in previous company projects and could comprehend multiple or individual products. The following chart was produced with the objective of presenting a better perception of how each Line of Business was composed. This chart took into account active policies and clients of each LoB at the date of **15th of February 2018**, which was also the reference date taken into account for most of the presented results in this document.

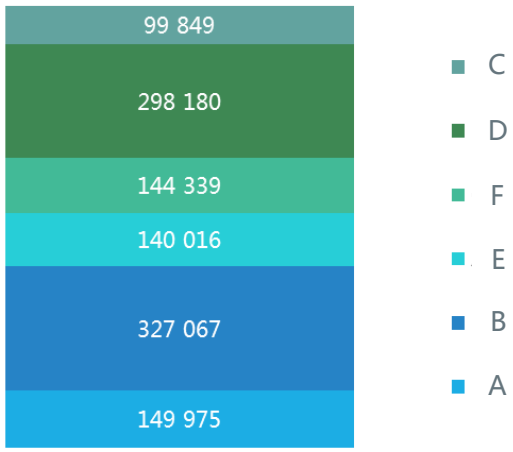



*Chart 5 – Number of clients per Line of Business*

Given the chart above, it is important to mention that, because each client could have multiple policies across different Lines of Business, the sum of clients from each Line of Business **is not** the true amount of considered clients for this project.

After properly defining which components and lines of business were going to be taken into account, it was created a high-level view table, which crossed these two elements. The main idea was to understand, for each line of business, which components had to be incorporated (not all lines of business required all components) and from those, which were already available to be integrated (available models), which ones were simultaneously being developed by other teams, and which had to be estimated based on historical data (simple estimations) because no other alternative was available. Additionally, if a given component corresponding to a certain line of business had a stakeholder assigned to it, the name of that person would also be part of the high-level table, in the slot corresponding to the component he/she was responsible for. However, to avoid revealing confidential information regarding the company's maturity in terms of Data Mining models, the

table below did not contain any compromising information, other than the high-level view of which components had to be estimated by each LoB.

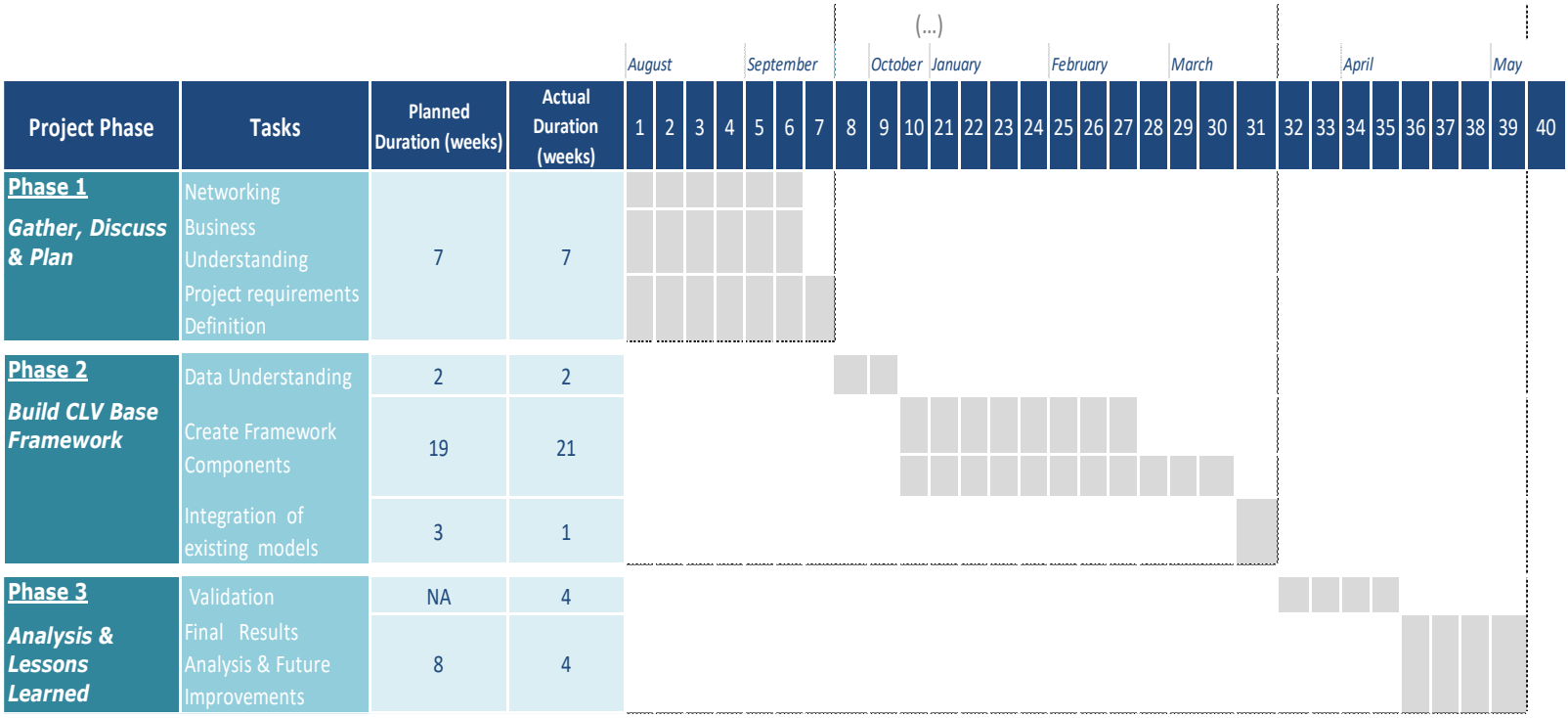
| LINE OF BUSINESS | Upsell Probability | Upsell Value | Cross-sell Probability | Cross-sell Value | Churn | Risk |
|---|---|---|---|---|---|---|
| A | | | | | | |
| B (*Active Sale*) | NA | NA | | | | |
| B (*Associated Sale*) | NA | NA | | | | |
| C | | | | | | |
| D (*Active Sale*) | | | | | | |
| D (*Associated Sale*) | NA | NA | | | | |
| E | NA | NA | | | | |
| F | NA | NA | | | | |

*Table 3 - High-level view of which components to cross by LoB*

There were some lines of business (B and D) which presented different behaviors in their clients depending on the type of sale they belong to (Active or Associated), thus different approaches had to be taken for some components. Besides that, regarding the upsell component, not all lines of business were applicable either because the characteristics of its products did not allow upsell (e.g., associated sale products or the product did not have several packages/options to choose from), or the amount of data representing this type of event was insignificant.

## 3.3 Project Roadmap

Before making the transition to the practical part of the project, it was crucial to properly define a roadmap with all phases which were going to be executed, their time allocation and the sequence of tasks to be followed. The figure below intends to present the adopted roadmap without forgetting to make the distinction between the planned and the actual time frame.




| Project Phase | Tasks | Planned Duration (weeks) | Actual Duration (weeks) |
|---|---|---|---|
| **Phase 1** ***Gather, Discuss & Plan*** | Networking<br>Business Understanding<br>Project requirements Definition | 7 | 7 |
| **Phase 2** ***Build CLV Base Framework*** | Data Understanding | 2 | 2 |
| | Create Framework Components | 19 | 21 |
| | Integration of existing models | 3 | 1 |
| **Phase 3** ***Analysis & Lessons Learned*** | Validation | NA | 4 |
| | Final Results Analysis & Future Improvements | 8 | 4 |

*Figure I - Adopted Project Roadmap*

This was not the originally proposed roadmap, but unforeseen events forced some tasks to be excluded and/or added. One of the phases that was initially planned was a research on how Survival Analysis models could be integrated in the future to enable the computation of a Customer Lifetime Value within a timeframe bigger than 1 year which, due to time constraints was passed to an improvement to be added in future versions of this project.

## 3.4 Output Example

With the idealized CLV framework built, it was also necessary to think how the final output would be presented. In this sense, it was defined that all the information would be aggregated at the policy-holder level since the main tables of this project (client and policy-level data marts) had their information aggregated at that same level. Each policy-holder was then assigned classified in 3 different levels of granularity: Global-level (Company-wise), Line of Business-level and component-level.

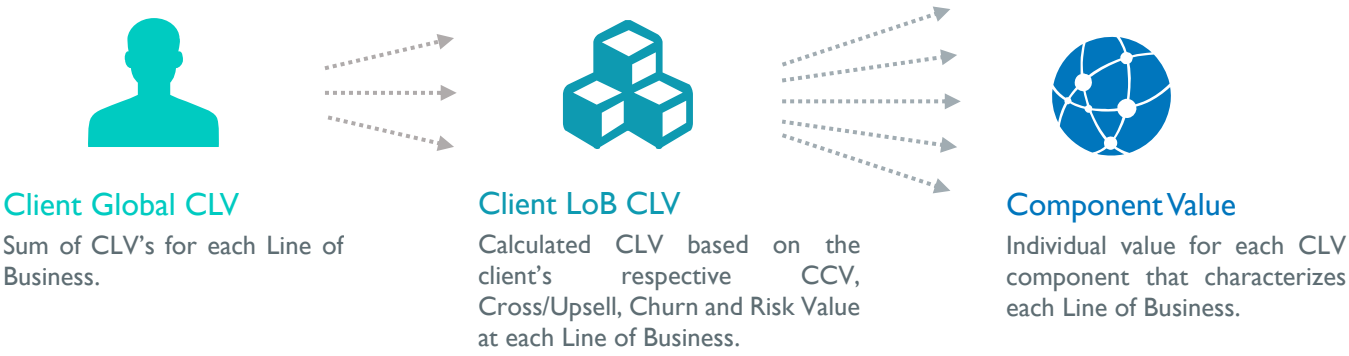



*Figure II – The three granularity levels considered in the final CLV framework output*

After having the required output values to complete the previous 3-levels of granularity, the information of each policy-holder was then presented according to the following example.

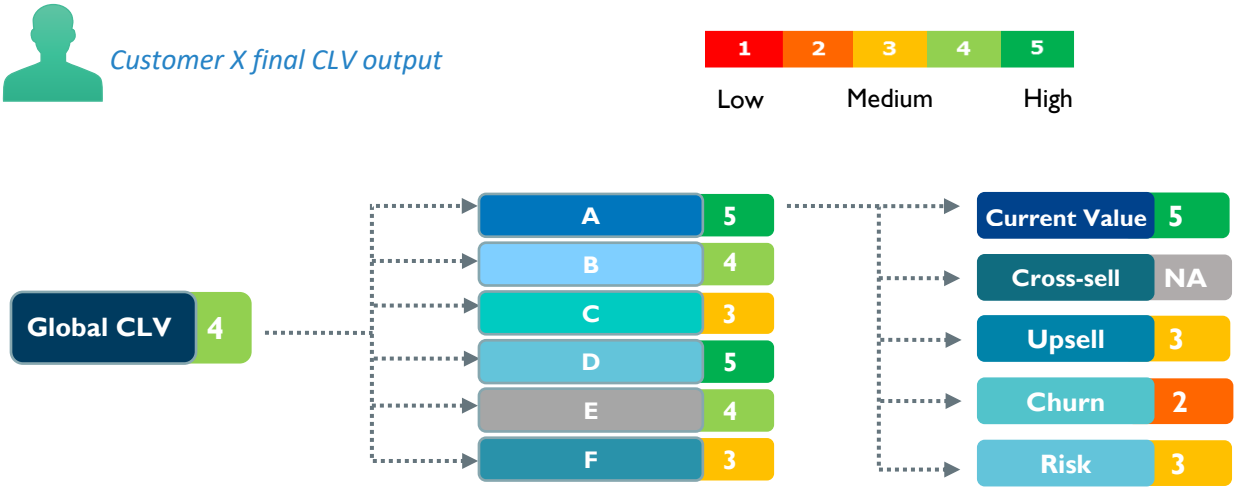



*Figure III – Detailed representation of the final output*

In the example above, it was chosen a Low-to-High labeling system to simplify interpretability.

