University of Padova

Department of Mathematics

Master Thesis in Data Science - Big Data Management and Analytics

Customer Base Segmentation and a Baseline Approach to Mixed Effects Survival Forest

Supervisor
Professor Bruno Scarpa
University of Padova

Co-supervisor

Master Candidate
Pietro Ferrazzi

Academic Year

2022-2023

Abstract

This thesis is structured into two distinct parts. The first revolves around proposing a data driven approach for segmenting a customer base. The organization that contributed the data currently employs an expert based segmentation strategy they want to substitute with a quantitative approach. Here we implemented a methodology that combines principal component analysis with survival models. A critical prerequisite was the utilization of a two-dimensional framework, encompassing the dimensions of customers' economic value and their likelihood to make repurchases (propensity). We formulated two separate indexes, each corresponding to one of these dimensions, employing diverse analytical techniques to determine optimal approaches in terms of performance. The outcome of this segmentation effort furnishes the business with a valuable tool, ultimately aimed at enhancing customer relations and facilitating business growth.

The second part establishes the foundation for extending mixed effects models into the realm of survival forests. At the best of our knowledge, this problem has not been addressed yet in the literature. The objective here was to expand the methodology employed in mixed effects Cox models for handling random effects, and adapt it to the context of survival forests, exploring its potential applicability to the segmentation problem under consideration. Two different approaches were experimented with. The former approach, which ultimately yielded the final outcome, involves the integration of Martingale residuals. The achievement of this endeavor establishes a baseline from which future researchers can delve into more sophisticated and efficient implementations.

Contents

Abstract v

List of figures ix

List of tables xiii

Listing of acronyms xv

1 Introduction 1
1.1 Customer Base Segmentation 2

2 Customer Base Segmentation Definition 5
2.1 Introduction 5
2.2 Data description 5
2.2.1 Source Data Aggregation 6
2.2.2 Segmentation-required data split 7
2.3 Customer Life Time Value (CLTV) 8
2.3.1 CLTV calculation: the implemented solution 9
2.4 Segmentation Approach 10

3 Value and Propensity Indexes 13
3.1 Value 13
3.1.1 PCA as Value Index generator 14
3.2 Propensity 18
3.2.1 Survival Models 19
3.2.1.1 Survival Function 19
3.2.1.2 Hazard Function 20
3.2.1.3 Non parametric estimators 20
3.2.1.4 Cox Model 20
3.2.1.5 Survival Forest 21
3.2.1.6 Evaluation Metrics 23
3.2.2 Preprocessing for Survival Models training 25
3.2.3 Survival Models Training and Selection 28
3.2.3.1 Survival EDA 28
3.2.3.2 Cox Model Variables Selection 31

3.2.3.3 Selected Cox Model 32
3.2.3.4 Survival Forest Training and Selection 34
3.2.3.5 SRF versus Cox Model 38
3.2.4 Indexes Calculation 38
3.2.5 Index Selection 40

4 Segmentation Results 47
4.1 Segmentation based on Value and Propensity Indexes 47
4.1.1 Cutting planes 48
4.1.2 Segmentation 49
4.1.3 CLTV model integration 51

5 Mixed Effect Survival Forest 55
5.1 Introduction to Mixed Effects 55
5.2 Mixed Effects Models Overview 56
5.3 Cox Mixed Effects 56
5.3.1 Cox Mixed Effects Model Estimation 57
5.4 Mixed Effects Survival Forest 57
5.4.1 Model Specification 57
5.4.2 Model Estimation 59
5.4.2.1 Expanded-Baseline Iterative Approach 60
5.4.2.2 Survival Residuals 64
5.4.2.3 Adding the Martingale residuals information 66
5.4.2.4 Mixed effect Survival Forest: Case Study 70

6 Conclusion 73
6.1 Customer Base Segmentation 73
6.2 Mixed Effects Survival Forest 74

7 Appendix 75
7.1 R code for mixed effect survival forest 75
7.1.1 Functions 75
7.1.2 Fitting 81
7.2 Tables 83

References 85

Acknowledgments 91

Listing of figures

2.1 **Top graph**: datasets as initially provided by the company **Bottom graph**: datasets as aggregated for the analytical purposes 6