Without entering in too much detail, in this example it is shown a client that could be characterized globally by having a Medium-high CLV, supported by a High CLV in LoB A and D, and a Medium-High

CLV in B and E. Going to a higher level of detail, regarding LoB A, this client could be characterized by being within the most valuable clients due to its CCV rank. Besides, it is also characterized by having a Medium Risk and Upsell potential associated and Medium-Low churn propensity. Clearly, this is an example of a client that should be retained by the company and could justify an extra effort to do so, given its higher ranks in some LoB's.

## 3.5 Data preparation

As a first step from the more practical group of tasks, the set of required filters, mentioned in the project requirements (section 3.1), were applied to generate the eligible set of policies and clients. At this point there two crucial sets for each perspective of clients and policies: one with historical data which, included active and inactive observations and another with the set of policies to be classified at the level of each component to then be aggregated at the client-level.

After applying the previous set of filters, additional variables were created to aid the next steps of CLV implementation, namely regarding components. Those were essentially possession variables i.e., variables characterizing a policy or a client regarding its past or current presence in each line of business, distribution channel, sale type, etc. In the next sections, the development of the considered CLV components will be further scrutinized.

## 3.6 Risk

The first component to be dealt with was Risk. As previously mentioned in section 2.5, this component could be explained by the term "Risk Premium" which is defined as being the minimum amount of money necessary to be paid to cover the risk that is being taken by the company on a given policy. In LoB's where Risk models were available, the current risk premium (RP) was generated automatically for their respective active clients. On the remaining cases, it was used a simplified set of formulas to calculate this component. Based on the equations presented in the literature review section (equations ***( 13 )***, ***( 14 )*** and***( 15 )***), it was possible to obtain an RP by looking to a 2-year or 3-year history of claims. When estimating RP, some products considered different groups of variables in order to explain higher or lower levels of risk. Because simple estimations were applied, and some products did not have always a significant amount of data assigned to them, it was not possible to go beyond 2 variables, otherwise, overfitted groups would be built with no representative amount of observations. The risk premium from each period of analysis was the respective observed average risk premium between the observations within each group. If no group was formed, then it was the average risk premium of a given product's population. In the end, the final risk premium of each product was the average of its analyzed years. In cases where variables were assigned to create risk estimates, those were mainly related to bank-wise behavior, demographic or product usage. As a final step, policies had a commission percentage added to the original premium. This commission represented a percentage charged by the distribution channel where each policy was sold.

The following process flows summarize the methodology applied to estimate the Risk Premium.

***For each analyzed period*** $t$

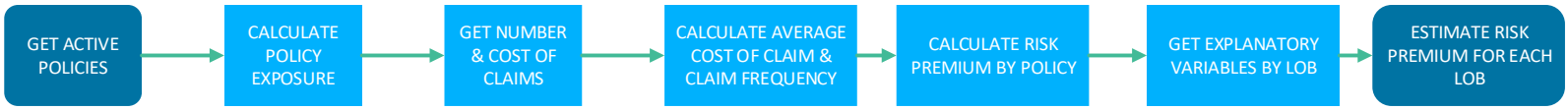



*Figure IV – Steps taken to estimate risk premium by LoB*

***For each considered line of business*** $l$

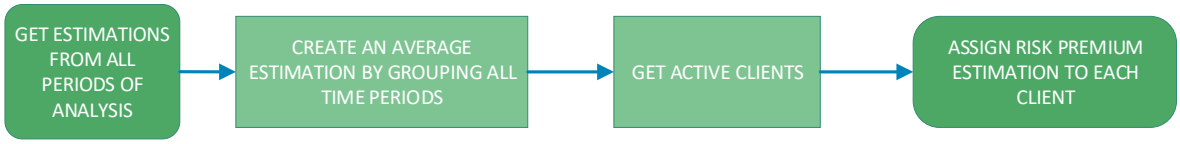



*Figure V – Steps taken to get final RP estimations and assign them to clients*

## 3.7 Cross-sell

Another component that was part of the CLV framework was cross-sell, defined as the sales of additional items related (or sometimes unrelated) to a previously purchased item (Kamakura, 2007). Within the scope of the project, the main objective of this CLV component was to identify three distinct events:

1) ***Probability*** – how likely a given customer is to acquire a policy from a Line of Business where he/she does not have any?
2) ***Value*** – If a given customer does cross-sell for a given Line of Business, how much is he/she likely going to pay for it?
3) ***Risk*** – If a given customer does cross-sell for a given Line of Business, what is the risk assigned to him/her?

However, before analyzing any of the events above, it was necessary to correctly identify, for each LoB, the customers that did not have any product and meanwhile bought at least one in a LoB they weren't present in, i.e., which customer did cross-sell (or not) during the analyzed periods. This occurrence was analyzed in two different 1-year periods, i.e., 1 year ago *VS* Reference date and 2 years ago *VS* 1-year ago. Events where the policy sold was associated to the acquisition of a bank product (i.e., associated sale) were not considered as cross-sell, since these policies resulted from other bank sales and little had to do with the insurance company effort to upgrade its current client's relationship.

To obtain all clients that cross-sold to a given LoB two events were identified:

- **HAD_LoB** $_{i,l}$ – Indicated whether (1) or not (0) each client $i$, had products within a given line of business $l$ at the start date of analysis.
- **BOUGHT_LoB** $_{i,l}$ - Indicated whether (1) or not (0) each client $i$ bought products within a given line of business $l$ at the start date of analysis.

After creating the indicators above, the cross-sell targets for each line of business were identified as follows:

| HAD_LoB $_{i,l}$ | BOUGHT_LoB $_{i,l}$ | CROSS-SELL_ LoB $_{i,l}$ ***(Target)*** |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 1 | **0** |

*Table 4 – Cross-sell target identification for each considered Line of Business*

Having the cross-sell target well defined, it was then possible to compute each one of the previously mentioned events (probability, value and risk).

### 3.7.1 Cross-sell probability

Identically to what happened with the RP computation regarding the risk component, for specific LoB's, the probability of cross-sell was generated by existing Data Ming models. However, these models had specific filters within their process which rejected clients with a given set of characteristics (e.g., super high claim rate, very low tenure, bad credit score, etc.). Because of this situation, simple cross-sell probability estimations were done not only for the LoB's that did not have any probabilistic cross-sell model but also for the ones who did. This enabled every eligible observation of the universe to be classified with a cross-sell probability, complementing a brief lack of coverage from Data Mining models. Similarly, to what was done with the risk component, these 2 years of analysis were divided into two 1-year periods to match the 12-month prediction horizon.

The approach taken to estimate the probability of cross-sell started by identifying and group together the set of clients that did not have any product at the start of any of the analyzed lines of business. Afterwards, for each LoB, it was applied a set of variables that better explained the cross-sell phenomena. In this case, when estimating probabilities, the chosen group of variables was the same for all LoB's, which were related to bank behavior and client's demographic characteristics. The combination of these two types of variables was well known within the business and used to explain purchasing behaviors, therefore they were a solid choice to define this and other components probabilities.

Finally, for each past period analysis and for a given Line of Business, each pair variables had a probability assigned to it, computed by:

$$Cross\text{-}Sell\ Prob_{\,l,b,a,\ T-n} = \frac{\#\ Cross - sell_{l,b,a,T-n}\ (target\ = 1)}{\#\ Obs._{l,b,a,T-n}}$$

**( 21 )**

Where,

$l$ – Line of Business

$b$ – variable 1

$a$ – variable 2

$T$ – Current Year

$n = 1, 2 \ldots \mathrm{N}, \in \mathbb{N}$

After calculating the average cross-sell probability for each period of analysis, the final probability for each $l$, $b$ and $a$ was the average of the corresponding "$l\ b\ a$" groups among the total number of considered periods, $N$. This could be explained by the following formula.

$$Cross\text{-}Sell\ Prob_{\,l,b,a} = \frac{\sum_{n=0}^{N} Cross\text{-}Sell\ Prob_{\,l,b,a,\ \ T-n}}{N}$$

**( 22 )**

As mention before, some LoB's had two cross-sell probabilities assigned to them since their Data Mining models were not able to fully cover all clients. If a given client had already a cross-sell probability given by a Data Mining model regarding a LoB where he/she had no presence in, then that probability would be assigned to that same client regarding the LoB it was meant for. For the remaining LoB's, where the client did not have any cross-sell probability assigned by a model, simple estimations would be applied.

### 3.7.2 Cross-sell value

Considering the rationale described in the section above, the cross-sell value was also estimated by analyzing historical data of clients that cross-sold in the last 2 years.
For each period of analysis and each LoB, the first step to estimate this value was to filter out all clients with cross-sell target = 0. Then, similarly to what was done with the probability component, two highly explanatory variables were chosen to make groups of higher/lower cross-sell value. These were related to bank behavior and client's demographic characteristics

After obtaining the cross-sell value for each LoB and their respective pair of explanatory variables in each period of analysis, the final cross-sell value was the average between the corresponding groups among the considered time periods.

### 3.7.3 Cross-sell risk

With the components of cross-sell value and probability estimated, it was also necessary to assign a risk value to each potential cross-sell occurrence. Because clients that were going to be assigned this estimation had no presence in the analyzed lines of business, they did not have their own risk premium, since this value is usually generated at the moment of a simulation and allocated to a policy after it being purchased. With this in mind, the alternative found was to gather all policies corresponding to cross-sell events (i.e., policies with cross-sell target = 1) in each LoB and estimate an average risk premium per pair of values from each explanatory variable, based on the one assigned to those same policies. The previous situation was only applied for LoB's where the RP was generated by a risk model. For the remaining LoB's, the estimation of cross-sell risk premium was the same as the original estimation, presented in section 3.6. Finally, after estimating each cross-sell component the final cross-sell value of each client was calculated based on equation ***( 12 )***.

Overall, the full cross-sell methodology could be summarized by the following process flows.