2.2 Representation of the CLTV calculation pipeline 10

2.3 Example of ideal segmentation for the customer base of a generic company. Four groups are identified based on the Value and Propensity indexes. Meaningful names are assigned to them to add business sematic to the segmentation. 11

3.1 Correlation matrix between the variables selected for the calculation of the Value Index. They are all higly positively correlated excluded the number of purchases per customer. This suggests that: A) high levels products are usually sold for higher prices, B) the average value of the purchases is mainly determined by the first one, C) the number of purchases is not correlated to the level and the amount of the purchases. 15

3.2 The observed value of the client at the end of the first five year period of observation against the scores of the PCA interpreted as Value Index. The **left plot** represents all the customers, while the **right plot** represents a random subset of 5000 customers for a clearer visualization. The red line represents the linear relation between the two features. The PCA scores are significant in explaining the real value with a $pvalue < 2 \cdot 10^{-16}$ for the test against the null model. It is worth noting that a linear relationship is evident in the lowest levels of the real value after the first 5 years. This occurrence is because the value is determined by the cumulative sum of all purchases, including the initial one. Since the scores are influenced positively by the first purchase amount, there cannot be observations with high scores and an overall sum of purchases below a certain threshold. 17

3.3 Survival curves for the groups defined by the levels of 11 variables. See the Appendix for more details about the variables. For each plot the *p value* of the log-rank test is printed on the bottom-left for the hypothesis of all the curves being equal. The Bonferroni-adjusted threshold to maintain a confidence interval of 95% is 0.0045. For this reason, AVG_AMOUNT_PURCH is considered as not significant in this regards, since $0.0055 > 0.0045$. 30

3.4 Test to assess whether the coefficients β of the Cox model are time-independent. If this test confirms the hypothesis, it verifies the assumption of proportional hazards. Visually, if the hypothesis holds true, the plot should exhibit a horizontal line. The only instances where the test does not reject the hypothesis of independence are for the first (*tasso interaz last 3 Y dopo 5y*) and seventh (*GG DISTANZA INTERAZ PRIMO ACQ*) cases, with p-values of 0.21 and 0.19 respectively. However, the global test indicates that the assumption of proportional hazards is rejected. 34

3.5 **Top left**: One's complement of a survival curve for two customer with different values of the explicative variables **Top right**: propensity index as area under the curve **Bottom left**: propensity index as cumulative probability of repurchasing after a time threshold = 2000 days **Bottom right**: propensity index as time after which the probability of repurchasing is equal to a probability threshold = 30% 42

3.6 Performances of the three methods proposed to summarize the information contained in the survival curve. The x axis represent the value of the index calculated for the given method. The best method is the one that better divides the grey observations from the red ones. The one that seems to do that is the one based on the probability threshold. Note that for the first two a value must be assigned to the threshold; here we used prob.threshold = 0.3 and a time.threshold=2000. The right bar of the probability threshold distribution represent units that never reach a cumulative probability of 0.3. 43

3.7 Distribution of the propensity index calculated for observations in the test set, and how it varies with different values of the probability threshold. As the threshold increases, the number of observations that do not reach that level of the probability of repurchasing increases significantly, represented by bars with values equal to 9999. We observe a transition from one extreme (top-left plot) where the observations are mainly distributed over low values, to the other extreme (bottom-right) where almost all observations are located at high values. The scope is to identify the best threshold in terms of separation between red and grey observations. 44

3.8 Distribution of the propensity index calculated for observations in the test set, and how it varies with different values of the probability threshold. It can be noted that for any threshold there is an overlapping area in which it is not possible to identify clients that did a repurchase from the ones that did not. The scope is to identify the best threshold in terms of separation between red and grey observations. 45

4.1 **Top left**: distribution of the clients in DF10y given the two indexes. **Top right**: distribution of the clients in DF10y given the two indexes and the observed future revenues each client will provide to the company classified as high or low. **Bottom left**: cutting planes calculated by the classification tree having as target the classes identified by the observed future revenues as 'low' and 'high'. **Bottom right**: distribution of the clients in DF10y given the two indexes and the predicted future revenues as 'low' and 'high' 51
4.2 Segmentation of the Customer Base into 6 groups. The separation of the space is based on the cuts provided by the classification tree integrated by a semantic-based reasoning. 52
4.3 Different frameworks for CLTV prediction. **Framework 1** shows the one currently implemented that does not consider the data driven segmentation. **Framework 2** uses the data driven segmentation as a feature in the model that predicts the number of purchases. **Framework 3** as the previous, but also substitutes the expert based segmentation with the data driven one. **Framework 4** uses the segmentation only as a substitute to the expert based one. 54

5.1 Example of the impact of the random effect on the hazard function. The black line represents a baseline hazard function. If we add a random effected which impact is estimated as 1.5, the consequential hazard function is the red one. . 58
5.2 **Output 1**: baseline expansion of a row of the original data through a discrete hazard function $h_0(t)$. **Output 2**: baseline expansion of a row of the original data through a discrete hazard function $h_0(t)$ and the fixed effects of the Cox model $\exp(x_i\beta)$. 62
5.3 **Graph 1**: training dataset as expansion of a row of the original data through the Martingale residuals for all the possible time stamps. **Graph 2**: training dataset as the original one with a new column representing the new censoring information calculated trough the Martingale residuals. 69
5.4 Convergence criterion. The mixed effect survival forest is considered to converge when all the elements of the random effect modelling parameter $\hat{b}$ do not increase or decrease of more than 10% in one iteration. The plot shows how the convergence is not reached in the fixed maximum number of iterations, since none of the iterations provided a convergency measure lower than the threshold. 72

Listing of tables

2.1 Description of the varibles in the two datasets. DF2022 contains all the variables, while DF2017 only some of them 8

3.1 The table shows the variance explained by the PCA components and how they are impacted by the 4 variables. First 2 rows: variance explained by each component. The first component explain aroun 65% of the original variance, the first 2 together almost the 90%. Last 4 rows: PCA loadings. Each row shows how much each component is determined by a certain variable. E.g. the first component is determined almost equally by *LEVEL DS FIRST PURCHASE*, *NET AMOUNT FIRST SALE ACTUAL* and *avg value purchases actual*. The second one is determined by *N PURCHASES CUSTOMER* . . . 16

3.2 Evaluation metrics for survival models. Cox Partial-Likelihood Score (Cox PLS), Concordance index (C-index), Brier Score (BS), Integrated Brier Score (IBS), Integrated Absolute Error (IAE), Integrated Square Error (ISE), Mean Absolute Error (MAE). For a comprehensive overview see [1] 24

3.3 Summary of splitting criteria. For a comprehensive overview see [2] and [1]. . 26

3.4 Preprocessing pipeline. 27

3.5 Distribution of clients given the number of purchases they did. The table includes only customers from the DF10y dataset since this is the dataset on which the analysis will be performed. Around 65% of them never did a second purchase, 27% did exactly 2 purchases and around 7% did more than 2. 27

3.6 One-step-forward approach for Cox model variables' interaction selection. At each step, the interaction between a pair of variables is given as feature to the model and the performances on the validation set are calculated. If performances increase, the interaction term is kept in the next steps. 33

3.7 One-step-forward approach for Cox model variables' quadratic terms selection. At each step, the squared variable is given as feature to the model and the performances on the validation set are calculated. If performances increase, the interaction term is kept in the next steps. 33