***For each analyzed period*** $t$ *and* ***line of business*** $l$

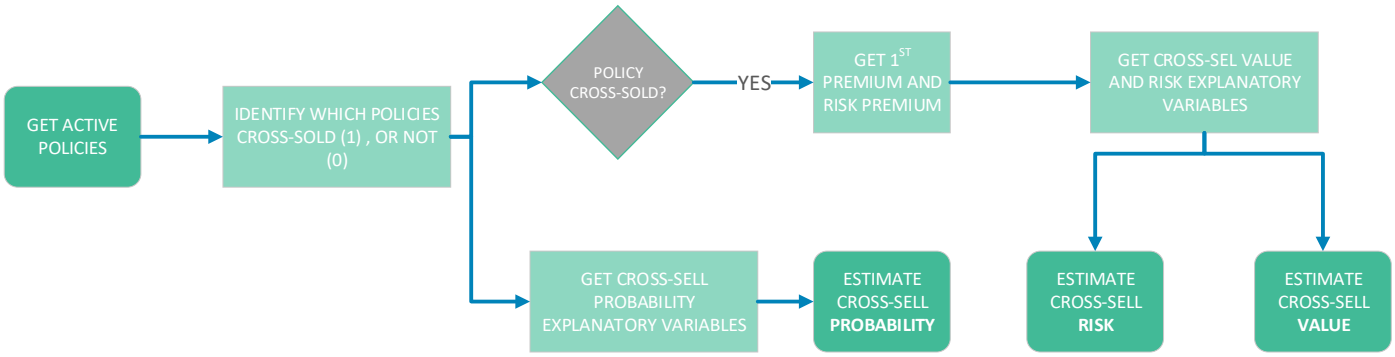



*Figure VI – Steps taken to estimate cross-sell probability, risk and value by Line of Business*

***For each considered line of business*** $l$ ***and cross-sell element*** (value, risk and probability)


*Figure VII – Steps taken to assign all cross-sell elements (value, probability and risk) to each client*

***For each client*** $i$ ***and LoB*** $l$ ***where a client has no presence in***


*Figure VIII – How multiple cross-sell probabilities were reduced to one per client*

## 3.8 Churn

Another important component that was part of the CLV framework was the churn propensity, which determined how likely a given customer was to abandon a Line of Business. In the scope of this project churn was defined as being any intentional form of defect, i.e., death, age limit, or other reasons that led to the "automatic" departure of a client, were not classified as churn.

To estimate churn, historical behaviors had to be analyzed, more specifically, churn behaviors during the past 12 months. This was done for each Line of Business, but also by type of sale (associated or active) when it made sense. This last distinction was applied because there were clear churn discrepancies between associate and active sale policies, as it is shown in the example below:

*By analyzing the tenure (in years) of the churned policies during the year of 2017 from a given LoB with active and associated sales, the % of churn between the two types of sales is presented by the chart below.*

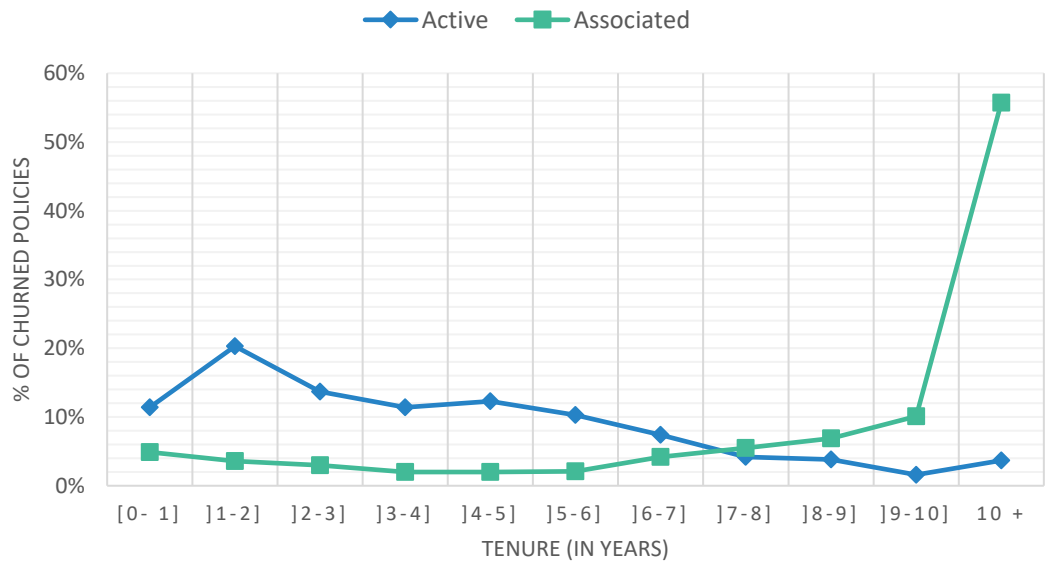



*Chart 6 – Percentage of churned policies in 2017 by Tenure (in years) and Sale Type (Active vs Associated)*

Clearly, it was possible to understand that associated sale clients churned more in high tenure values, while active sales clients had the opposite behavior (i.e., churned more in lower levels of tenure).

The first step to estimate churned policies was to identify which ones churned in the last 12 months, taking into account the two previous aspects (LoB and Type of sale). Afterwards, similarly to what was done with the estimation of previous components, variables were explored to find which could better explain this phenomenon. One common variable to all churn analysis was policy tenure. With this in mind, it was decided to make a brief analysis regarding this variable to better understand how it explained churn in the set of considered observations. By doing so, it was clear that churn events and tenure were highly linked to each other, however, distinct behaviors were observed

depending on each analyzed LoB. Because of this, several tenure groupings were formed according to higher or lower churn ratios in each line of business (see annex **A 2**). The final set of tenure groupings are presented in the table below.

| **A** | **B** *(Active)* | **B** *(Associated)* | **C** | **D** *(Active)* | **E** | **F** |
|---|---|---|---|---|---|---|
| [0 -2] | [0-3] | [0-2] | [0 -3] | ]1- 2] | [0-2] | [0 -2] |
| ]2-3] | ]3-4] **U** ]13-15] | ]2-3] **U** ]9-11] | ]3-5] | [0 – 1] **U** ]2-6] | ]2-5] | 2 + |
| ]3-5] | ]4-13] **U** 15 + | ]3 – 9] **U** 11 + | 5 + | 7 + | 5 + | |
| 5 + | | | | | | |

**LOW** **HIGH**

*Churn risk*

*Table 5 – Final set of Tenure groupings by Line of Business and Sale type (LoB B and D, only)*

There was no clear pattern indicating if tenure was relevant for churn in associated sales policies of LoB D, therefore this variable was not used in that group.

After analyzing tenure, an attempt to join other explanatory variables to each group was made. Once again, the most relevant ones were related to bank-wise behavior and demographic information.

The probability of churn for each LoB group and each set of variables was obtained by the ratio which divided the number of churned policies by the total number of policies within that set.

$$Churn\ Prob._{l,s_j} = \frac{\#\ Churn_{l,s}\ (target\ = 1)}{\#\ Obs._{l,s}}$$

***( 23 )***

Where,

$l$ – Line of Business

$s_j$ – set of variables $j$

$j = 1, 2 \ldots \mathrm{J} \in \mathbb{N}$

Since the final churn probability had to be assigned at the client level, but the analysis was made at the policy level, there were situations where a client could have more than one churn probability if he/she held more than one policy in the same line of business. To fix this situation, it was decided that, for each line of business, the final churn probability of a customer would be the minimum probability among all its policies. Business-wise this was the solution that made more sense because what was being measured was the likelihood of a given customer to churn from the whole Line of Business and not just one of the several products he/she hold on that same LoB. In this sense, it was assumed that if a customer churned one policy he/she would churn to the remaining policies of the

same LoB. This situation may or may not be the ideal one but, with the existing time limitations, it was the one which had higher acceptance from business stakeholders.

Another aspect that was affected by time constraints was the analysis of cannibalization events. These events could be identified as "false churn" since clients who had this type of behavior cancelled their policies and acquired a new one after a small period of time with better conditions (price, coverage amount, number of coverages, etc...). This situation can happen in any line of business, however, it is known there are some where this is more common to occur. Because cannibalization was not removed, it was known some policies were falsely identified as churn, however, these events represent a small portion of the whole churn universe (e.g., 3,6% in 2017 for LoB D), so their effect is not likely to significantly impact the created estimations.

Overall, the churn estimation methodology could be summarized by the following process flow.

***For each considered line of business*** $l$

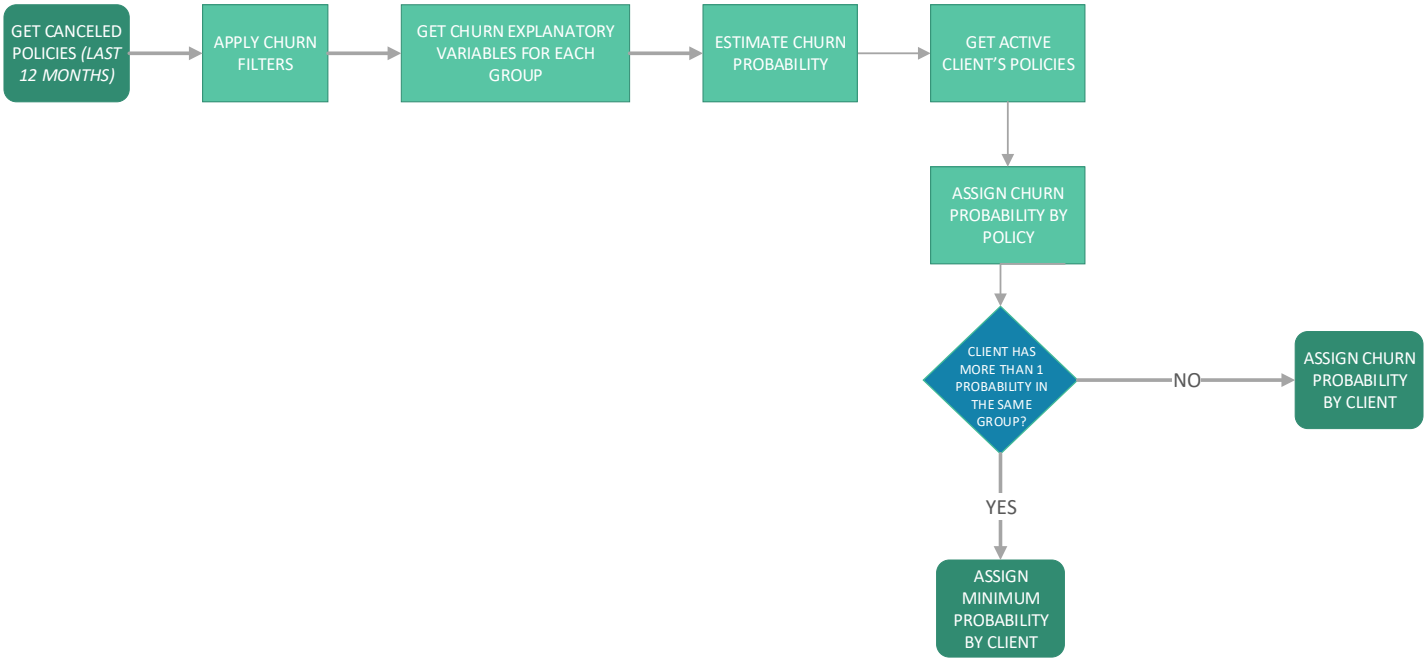



*Figure IX – Steps taken to estimate churn probability by Line of Business at the client level*

## 3.9 Upsell

The last component to be built was upsell. Upselling could be briefly defined as being the practice in which a business tries to persuade customers to purchase a higher-end product, an upgrade, or an additional item in order to make a more rewarding sale (dun & bradstreet, 2016). This component only took into consideration LoB A, C and D because these were the ones where upsell occurred with more frequency due to the wide variety of options in terms of products. Only active sale policies were considered since the conditions of associated sale ones are static for the remainder of the contract with the bank.

Within the scope of this project upsell was identified by an increase of premiums followed by at least one of the scenarios below:

1) Increase in **coverage amount** with the same policy;
2) Increase in the **number of coverages** with the same policy;
3) Increase in the **number of insured people/objects** with the same policy;
4) Increase the **number of policies** in the same line of business;

One important aspect to mention is the fact that there were some situations where at least one of the four scenarios above occurred but not due to an upsell action. It is possible to call them "false upsell" scenarios.

| FALSE UPSELL SCENARIO | REASON |
|---|---|
| Increase in the number of coverages | *New coverages with lower amount assigned to them i.e., lower risk for the company and lower price.* |
| Premium increase | *Commercial tariff alterations and/or uninventable events (e.g., age increase)* |
| Increase in the coverage amount | *Commercial tariff alterations and/or changes in the law* |

*Table 6 – False upsell situations and the reasons behind them*

Naturally, it was difficult to properly identify upsell situations not only because there weren't clearly defined business rules to do so, but a lot of scenarios were also difficult to access as true or false upsell occurrences. Baring this in mind, in order to reduce the number of situations falsely identified as upsell, the adopted target changed according to each analyzed LoB. The table below presents which set of conditions were identified (**1**), or not (**0**), as upsell events.

| LINE OF BUSINESS | CONDITIONS | UPSELL? |
|---|---|---|
| C | # COVERAGES INCREASED ∧ PREMIUM INCREASE | 1 |
| | COVERAGE AMOUNT (€) INCREASED ∧ PREMIUM INCREASE | 1 |
| | # INSURED OBJECTS/PEOPLE INCREASED ∧ PREMIUM INCREASE | 1 |
| | # POLICIES INCREASED | 1 |
| | OTHERWISE | 0 |
| D | # COVERAGES INCREASED ∧ PREMIUM INCREASE | 1 |
| | COVERAGE AMOUNT (€) INCREASED ∧ PREMIUM INCREASE | 1 |
| | # INSURED OBJECTS/PEOPLE INCREASED | 1 |
| | # POLICIES INCREASED | 1 |
| | OTHERWISE | 0 |
| A | # COVERAGES INCREASED ∧ COVERAGE AMOUNT (€) ∧ INCREASED PREMIUM INCREASE | 1 |
| | # INSURED OBJECTS/PEOPLE INCREASED ∧ PREMIUM INCREASE | 1 |
| | # POLICIES INCREASED | 1 |
| | OTHERWISE | 0 |

*Table 7 – Adopted upsell conditions for each LoB*

The presented upsell conditions of each line of business differed essentially according to their respective product characteristics.

After identifying the several upsell scenarios, simple estimations were built similar to the ones of cross-sell and churn, where pairs of explanatory variables defined how higher or lower the probability and value of upsell would be. No element of risk was assigned to potential upsell scenarios because there was no available process to retrieve past risk premiums for all considered LoB's in order to make the difference between the current and previous risk assigned to each upselling policy.

It is important to mention that cannibalization was also not considered in the upsell process. Even though this phenomenon is mostly related to churn events, there could be some cases where a new and upgraded policy with better conditions was subscribed to replace an older one.

### 3.9.1 Upsell Probability

The probability element of upsell was estimated through a process similar to the one implemented for cross-sell probability. The set of chosen explanatory variables to estimate this element were once again related to bank-wise behavior and demographic information. Bearing in mind the whole set of considered policies, upsell probabilities were estimated by looking at the proportion of upsell occurrences of the last 12 months in each pair of explanatory variables. In this case, only one period of time was analyzed, so there was no need to make any comparison with other periods, unlike what happened with cross-sell or risk premium estimations.

In the end, the probability of upsell was assigned to each active client based on its presence within each considered LoB, matching him/her characteristics with the corresponding upsell probability group. Because the upsell probability explanatory variables were at the client-level, when a customer had multiple policies in the same LoB its upsell propensity would not suffer any change since the variables would be the same independently on the number of policies a given client had.

### 3.9.2 Upsell Value

The process to estimate upsell value only took into account policies/clients linked to upsell scenarios, i.e., policies or clients which had upsold during the period of analysis. In this case, the chosen explanatory pair of variables was the same as the ones of upsell probability.

To estimate the value of upsell, two different approaches were taken according to the observed upsell scenario:

1. **New policies from the same LoB** – Value of upsell was considered to be the premium of the new policy.
2. **Remaining Upsell scenarios (New Objects, covers, etc.)** – Value of upsell was considered to be the difference between the older and new premium to be paid by a given client.

In the end, the value of upsell was assigned to each active client based on its presence within each considered LoB and by making the match by age group and bank segment with the produced estimation.The final upsell value of each client was then given by equation **( *11* )**, presented in the literature review section

The processes below summarize the set of steps taken to obtain a final upsell value to the universe of clients.

***For each considered line of business*** $l$

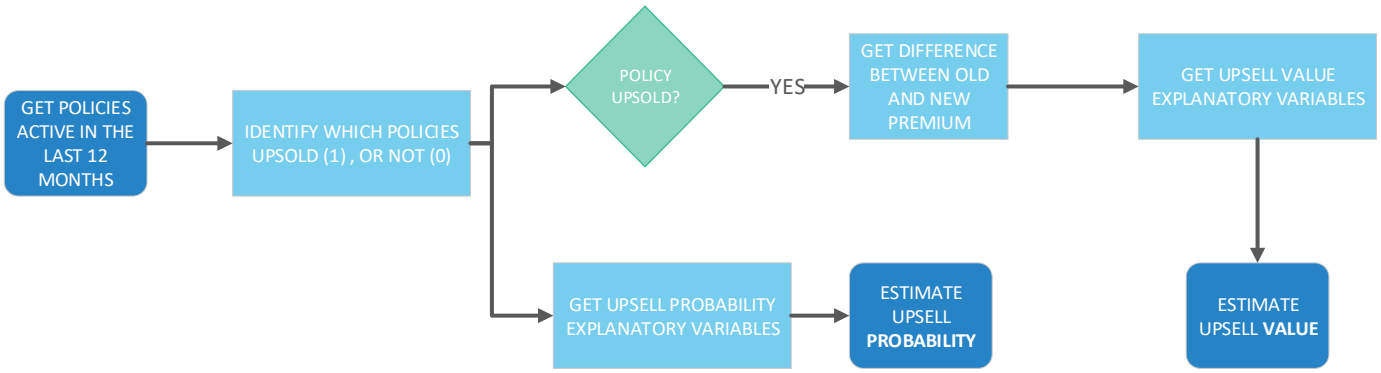



*Figure X - Steps taken to estimate upsell probability and value by Line of Business*

***For each considered line of business*** $l$

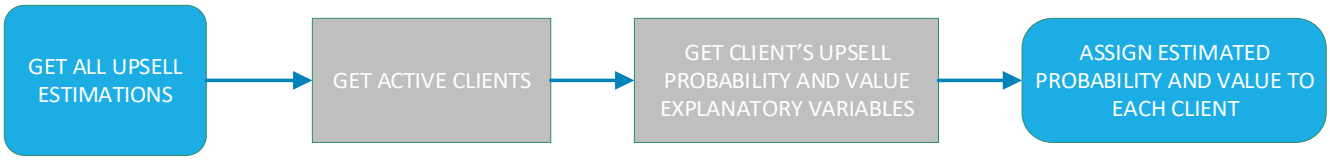



*Figure XI - Steps taken to assign all upsell elements (value, probability and risk) to each client*

## 3.10 Data discretization

After building each component, either through a complex *SAS* model, a simple estimation or just by gathering data from business tables, to calculate CLV for all defined perspectives (Global, or by Line of Business) following equation **( *16* )**, it was still necessary to produce a user-friendly output, easily interpreted by technical and non-technical business stakeholders and would clearly characterize customers in three different levels of detail: Globally, LoB-wise and Component-wise (Figure II and Figure III) . To obtain such output, it was implemented a process of data discretization, used to reduce the number of values for a given continuous attribute, by dividing the range of the attribute into intervals (Kurgan & Cios, 2001). A normal discretization process specifically consists of four steps (Liu, Hussain, Tan, & Dash, 2002):

1) Sort all the continuous values of the feature to be discretized;

2) Choose a cut point to split the continuous values into intervals;

3) Split or merge the intervals of continuous values;

4) Choose the stopping criteria of the discretization process;

Several data discretization methods exist however, time limitations did not allow a thorough research on this topic, which led the adopted method to be one already familiar in similar business processes: Equal Frequency Binning. This is one of the most well-known and simple binning methods and it is characterized by dividing a continuous-valued attribute into a specific number of bins (Liu, Hussain, Tan, & Dash, 2002), each defined by a numeric interval and, as the name infers, the number of observations in each bin is similar. However, when using the previous binning method, not always its output made sense business-wise. Below is presented an example that supports the previous argument.

*e.g., Consider a set of 50 000 clients to be classified from 1 to 5 (Low to High) according to their respective global CLV. Using a binning method completely aligned with the previously presented theoretical concept those clients would be classified as follows:*

| Bins | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Number of Observations | 10**K** | 10**K** | 10**K** | 10**K** | 10**K** |

*In theory, CLV values could be within the range of* ]- ∞; ∞ +[. *With this in mind, it is very likely there will be a bin which will have simultaneously clients with negative and positive CLV values, as presented in the figure below.*

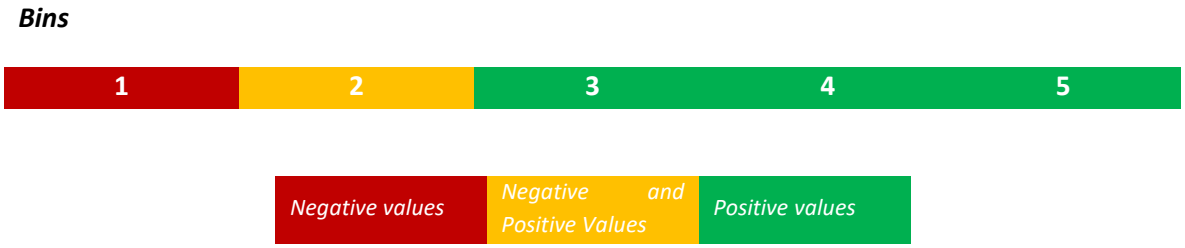



*Figure XII – Default implementation of the Equal Frequency Binning method considering 5 bins*

*Business-wise, it doesn't make sense to mix those two types of clients in the same bin, because customers with negative future value shouldn't be dealt in the same manner as clients with low, but positive future value.*

To face the problem presented in the example above, it was defined that clients with negative value would all be part of the lowest bin (bin 1) and from bin 2 onwards (positive values only) clients would be classified according to the Equal Frequency Binning method.

The implementation of the previous data discretization method was done through the *SAS* procedure *proc rank.* This procedure computes ranks for one or more numeric variables across the observations of a SAS dataset and outputs the ranks to a new one (SAS, s.d.). Because *proc rank* output were ranks (or rankings), from this point forward, that will be the term adopted to identify what was previously called as bins.

As mentioned before, the outputs of three distinct perspectives (Global, LoB-wise and Component-wise) were going to be ranked from 1 to 5. The label of each rank is presented below.