3.8 Variables selected in the final Cox model. For continuous variables, the product of $exp(\beta)$ and the observed value of the relative variable can be interpreted as a multiplication factor effect on the baseline hazard function. The same interpretation holds for qualitative variables when the feature itself is replaced by an indicator function equal to 1 when the level of the respective variable is observed. For instance, consider a client i with *aging client-customer relationship* level = 'HOT' and *int rate last 3 y* = 0.3. The estimated hazard function for this client is given by $h_i(t) = h_0(t) \cdot \exp\{1 \cdot 0.810\} \cdot \exp\{0.3 \cdot 1.08\} \cdot \exp\{1 \cdot 0.3 \cdot 1.199\}$. The three terms represent the effects related to *aging client-customer relationship*, *int rate last 3 y*, and their interaction, respectively. 35

3.9 Validation set C-index calculated on the prediction given by different values of the number of trees and minimum nodes size. The pair of parameters that results in the best metric is number of trees = 150 and minimum node size = 40. The relative cell is highlighted in green. 36

3.10 Survival Random Forest hyperparameters. The *number of trees* and *mtry* were selected through a fine-tuning procedure, while the others were chosen based on literature recommendations. 37

3.11 Comparison of different evaluation metrics between the selected Cox model and Survival Forest. In almost all the cases, the Survival Forest performs better than the Cox model. Note that the Integrated Brier Score (IBS) has been approximated by the selection of a random subset of observations to reduce its high computational cost. 38

3.12 Possible choices to calculate the propensity index as a summary of the information contained in 1 - the survival curve. The selected one is the probability threshold that provides a straight forward interpretation and properly divides the observations that actually did a repurchase from the ones that did not. . . 40

4.1 Classification Models to identify groups of customers similar in terms of overall revenues 50

5.1 Training computational cost 71

5.2 Environment specifications 71

7.1 Description of the varibles in the two datasets. DF2022 contains all the variables, while DF2017 only some of them 84

Listing of acronyms

1
Introduction

The hereby presented work takes place within the context of the consultancy service provided by Alkemy to a company operating in the hearing aids market. Specifically, the company specializes in selling devices designed to assist individuals who have experienced hearing loss. The main driver of the business revolves around a single type of product positioned at a high price point, typically ranging in the thousands of euros. The clients are people that have a medical need and are looking for a solution to it. The customer base is composed by clients with an high average age and, despite offering a limited range of products, the company has recognized that customers exhibit varying behaviors concerning repurchases, responsiveness to marketing campaigns, and their willingness to invest in the products. This diversity in customer behavior suggests the need for segmentation to understand the preferences and needs of different customers' groups, providing them with tailored experiences that improve customer satisfaction, increase the productivity and strengthen the company's position in the market. For example, when a marketing campaign is launched, the segmentation can be the driver to assess whether it should be targeting all the customers or any specific group, based on the expected outcomes.

The company which data I analyze in this thesis wanted to implement a segmentation of the Customer Base for a specific country they operate in. The aim is to group the clients on the basis of two relevant dimensions they expected to extrapolate from the data: the economical value of clients and their propensity to repurchase the company's products. These dimensions were selected due to their relevance in the broader context of predicting Customer Lifetime Value (CLTV), an indicator of how much a customer is expected to generate during their relationship with the company over time calculated on the first five years of collected information. Consequently, to align with this purpose, the usable data for the segmentation was restricted to the first five years from first purchase of each customer. Furthermore, the scope of such a segmentation goes behind the grouping itself, addressing the need for a new reliable predictor to be considered by the CLTV predictive model.

Up until now, the segmentation process relied on experts knowledge, dividing clients into groups based on their age range, time since last purchase, and the performance level of their first device, indicative of their hearing aid quality. The assumption was that these three features could be informative enough for the identified groups to be properly grouping the customers. However, they sought to improve this approach by transitioning to a data-driven solution. The aim of this project is to provide a robust segmentation approach to divide the customers into groups based on the two identified dimensions, constructed on the first five years of information of each client. This methodology has to be an evidence-based, reproducible procedure that is expected to prove more beneficial than the expert-based approach.