*Figure XIII – Rank label*

There were several reasons why a 5-level ranking approach was chosen, the most relevant being the fact that it was aligned with past product/client classifications made in other company processes and analysis. Another strong reason supporting this approach was the fact that a classification with 5 levels gave a good enough range to properly distinguish high expected future value clients from lower ones, without losing the capability of technical and non-technical stakeholders to easily interpret the results. If applying, for example, a 3-level labelling system of Low, Medium and High, by following the same rationale as before, level 1 would have negative value clients and only 2 levels would exist to characterize positive clients. This way, groups would be too big to clearly identify top and bottom positive-valued customers, and almost no business decisions could be made. Baring everything in mind, between the three perspectives to be classified, two distinct rationales were applied. Globally and LoB-wise, the rank assignment corresponded the one where rank 1 had all clients with negative values and from rank 2 to 5 clients with positive were divided into quartiles and were assigned to the rank that matched their quartile. Rank 2 matched the 1$^{st}$ quartile, rank 3

matched the 2nd quartile and so on. Component-wise, CCV ranks were assigned by following the previous approach, however, the remaining components (upsell, churn, risk and cross-sell) had their ranks assigned in a different manner. This happened because in those scenarios almost no negative values were produced due to the characteristics of each component: i) churn probability or risk value would never be negative; ii) upsell and cross-sell events are ways of increasing revenue so, in theory, those were always positive. Consequently, instead of using rank 1 for negatively valued clients and diving the remaining set of clients (positive values) into groups of 25%, they were divided into groups of 20% instead i.e., quintiles. Afterwards, for each component, each client was assigned to the rank matching its respective quintile, where rank 1 corresponded to the 1st quintile and so forth. Because churn and risk had a negative impact on a client's CLV (the higher they were, the worse), the rank assignment was inverse to the one previously stated, i.e., 1st quintile rank 5, 2nd quintile rank 4, and so on.

The table below summarizes the different ways the ranking process was applied, according to each perspective and component.

| **Ranking Perspective** | **Ranking approach** | **Ranking order** *(Best to Worst)* |
|---|---|---|
| Globally<br><br>Line of Business<br><br>Component (*CCV*) | *Rank 1* – Negative value clients<br>Rank 2 – The 25% lowest clients with positive value<br>Rank 3 – The next best 25% clients with positive value<br><br>*Rank 4* – The next best 25% clients with positive value<br>Rank 5 - The best 25% clients with positive values | *Low to High* |
| Component (*Upsell and Cross-sell*) | *Rank 1* – The 20% lowest clients<br>Rank 2 – The next best 20% clients<br>Rank 3 – The next best 20% clients<br><br>*Rank 4* – The next best 20% clients<br>Rank 5 - The best 20% clients | *Low to High* |
| Component (*Churn and Risk*) | *Rank 1* – The 20% best clients<br>Rank 2 – The next lowest 20% clients<br>Rank 3 – The next lowest 20% clients<br><br>*Rank 4* – The next lowest 20% clients<br>Rank 5 - The lowest 20% clients | *High to Low* |

*Table 8 – Summary of all ranking approaches per perspective and their respective ranking order*

## 3.11 Validation

After producing the results of CLV it was important to understand how reliable they were, not only globally, but also for each Line of Business. In that sense, to validate the outputs, it was decided to do a back-test validation, which consisted in applying the CLV Framework to the classified clients based on their characteristics 12-months before the reference date of analysis (**15th of February 2018**) and compare their future value (CLV) prediction with the observed current value (CCV). The following scheme summarizes how the back-test validation process was developed.

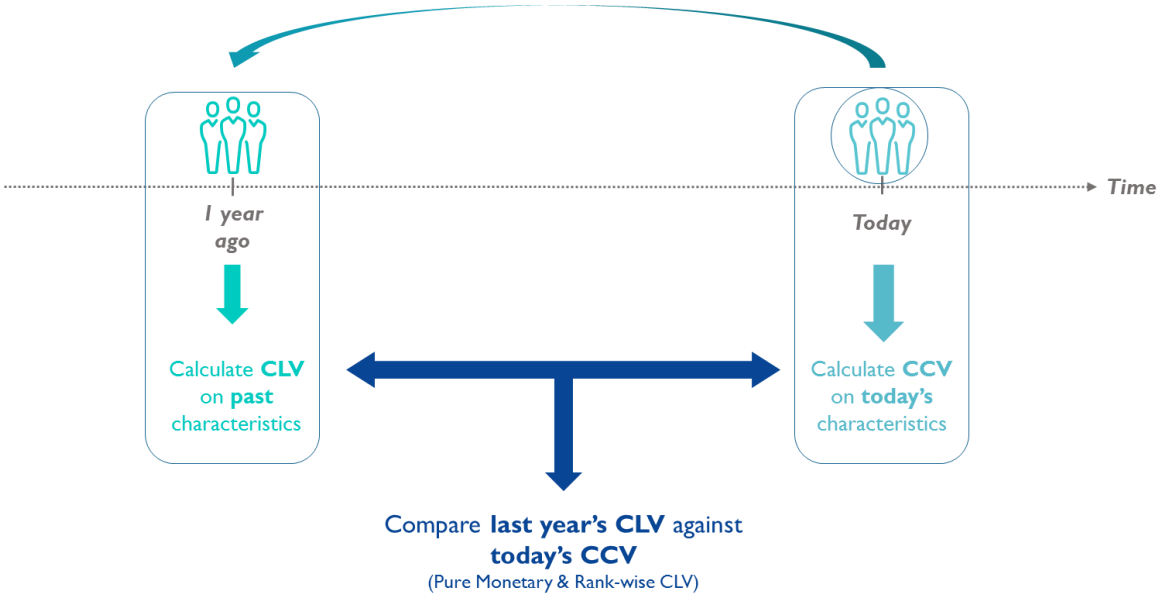



*Figure XIV – Back-test process scheme*

Out of 713 125 clients that were given a CCV, 642 798 were active 1 year ago from the reference date and/or eligible to be classified by the CLV framework. The reason why fewer clients were taken into account when moving 1-year back could be explained by the following scenarios:

- The client was not active 1 year ago;
- The client did not comply with the CLV framework eligibility conditions (Table 1);

To compare these two time periods, it was necessary to look back to the past values of some variables. These variables represented the past characteristics of a client (demographic or bank behavioral data) or its policies. This process of looking back to the past value of policy and client variables was only possible using the Analytic Base Table (ABT) which was designed to retrieve information given a specific date of analysis. The following table presents which group of variables had their past values checked in order to enable all CLV components to be re-calculated.

| CLIENT-LEVEL | POLICY-LEVEL VARIABLES |
|---|---|
| ▪ Demographic variables<br>▪ Bank variables<br>▪ LoB Ownership Variables (e.g., Has/Had LoB_A/B/C …); | ▪ Premiums<br>▪ Claims<br>▪ Commissions<br>▪ Number of insured people/objects |

*Table 9 – Variables that had their past values checked*

Having explained how the CLV outputs were evaluated, in the following section performance results will be presented and discussed.

# 4. Results

As it was already mentioned, both CLV and CCV values were ranked from 1 to 5, according to a well-defined logic, which assigned the rank 1 to clients with CLV/CCV less or equal to 0 and the remaining ranks (from 2 to 5) were assigned based on the quartile distribution of the positive CLV/CCV values. With this in mind, two performance evaluations were made:

- **Continuous value performance**: Comparing CLV and CCV original values, based on error measurements;
- **Rank-wise performance**: Comparing CLV and CCV ranks as if they were discrete target values (with 5 levels);

For both cases, the predicted target was the value produced by the CLV of 1 year ago, while the observed target was the value obtained by the CCV. This evaluation rationale was applied both at the Global and the LoB-levels.

## 4.1 Continuous value performance analysis

The first performance results analyzed were related to the estimated future monetary value (in €) that each customer would be generating during the next 12 months. This analysis was done both at a global and at a Line of Business level. In this particular scenario, the final output was a continuous value, so the chosen evaluation procedure was to measure how far-away the estimated output (predicted value) was from the observed value (true value). The error measurements used to evaluate the error *Mean Absolute Error* (MAE) and *Relative Absolute Error* (RAE) since those were two very well-known metrics which complemented one another and enabled errors to be interpreted. With the error measures determined, the outputs were then evaluated. The obtained results are presented within the following table.

| PERSPECTIVE | METRIC | MINIMUM | MAXIMUM | AVERAGE | STANDARD DEVIATON | MAE | RAE |
|---|---|---|---|---|---|---|---|
| **GOLBAL** | **CCV** | - 498 483,16 | 85 450,37 | 790,81 | 2 790,19 | ± 618,13 | 0,63 |
| **GLOBAL** | **CLV** | - 264 406,91 | 123 952,76 | 1 231,75 | 2 407,89 | | |
| **A** | **CCV** | -289 516,77 | 28 687,64 | 717,10 | 3 252,95 | ± 1172,36 | 0,88 |
| **A** | **CLV** | -115 922,72 | 41 299,27 | 1 584,36 | 2 747,18 | | |
| **B** | **CCV** | -498 483,16 | 65 633,99 | 895,18 | 2 843,00 | ± 418,33 | 0,47 |
| **B** | **CLV** | -119 826,55 | 123 562,16 | 1 203,01 | 2 003,31 | | |
| **C** | **CCV** | -303 197,51 | 17 551,75 | 326,53 | 2 236,66 | ± 354,38 | 0,73 |
| **C** | **CLV** | -265 181,01 | 16 254,48 | 347,82 | 1 928,84 | | |
| **D** | **CCV** | -193 369,99 | 18 806,18 | 237,82 | 1 178,08 | ± 138,06 | 0,51 |
| **D** | **CLV** | -174 033,31 | 21 475,75 | 277,16 | 1 066,41 | | |
| **E** | **CCV** | -48 869,02 | 17 146,08 | 180,14 | 272,14 | ± 50,02 | 0,49 |
| **E** | **CLV** | -24 895,41 | 17 160,28 | 204,63 | 226,89 | | |
| **F** | **CCV** | -60 594,62 | 10 864,40 | 198,47 | 623,49 | ± 289,36 | 1,18 |
| **F** | **CLV** | -71 334,19 | 17 226,07 | 416,11 | 796,57 | | |

*Table 10 – Customer Lifetime Value Back-test performance in terms of its continuous value*

Based on the table above it is possible to observe the framework as a lower RAE in LoB B, E and D with 0,47, 0,49 and 0,51, respectively. Even though these areas show the best performance, by looking to their respective MAE's, by trying to make sense of them business-wise, at first glance, one can naively conclude that the error value is still high for a given output to be considered truly reliable. For example, by taking LoB B MAE of $\pm$ 418,33, what is being said is that, on average, the calculated CLV for this Line of Business is 418,33€ away from its true value. Depending on the analyzed LoB, this value could be close to the average annual premium paid, which means any future estimations made to the presented could be on average close to $\pm$ 1 annuity from its true value.

Nevertheless, part of these higher error rates could be justified by fact that CLV is impacted by churn, which always reduces the future value of a customer depending on the higher or lower values of this component, while CCV is not. Given that a customer stays within a LoB between two consecutive time periods, this situation leads to an increased absolute differential between CLV and CCV, which in the end will also increase the observed error. Additionally, another aspect which could explain an increase in the observed errors is the fact that there were LoB's with a significant number of components built by simple estimations. Although, while building these estimations it was made an effort to maintain contact with business stakeholders to better understand which variables could be most explanatory for each considered component, since these estimations resulted from very simple processes, it was known that most of them could not fully reflect the complexity of the interactions that were being explained. Lastly, the fact that no outliers were removed since every customer had to be classified, could also have led extreme value clients to negatively impact the performance results. Overall, considering the previous three aspects, it can be said that the obtained continuous values are slightly positive and could be reliable if interpretations are made having in mind its limitations.

Nonetheless, it was still necessary to analyze the rank-wise results, on which there were higher expectations in terms of its Overall and Line of Business performance, justified by the fact that this second output had greater resistance against the limitations affecting the continuous value performance. Further details are given in the next section.

## 4.2 Rank-wise performance analysis

The fact that this output dynamically grouped client's CLV into ranks according to their distribution, made it resilient to the limitations affecting the continuous value output. Churn had a lower impact on the final result because CLV and CCV were ranked separately according to their own distributions. This meant, even though churn could significantly reduce a client's CLV it could still match the same CCV rank. On the other hand, extreme value clients had less effect in the rank-wise performance, since they would either belong the highest (5) or lowest rank (1), without the difference of CLV and CCV affecting that much.

Among the 642 798 customers considered to make the rank performance evaluation, this was how they were distributed across CCV and CLV (of last year) rank.

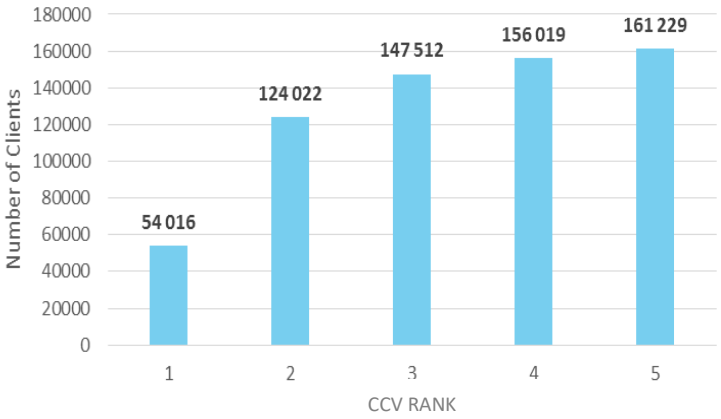



*Chart 7 – Client distribution among last year's Current Customer Value (CCV) Rank*

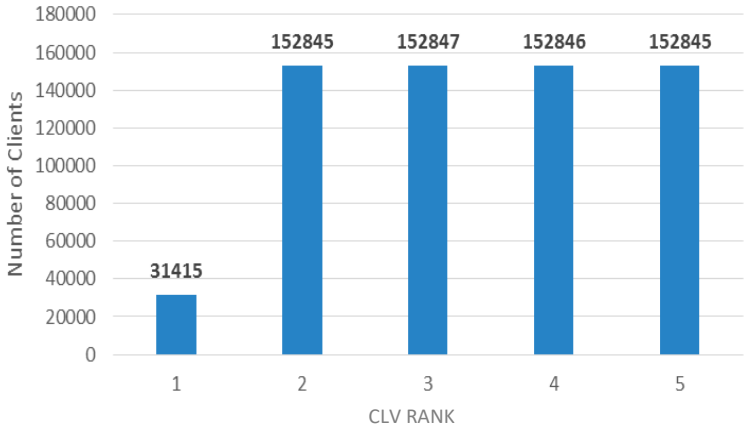



*Chart 8– Client distribution among last year's Customer Lifetime Value (CLV) Rank*

Each time the CLV rank that was equal to CCV rank, it was considered as a correct prediction (IS_CORRECT = 1), otherwise, it was considered an incorrect one (IS_CORRECT = 0). The global accuracy of the CLV framework is presented in the following table.

| IS_CORRECT | Number of Clients | % of Clients |
|---|---|---|
| 0 | 220 439 | 34,29 |
| 1 | 422 359 | 65,71 |
| TOTAL | 642 798 | 100 |

*Table 11 – Global CLV Back-test performance results*

Looking at the table above it is possible to understand that approximately 66% of the predicted CLV ranks were correctly assigned when compared to their respective CCV ranks. However, there is still a significant number of clients (≈ 220,4K) which had their predicted rank incorrect. One scenario that could have led the incorrect rank values to increase, could have been the different split values between the corresponding CLV and CCV ranks. Because ranks are produced based on the quartile

distribution, this could have led to discrepancies and mismatches, since the CLV and CCV distributions are distinct from each other.

To understand with higher detail the performance achieved on the prediction of each target level, it was produced the following confusion-matrix, where the green highlighted cells represent the CCV rank with higher correct observations per CLV rank.

| **CLV RANK** (Predicted Value) | **CCV RANK** (Observed Value) 1 | 2 | 3 | 4 | 5 | **TOTAL** |
|---|---|---|---|---|---|---|
| 1 | 17 172 | 8 242 | 2 549 | 2 249 | 1 203 | 31 415 |
| 2 | 9 704 | 98 782 | 39 781 | 3 796 | 782 | 152 845 |
| 3 | 9 231 | 10 227 | 88 266 | 43 305 | 1 820 | 152 849 |
| 4 | 9 309 | 4 639 | 12 967 | 93 323 | 32 608 | 152 846 |
| 5 | 8 604 | 2 132 | 3 949 | 13 346 | 124 816 | 152 847 |
| TOTAL | 54 020 | 124 022 | 147 512 | 156 019 | 161 229 | **642 798** |

*Table 12 – Number of classified clients per CLV rank by CCV rank*

To better understand the proportions of the correct and incorrect number of observations per predicted target level, it was also produced the table below. Once again, the cells with the correct values were highlighted in green.

| **CLV RANK** (Predicted Value) | **CCV RANK** (Observed Value) 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0,55 | 0,26 | 0,08 | 0,07 | 0,04 |
| 2 | 0,06 | 0,65 | 0,26 | 0,02 | 0,01 |
| 3 | 0,06 | 0,07 | 0,58 | 0,28 | 0,01 |
| 4 | 0,06 | 0,03 | 0,08 | 0,61 | 0,21 |
| 5 | 0,06 | 0,01 | 0,03 | 0,09 | 0,82 |

*Table 13 - Proportion of classified clients per CLV rank by CCV rank*

As it is shown above, the proportions corresponding to the correctly predicted observations are always the biggest ones on each predicted target value, which is a positive insight. However, the framework clearly shows inconstant performance to predict different target values. While to predict the **target-level 5** the framework shows an **accuracy of 82%**, regarding the **target-level 1**, the performance **decreases by 27%**, in comparison.

Aside from the previous insights, there are 3 other relevant observed situations worth being discussed, those being:

1. If the predicted CLV rank was always equal to each client CCV rank of 12-months ago, the framework would have an overall accuracy of 71%. This means that when the framework tries to predict change, it loses 6% correct observations in comparison to its observed Global-wise accuracy. The table below presents some insights regarding this problem.

| SCENARIO | NUMBER OF CLIENTS | % OF CLIENTS | OBSERVATION |
|---|---|---|---|
| **LAST YEAR'S CCV RANK ≠ CLV RANK** | 118 038 | 18 | *Where the framework predicted change* |
| **LAST YEAR'S CCV RANK ≠ CLV RANK** AND **IS_CORRECT =1** | 30 802 | 5 | *Where the framework predicted change and was correct.* |
| **LAST YEAR'S CCV RANK = TODAY'S CCV RANK** | 458 005 | 71 | *Clients who maintained last year's rank* |
| **LAST YEAR'S CCV RANK < TODAY'S CCV RANK** | 114 574 | 18 | *Clients who decreased last year's rank* |
| **LAST YEAR'S CCV RANK > TODAY'S CCV RANK** | 70 225 | 11 | *Clients who increased last year's rank* |
| **CLV RANK ≠ TODAY'S CCV RANK** AND **(LAST YEAR'S CCV RANK ≠ CLV RANK)** | 87 236 | 14 | *Where the framework predicted change and was not right* |
| **CLIENTS PREDICTED TO CHANGE AND DIDN'T** | 66 356 | 10 | *Where the framework predicted change and clients maintained their rank* |

*Table 14 – Set of Insights on the CLV rank-wise back-test validation performance*

Based on the table above it is possible to realize 29% of clients (≈ 184,8 K) changed their global CCV rank, however, the framework was only able to correctly predict change on 5% (≈ 30,8K) of them, even though it originally predicted change of 18% (≈ 118 K) of the total observations. This indicates that among the clients who changed their rank, the framework was only capable to correctly identify 17% of the observations, translating into a poor capability of predicting change. On the other hand, the framework also predicted 82% of the observations would maintain their current rank, getting it right in 61%, out of the observed 71%, presenting an accuracy of ≈ 85% for this "static" scenario.

2. Clients with a negative CLV are very difficult to predict since they are usually linked to rare scenarios leading to an abnormal increase in claim costs, such as serious diseases, rare natural disasters, among other severe unfortunate events. The table below presents the inverse perspective of the previous confusion-matrix (Table 6), i.e., the proportion of observations per CCV rank by each CLV rank (analysis by column).

| | | **CCV RANK** (Observed Value) | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| **CLV RANK** (Predicted Value) | 1 | 0,32 | 0,07 | 0,02 | 0,01 | 0,01 |
| | 2 | 0,18 | 0,80 | 0,27 | 0,02 | 0,00 |
| | 3 | 0,17 | 0,07 | 0,60 | 0,28 | 0,01 |
| | 4 | 0,17 | 0,04 | 0,09 | 0,60 | 0,20 |
| | 5 | 0,16 | 0,02 | 0,03 | 0,09 | 0,77 |

*Table 15 - Proportion of classified clients per CCV rank by CLV rank*

Taking into consideration the table above it is possible to understand that the set of observations with CCV rank 1 is way more dispersed throughout the several CLV ranks,

supporting the fact that this rank is the most difficult to correctly identify due to the fact of often being associated with highly unpredictable events. Because of this scenario, the framework's performance was once more analyzed, this time filtering out this rank both at the CCV and CLV side.

The two confusion-matrixes below present the performance results with and without considering rank 1.

| CLV RANK (Predicted Value) \ CCV RANK (Observed Value) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0,55 | 0,26 | 0,08 | 0,07 | 0,04 |
| 2 | 0,06 | 0,65 | 0,26 | 0,02 | 0,01 |
| 3 | 0,06 | 0,07 | 0,58 | 0,28 | 0,01 |
| 4 | 0,06 | 0,03 | 0,08 | 0,61 | 0,21 |
| 5 | 0,06 | 0,01 | 0,03 | 0,09 | 0,82 |

| CLV RANK (Predicted Value) \ CCV RANK (Observed Value) | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 2 | 0,69 | 0,28 | 0,03 | 0,01 |
| 3 | 0,07 | 0,61 | 0,30 | 0,01 |
| 4 | 0,03 | 0,09 | 0,65 | 0,23 |
| 5 | 0,01 | 0,03 | 0,09 | 0,87 |

*Table 16 - Proportion of classified clients per CLV rank by CCV rank (considering rank 1 VS not considering)*

Based on the results of the table above it is possible to notice that the overall performance increases for every rank by 3% to 5%. This leads to an **increase in overall accuracy from 65,71% to 70,5%**.