1.1 Customer Base Segmentation

A classical approach to customer segmentation involves identifying major areas of interest that can effectively divide a company's customer base into valuable groups, regardless of their interactions with the company itself. This segmentation aims to group customers based on specific characteristics or attributes without considering their past behavior or engagement with the company's products or services. Several methods have been proposed, as discussed in [3]. Geographic segmentation involves dividing customers based on their location, considering different granularities such as states, regions, cities, or neighborhoods. Demographic segmentation, on the other hand, relies on demographic variables such as age, family size, gender, income, social class, or generation, assuming that consumer preferences and desires are often associated with these demographic factors. Psycographic segmentation focuses on understanding customer lifestyles, personalities, and values, as these elements can play a significant role in shaping their buying behavior and preferences [4]. Finally, behavioral segmentation involves classifying customers based on their knowledge of, attitude toward, usage of, or response to a particular product or service, reflecting their behaviors and interactions in relation to the offering.

By employing these segmentation approaches, companies can gain insights into different customer groups and tailor their marketing strategies and product offerings to better meet the specific needs and preferences of each segment.

On the other hand, the relationship that a company has with its customer base is crucial for the business. Every company aims to proactively enhance clients' retention and loyalty, increase their profitability, create value for them customizing products and services [5]. In order to assess these results, many companies necessitate to measure their customers' value [6], identifying common characteristics among them. One widely adopted value measure is the Customer Lifetime Value (CLTV) [7], which quantifies the overall worth of a customer to the company over their entire customer lifetime. Employing CLTV, businesses can conduct customer segmentation based on the customers' individual value or integrate it with additional information. A standard technique involves grouping customers according to percentiles of the CLTV values distribution, allowing for clear delineation of high-value, medium-value, and low-value customer segments. Extending this approach, [8] proposed a more sophisticated segmentation based on three dimensions extrapolated from CLTV: the current value, the potential value, and customer loyalty.

Furthermore, by incorporating managerial information from other sources, companies can augment their segmentation strategies beyond the dimensions used in the CLTV-based approach. These additional features could encompass more classical segmentation criteria such as geographic, demographic, psychographic, or behavioral attributes. By amalgamating the diverse information available, businesses can create richer customer segments and fine-tune their marketing strategies to address the specific needs and desires of each segment more effectively.

When conducting customer base segmentation, there are several key characteristics [3] [9] [10] that a segmentation should ideally possess to be considered valuable and effective:

- Substantial: The segments should be of sufficient size and importance to be economically viable for the company. A segment with a small number of customers might not significantly impact the company's overall business performance.

- Profitable: The segments should identify customers who are profitable to the company. Identifying high-value segments that generate substantial revenue or have the potential to do so is crucial.

- Reachable: The customers within each segment should be reachable and accessible through various marketing channels. If a segment is difficult to target or communicate with, it might not be practical for marketing efforts.

- Differentiable: The segments should be conceptually distinguishable from each other in terms of their behaviors, preferences, and responses to marketing strategies. This allows the company to tailor specific marketing approaches to each segment.

- Actionable: The segmentation should provide actionable insights that enable the company to implement effective marketing programs to attract and retain customers within each segment. If the segmentation lacks practical applications, it might not yield tangible results.

- Stable: The segments should be relatively stable over time, meaning that customers' behaviors and characteristics within each segment do not fluctuate dramatically. A stable segmentation allows the company to plan and execute long-term marketing strategies.

- Measurable: The characteristics of each segment should be quantifiable and measurable, allowing the company to track and evaluate the performance of each segment over time.

- Valid: The segmentation should be based on sound analytical techniques and validated through rigorous testing to ensure its accuracy and reliability.

By considering these characteristics, a company can assess a customer base segmentation approach, improve it and further extend it.

In this context, the primary objective of this thesis is to create a segmentation framework that possesses all the desired properties mentioned earlier. Additionally, the framework aims to seamlessly integrate the dimensions of Customer Lifetime Value (CLTV) to further enhance the segmentation process. By achieving these goals, the thesis seeks to provide the company with a powerful tool to effectively understand and target different customer groups, ultimately contributing to improved the CLTV calculation.