3. On every CLV rank, the 2nd CCV rank with higher proportion is always the rank right after to the correct one (e.g., for CLV rank 2 the CCV rank that has the 2nd highest proportion of observations is CCV rank 3, and so on.). CLV rank 5 has the same behavior, but for its previous CCV rank. This could imply that incorrectly classified observations, in most cases, were only 1 rank away from its correct value. To better understand if this situation was in fact true, incorrect observations were analyzed.

   In the test set, it was verified how many observations differed just 1 CLV rank from their observed CCV rank. With this approach, it was found that, at the global level, **23,7%** of the observations (≈152K) their **CLV rank differed from their CCV rank by just ± 1 rank**. This indicated the CLV framework most of the times misclassified by the minimum rank margin of error.

Regarding the 3 notes above, a similar approach was applied to analyze CLV rank-wise performance for each Line of Business, however, since most behaviors were similar to what was observed at the global-level, the LoB results were made available in the appendix section (see annex **A 3**). The following table summarizes the obtained performance regarding the back-test validation process, given each Line of Business.

| LINE OF BUSINESS | TOTAL OBSERVATIONS | % CORRECT | % CORRECT (WITHOUT RANK 1 ) | % ± 1 RANK (WITHOUT RANK 1 ) | % RANK > ± 1 (WITHOUT RANK 1 ) |
|---|---|---|---|---|---|
| GLOBAL | 642 800 | **65,7** | **70,5** | **23,7** | **5,8** |
| A | 131 766 | **55,4** | **60,1** | **32,2** | **7,7** |
| B | 298 927 | **61,0** | **69,3** | **25,9** | **4,8** |
| C | 90 076 | **49,0** | **50,9** | **35,6** | **13,5** |
| D | 279 620 | **69,9** | **71,8** | **22,2** | **6,0** |
| E | 132 212 | **73,2** | **73,4** | **22,3** | **4,3** |
| F | 131 591 | **56,7** | **68,0** | **26,0** | **6,0** |

*Table 17 – Back-test validation performance globally and by Line of Business*

Considering the table above, it is possible to notice that LoB E has the greatest performance (**73,2%**). Even disregarding rank 1 makes little impact and almost observations (**95,7%**) are either correct or just one rank away from the true value, which business-wise should not have a big impact. On the other hand, LoB C has the lowest accuracy (**49%**) and is the highest with incorrect observations within 1 rank of difference from the true rank (**35,6%**), as well as differences of > 1 rank (**13,5%**). Further analysis should be conducted to comprehend why this Line of Business has such a low accuracy compared to the remaining ones.

# 5. Conclusion

Transitioning academic concepts to a corporate environment is rarely an easy process, especially due to the complexity of the day-to-day processes that businesses rely on. This project was one of those examples, where for 9 months, a Customer Lifetime Value estimation over a 12-month horizon was calculated to all individual bancassurance clients belonging to 6 different lines of business from one of the top bancassurance companies operating in Portugal. As determined, the final output was assigned to the main representative of each policy (policy-holder). Additionally, the developed metric successfully incorporated potential customer interactions, namely churn, cross-sell, risk and upsell. This structure was only possible to be achieved by integrating pre-existing data mining models and building simple estimations to cover the remaining scenarios models were not able to.

As of February of 2018, the final output was delivered to 713 125 unique customers into two distinct forms – continuous and rank value – each of those having their unique performance results, evaluated through a back-test validation. Regarding the continuous CLV results globally, *Relative Absolute Error* was 0,63 followed by a corresponding *Mean Absolute Error* of ± 618,33. At first glance, these results may not seem the most appealing since business-wise, this margin of error could be close to the annual price a customer might pay for his/her insurance. Nevertheless, it is relevant to mention they had some associated constraints, such as: *i*) the significant impact of the churn component, which increased the difference between CCV and CLV; *ii*) the fact that a lot of components were only covered by simple estimations which possibly were only able to explain part of the reality they were assigned to; *iii*) no extreme value clients were filtered out when evaluating the model performance, which could've also impacted its results. Rank-wise, outputs were more resilient to the previous continuous value constraints, which lead to the positive global results of 65,71% accuracy and of 70,5% if rank 1 (related to rare unfortunate scenarios) was not considered. Nonetheless, through the analysis of this second evaluation method, it was possible to understand the model was only able to correctly predict change 17% of times, however, 23,7% of incorrect observations differed from their true rank by the minimum distance of 1.

Insurance-wise, even though this version of CLV still needs to be fine-tuned, its outputs could be applied in several initiatives across multiple departments to help the business thriving, such as: *i)* leads prioritization, guiding campaign decision-making; *ii*) Affinity programs, or new product offering/design, rewarding customers who have higher potential or contributing towards the business; *iii*) Service quality or Claims handling, by providing superior assistance to clients presumed to generate higher future value; i*v*) Pricing, employing CLV as a premium rating variable; etc. However, for any of the previous interactions to be successful, stakeholders must be engaged and aware of the potential benefits of CLV. This way, the company will be able to get proper insights on its clients, strengthen their relationship and, consequently, increase their lifetime value, while continuously generating more profitability.

Academic-wise, there is set of explored aspects in this project which make it unique, those being: *i)* The covered business sector – insurance – lacks well-documented examples concerning the

implementation of CLV in comparison to others, such as retail or wholesale. Because CLV implementations vary according to each sector's characteristics, it is important to have a diversified range of examples across several types of businesses. This particular one seeks to enrich the knowledge base around CLV in insurance; *ii)* The business granularity achieved in this project was one of its most differentiating factors. Previous implementations only took into consideration one level of granularity by focusing on calculating CLV for one line of business, independently of the business sector they were applied to. In this project, CLV was computed considering 2 granularity levels - Globally and by LoB – and encompassed 7 distinct business areas (one company-wise plus one for each considered LoB), without disregarding each one's characteristics. This allowed the execution of a CLV project with a substantial scale that few other implementations achieved up to this date; *iii)* Integration of advanced Predictive models to better align propensities of cross-sell, upsell, risk or churn with reality of each customer was also a relevant factor which highly distinguished this particular implementation. In most previous applications, these components were simply estimated by looking at past events and joining one or two highly explanatory variables, without exploring advanced data mining and pattern discovery techniques.

# 6. Limitations & Future Improvements

In this section, project limitations and future improvements will be addressed. The first group will denote which aspects constrained the project implementation and possibly affected the end results. Subsequently, the second group will mention which improvements should be applied in future implementations in order to try to achieve better results as well as a model more aligned with the business reality.

## 6.1 Limitations

The fact that this project was developed in a business environment originated several limitations which were common to other projects made in these circumstances. Defined deadlines for every task and dependency on stakeholder's knowledge and availability were two great setbacks. Considering this CLV model had to cover almost every line of business, a 9-month duration was short to plan, execute and analyze everything. Ultimately, in order to avoid delays, some tasks were simplified (e.g., simple estimations to cover future components of CLV) and some decisions were taken which did not completely reflect the business processes.

Another limitation of this project was related with the fact the current customer value only considered historical data from the last 3 years when the ideal scenario would be to consider all main sources of revenues and costs since each customer entered the company. This situation was justified by data availability problems which arose during the development of this project. By only considering this 3-year window, good customers could be harmed by an unfortunate event with a large claim, which could greatly decrease a client's CLV, even if he/she had no major claims registered in older periods of its lifetime.

## 6.2 Future Improvements

In future versions of this project, there are indeed improvements to be applied. Naturally, the first set of improvements to consider is to work on previously identified limitations, either to erase them or to reduce its impact.

Secondly, it is necessary to integrate more advanced models, capable of better explain complex phenomena in ways which the adopted simple estimations cannot. Furthermore, the adopted data discretization process should be reviewed. Currently, the data universe with positive values is being classified according to the quartile distribution, but perhaps instead of dividing this subset into approximately equal parts of 25%, each rank could have distinct proportions assigned (e.g., 10%-40%-40%-10% for rank 2 to 5, respectively), or even test completely new rationales.

Additionally, another improvement to take into consideration is the inclusion of further cost sources, not only to complement the present value of a customer (CCV) but also its future value. By doing so, the cost component would be more aligned with business reality.

Finally, a component which could be interesting to analyze and possibly add to the current framework would be *down-sell*. This customer interaction was ignored in almost every reviewed

CLV approach, however, it is known that in some business sectors, such as insurance, product/service downgrades are sometimes preferred to avoid profitable customers to leave. This way the set of potential interactions a customer could have with the company, would be more complete.

# 7. Bibliography


1. Abdolvand, N., Albadvi, A., & Koosha, H. (January, 2014). Customer Lifetime Value: Literature Scoping Map, and an Agenda for Future Research. *International Journal of Management Perspective, 1*(3), 41-59.

2. Alexandre, A. M. (2009). *Customer Lifetime Value na banca.*

3. Anderson, J. F., & Brown, R. L. (2005). Risk and Insurance. Education and Examination Comitee of the Society of Actuaries.

4. Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: MArketing models and applications. *Journal of Interactive Marketing*.

5. Blattberg, R., & Deighton, J. (1991). Interactive Marketing: exploiting the age of addressability. *Sloan Management Review, 33*, 17-30.

6. Borle, S., Singh, S. S., & Jain, D. C. (2008). Customer Lifetime Value Measurement. *Management Science* , 100-112.

7. Chatley, P. (2014). *Effectiveness of bancassurance as a channel of selling life and non-life insurance products.* Retrieved from http://hdl.handle.net/10603/57443

8. D One. (2013). Measuring Customer Lifetime Value.

9. Donkers, B., Verhoef, P., & Jong, M. (2007). Modeling CLV: A test of competing models in the insurance industry. *Quantitative Marketing & Economics*, 163-190.

10. dun & bradstreet. (15th of January, 2016). *How to cross-sell and upsell*. Retrieved in 30 of July, 2018, from https://www.dnb.com/perspectives/marketing-sales/how-to-cross-sell-up-sell.html

11. Fader, P., Hardie, B. G., & Ka, L. L. (2008). "Counting Your Customers" the Easy way: An Alternative to the Pareto/NBD Model. *Marketing Science*, 275-284.

12. Gupta, S., Hanssens, D., Hardie, B., & Kahn, W. (2006). Modeling Customer Lifetime Value . *Journal of Service Research*, 139-155.

13. Gupta, S., Lehmann, D., & Stuart, J. (2004). Valuing Customers. *Journal of Marketing Research* , 7-18.

14. Harju, T. (2015). *Accuracy of noncomplex customer lifetime value models in the medical service context - the case of a telemedicine service provider.*

15. Hoekstra, J. C., & Huizingh, E. R. (1999). The Lifetime Value Concept in Customer Based Marketing. *Journal of Market Focused Management*, 257-274.

16. Holm, M., Kumar, V., & Rohde, C. (2012). Measuring custimer profitability in complex enviroments: An interdisciplinary contingency framework. *Journal of the Academy of Markerting Science*.

17. Jain, D., & Singh, S. (2002). Customer lifetime value research in marketing: A review and future directions. *Journal of Interactive Marketing, 16*, 34-46.

18. Kamakura, W. A. (2007). *Cross-selling: Offering the right product to the right customer at the right time.*

19. Kiss metrics. (s.d.). *Kiss Metrics blog*. Retrived in 3 of August, 2018, of https://neilpatel.com/blog/how-to-calculate-lifetime-value/

20. Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T., & Tilmanns, S. (2010). Undervalued or Overvalued Customers: Capturing total Customer Engagement Value. *Journal of Service Research*, 297-310.