2

Customer Base Segmentation Definition

2.1 Introduction

The aim of the project is to identify groups of clients to provide a data-driven segmentation of the customers base. There are two final objectives of such a segmentation: its integration into the Customer Life-Time Value's estimation pipeline (Section 2.3) and an explainable separation of the clients based on indicators that are valuable from a business and marketing prospective.

2.2 Data description

The statistical units contained in the database of interest are all the people that have performed at least one purchase. Since the format of the data collection has changed through the years, the available information can differ based on the date their first sale was registered in the system. In order to homogenise the procedures, only data between 2012 and 2022 has been extracted from the relational data warehouse by mean of an SQL query [*].

[*] As the data was forwarded me into an .xlsx format, the extraction process falls outside the scope of this project.

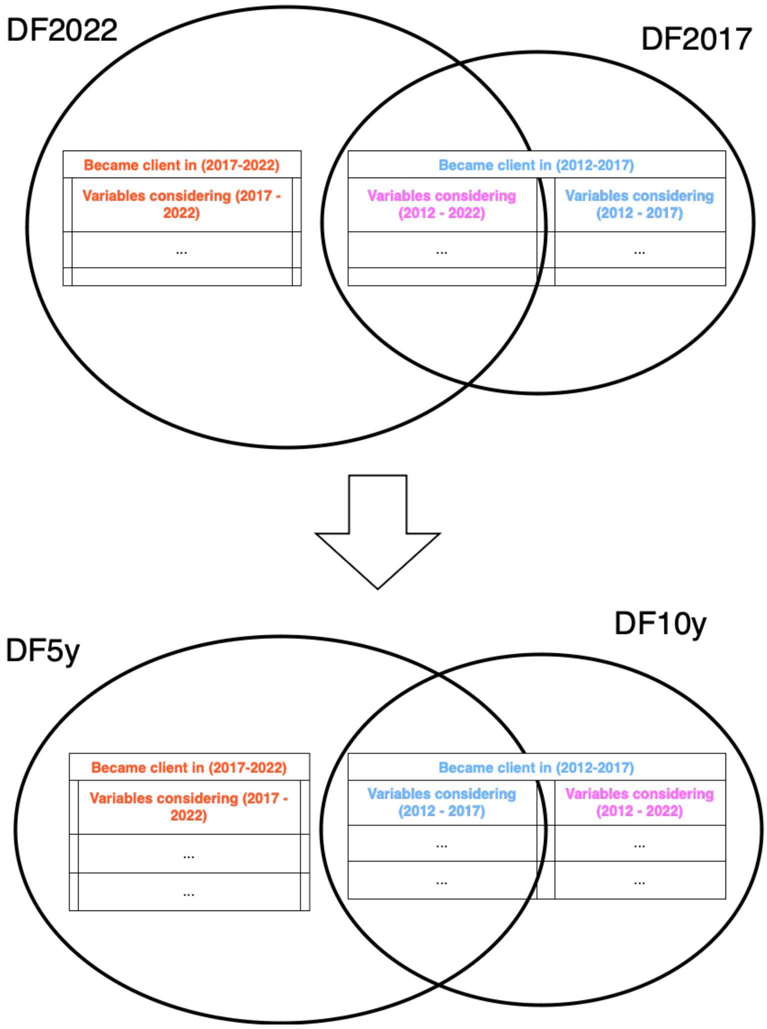

DF2022
DF2017
Became client in (2017-2022)
Variables considering (2017 - 2022)
...
Became client in (2012-2017)
Variables considering (2012 - 2022)
Variables considering (2012 - 2017)
...
...
DF5y
DF10y
Became client in (2017-2022)
Variables considering (2017 - 2022)
...
...
Became client in (2012-2017)
Variables considering (2012 - 2017)
Variables considering (2012 - 2022)
...
...
...
...