21. Kurgan, L., & Cios, K. J. (2001). Discretization algorithm that uses class attribute interdependence maximization. *Internation Conference on Atificial Intelligence*, (pp. 980-987). Las Vegas.

22. Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining Knowledge Discovery*, 393-423.

23. Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: an empirical investingation and implications for marketing. *Journal of Marketing*, 17-35.

24. Reinartz, W. J., & Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing*, 77-99.

25. Renard, P., Alcolea, A., & Ginsbourger, D. (2013). Stochastic versus Deterministic Approaches. Em J. Wainwright, & M. Mulligan, *Enviromental Modelling: Finding Simplicity in Complexity* (pp. 133-149). John Willey & Sons.

26. Rodne, L. (2009). *Introduction to Insurance.*

27. Romero, J., van der Lans, R., & Wierenga, B. (2013). A Partially Hidden Markov Model of Customer Dynamics for CLV Measurement. *Journal of Interactive Marketing*, 185-208.

28. Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on Marketing: Using Customer Equity to Focus Marketing Strategy. *Journal of Marketing*, 109-127.

29. SAS. (s.d.). *Proc Rank technical documentation*. Obtido em 24 of July of 2018, from http://support.sas.com/documentation/cdl/en/proc/61895/HTML/default/viewer.htm#rank-overview.htm

30. Seyerle, M. (2001). *Customer lifetime value and its determination using the SAS Entereprise Miner and the SAS-OROS software.* Retrieved in 6 of August, 2018, from http://www.sascommunity.org/seugi/SEUGI2003/SEYERLE_LifetimeValue.pdf

31. Statsbot. (February of 2018). Calculating Customer Lifetime Value: SQL example.

32. Sweet tooth. (2015). *Everything you need to know about Customer Lifetime Value.*

33. Towers Watson. (March of 2015). Customer Lifetime Value - Opportunities and Challenges.

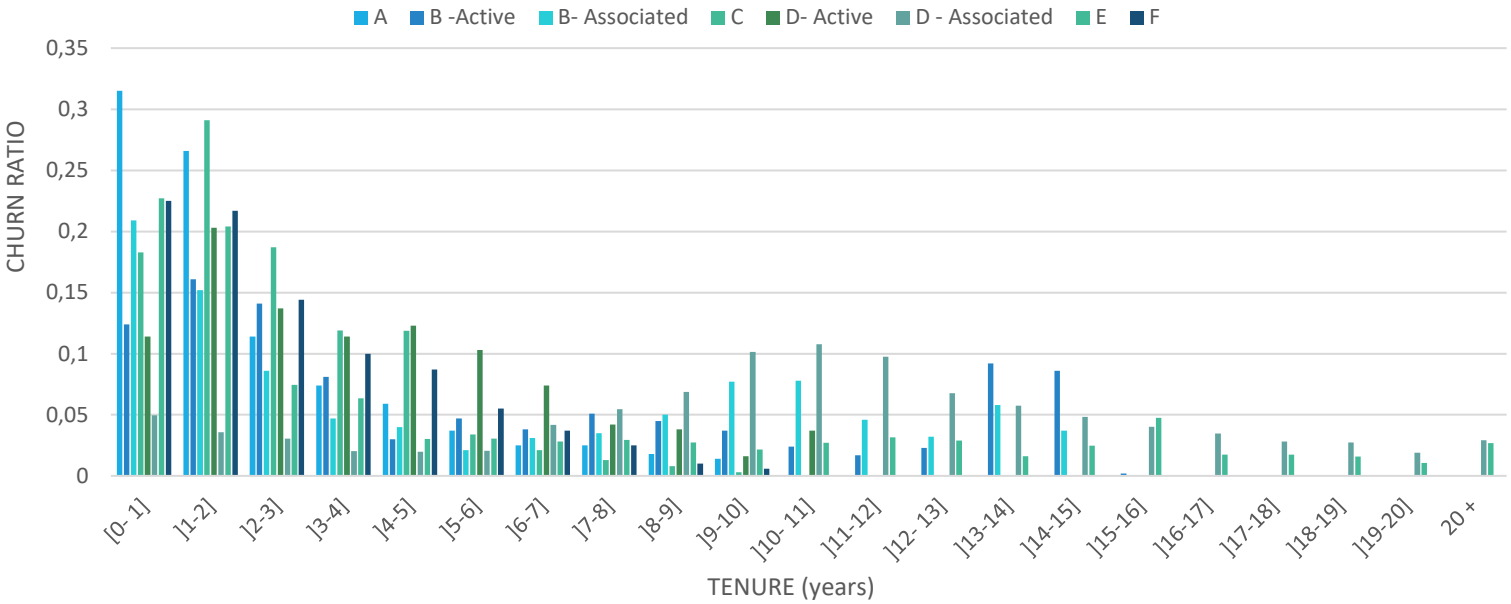
34. Venkatesan, R., & Kumar, V. (2004). A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. *Journal of Marketing*.

# 8. Annexes

*A 1 – CLV studies summary presented by Tuomas Harju (Harju, 2015)*

| Study | Context 1 | Context 2 | Measurement technique | Level of aggregation |
|---|---|---|---|---|
| Dwyer 1989 | Noncontractual | Always-a-share | Stochastic | Company |
| Blattberg and Deighton 1996 | Not applicaple | Lost-for-good | Deterministic | Company |
| Berger and Nasr 1998 | Not applicaple | Both | Deterministic | Company |
| Pfeifer and Carraway 2000 | Noncontractual | Always-a-share | Stochastic | Company |
| Rust et al. 2004 | Noncontractual | Always-a-share | Stochastic | Individual |
| Fader et al. 2005 | Noncontractual | Lost-for-good | Stochastic | Company and individual |
| Lewis 2005 | Contractual | Always-a-share | Stochastic | Individual |
| Reinartz et al. 2005 | Noncontractual | Lost-for-good | Stochastic | Individual |
| Haenlein et al. 2006 | Noncontractual | Always-a-share | Stochastic | Individual |
| Kumar et al. 2006 | Noncontractual | Always-a-share | Stochastic | Individual |
| Haenlein et al. 2007 | Noncontractual | Always-a-share | Stochastic | Segment |
| Venkatesan et al. 2007 | Noncontractual | Always-a-share | Stochastic | Individual |
| Borle et al. 2008 | Contractual (membership) | Lost-for-good | Stochastic | Individual |
| Kumar et al. 2008 | Noncontractual | Always-a-share | Stochastic | Individual |
| Ryals 2008 | Contractual | Not applicable | Deterministic | Individual |
| Homburg et al. 2009 | Noncontractual | Always-a-share | Stochastic | Segment |
| Jen et al. 2009 | Noncontractual | Always-a-share | Stochastic | Individual |
| Kumar et al. 2010 | Noncontractual | Always-a-share | Stochastic | Individual |
| Braun et al. 2011 | Contractual | Lost-for-good | Stochastic | Individual |
| Schweidel et al. 2011 | Contractual | Always-a-share | Stochastic | Individual |
| Rust et al. 2011 | Noncontractual | Always-a-share | Stochastic | Individual |
| Ascarza and Hardie 2013 | Contractual (membership) | Always-a-share | Stochastic | Individual |
| Romero et al. 2013 | Noncontractual | Always-a-share | Stochastic | Individual |
| Schweidel and Knox 2013 | Noncontractual | Always-a-share | Stochastic | Individual |
| Esteban-Bravo et al. 2014 | Noncontractual | Always-a-share | Stochastic | Individual |
| Ekinci et al. 2014 [1] | Noncontractual | Always-a-share | Stochastic | Individual |
| Jahromi et al. 2014 | Noncontractual | Always-a-share | Stochastic | Individual |

*A 2 – Last year's churn Behavior by tenure (in years) across considered LoB's and sales types*




*A 3 - Confusion matrix proportion results by LoB*

| CLV RANK (Predicted Value) | CCV RANK (Observed Value) 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0,74 | 0,10 | 0,06 | 0,06 | 0,04 |
| 2 | 0,12 | 0,70 | 0,35 | 0,05 | 0,01 |
| 3 | 0,09 | 0,11 | 0,45 | 0,33 | 0,03 |
| 4 | 0,09 | 0,06 | 0,09 | 0,52 | 0,25 |
| 5 | 0,11 | 0,06 | 0,04 | 0,08 | 0,70 |

*I - LoB A confusion matrix proportions*

| CLV RANK (Predicted Value) | CCV RANK (Observed Value) 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0,21 | 0,72 | 0,05 | 0,01 | 0,00 |
| 2 | 0,01 | 0,48 | 0,37 | 0,02 | 0,00 |
| 3 | 0,07 | 0,04 | 0,57 | 0,32 | 0,01 |
| 4 | 0,05 | 0,02 | 0,06 | 0,67 | 0,20 |
| 5 | 0,04 | 0,03 | 0,02 | 0,06 | 0,86 |

*II - LoB B confusion matrix proportions*

| CLV RANK (Predicted Value) | CCV RANK (Observed Value) | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0,61 | 0,15 | 0,06 | 0,09 | 0,08 |
| 2 | 0,06 | 0,38 | 0,45 | 0,09 | 0,02 |
| 3 | 0,05 | 0,13 | 0,39 | 0,36 | 0,08 |
| 4 | 0,05 | 0,11 | 0,13 | 0,47 | 0,23 |
| 5 | 0,05 | 0,12 | 0,05 | 0,11 | 0,67 |

*III – LoB C confusion matrix proportions*

| CLV RANK (Predicted Value) | CCV RANK (Observed Value) | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0,68 | 0,10 | 0,07 | 0,08 | 0,08 |
| 2 | 0,03 | 0,61 | 0,34 | 0,02 | 0,00 |
| 3 | 0,02 | 0,06 | 0,63 | 0,27 | 0,01 |
| 4 | 0,02 | 0,07 | 0,03 | 0,72 | 0,16 |
| 5 | 0,02 | 0,08 | 0,03 | 0,03 | 0,84 |

*IV - LoB D confusion matrix proportions*

| CLV RANK (Predicted Value) | CCV RANK (Observed Value) | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0,60 | 0,21 | 0,08 | 0,06 | 0,11 |
| 2 | 0,00 | 0,70 | 0,29 | 0,01 | 0,00 |
| 3 | 0,00 | 0,08 | 0,68 | 0,23 | 0,01 |
| 4 | 0,00 | 0,05 | 0,09 | 0,70 | 0,16 |
| 5 | 0,00 | 0,07 | 0,03 | 0,06 | 0,84 |

*V - LoB E confusion matrix proportions*

| CLV RANK (Predicted Value) | CCV RANK (Observed Value) | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0,37 | 0,55 | 0,02 | 0,04 | 0,02 |
| 2 | 0,10 | 0,64 | 0,23 | 0,02 | 0,02 |
| 3 | 0,14 | 0,05 | 0,51 | 0,26 | 0,04 |
| 4 | 0,15 | 0,03 | 0,08 | 0,54 | 0,20 |
| 5 | 0,22 | 0,04 | 0,04 | 0,09 | 0,62 |

*VI - LoB F confusion matrix proportions*