Figure 2.1: Top graph: datasets as initially provided by the company
Bottom graph: datasets as aggregated for the analytical purposes

2.2.1 Source Data Aggregation

The source data was initially aggregated into two dataframes based on the date of first purchase: the first dataset comprises customers who became clients between 2012 and 2017 (referred to as DF2017 hereafter), while the second dataset includes those who became clients between 2012 and 2022 (referred to as DF2022). It's important to note that the former set of customers is encompassed within the latter, whilst the features contained in the former dataset are not a subset of the ones contained in the latter. A graphical representation is shown in the top graph of Figure 2.1

Some consequences of this structure are remarkable:

- at a row level (customers), all the clients that are in DF2017 are also included in DF2022
- at a column level, DF2017 has the information collected between 2012 and 2017, DF2022 the aggregate information collected between 2017 and 2022 or, when available, between 2012 and 2022
- DF2017 has 41.661 rows, DF2022 127.244: which means that $127.244 - 41.661 = 85.583$ people became new clients between 2017 and 2022.
- DF2017 contains less variables than DF2022. See Table 7.1 in Appendix for more details on the features.

2.2.2 Segmentation-required data split

The distinction between clients that have less than 5 years of history and those that have between 5 and 10 years of history is necessary for the analysis.

As mentioned in Chapter 2, the segmentation has to be built on the first 5 years of history of each customer since it has to be integrated as an explicative variable in the more general analytical framework of the Customer Life Time Value calculation described in Section 2.3. For this reason, two new datasets have been created:

- **DF10y** containing all the clients (41.661) with (5, 10] years of history and all the information available about them joining DF2017 and DF2022,
- **DF5y** containing all the clients (127.244) and the information about their first (0, 5] years of history. To achieve this, information about clients in DF2017 is dropped from DF2022 (i.e., the data about (5,10] years of history is forgotten). Then, the result is merged with DF2017.

A summary of them is provided in Table 2.1 together with the basic preprocessing. A graphical representation is shown in the bottom graph of Figure 2.1

DF	rows	cols	source	Information	Preprocessing
DF10y	41.661	35	DF2017	all available information of clients with more than 5 years of history	**a)**remove non informative columns; **b)**filter out clients with age first purchase > 100
DF5y	115.613	48	DF2017 + DF2022	information about the first (0,5] years of history of all the clients	**a)**remove non informative columns; **b)**filter out clients with age first purchase > 100; **c)**coalesce columns present in both original df, keeping the first-added info

Table 2.1: Description of the varibles in the two datasets. DF2022 contains all the variables, while DF2017 only some of them

2.3 Customer Life Time Value (CLTV)

The CLTV is an indicator that summarizes the revenue that a business performs and will perform over time in regard to a certain customer [11]. In general, this metric can be viewed as a mean to indicate how much a client is and will be valuable for the business, enabling a company to make a reasonable projection of the earnings a customer will generate during the average estimated duration of its relationship with the business itself. For this reason, it can be seen as a bridge between marketing and finance [12], leading the development of quantitative-based approaches for advertising strategies, selling policies, sales campaigns etc. It is important to differentiate between the intrinsic economic value of a client and the Customer Lifetime Value (CLTV) as distinct concepts. While a client may have great financial means and purchase a top-quality product at a high price, it does not necessarily imply that they are willing to spend their money again on the company's products in the future. CLTV takes into account the entire customer relationship and forecasts their potential future value, considering their past behavior and purchasing patterns.

2.3.1 CLTV calculation: the implemented solution

As previously mentioned, one of the two scopes of this thesis is to develop a segmentation that can serve as a potentially valuable feature in the existing CLTV calculation pipeline. This Section is dedicated to present the structure of the analytical framework already implemented by the company.

The scope of the predictive procedure they constructed is to determine the total value of a client, which is the sum of their observed economic value up to the present and a forecast of their future value based on their past behavior. To obtain this, the CLTV was calculated on the data under study based on a two-layers paradigm: people that entered the customer base between 2012 and 2017 (i.e., they did their first purchase in this time interval) were used to develop and train the model as the real value they provided in terms of revenues was already known for the 'future' time period 2017-2022. Then, the CLTV was calculated for all the clients that entered the customer base between 2017 and 2022 thanks to that model.

The framework, presented in Figure 2.2, is composed by two branches that converge to yield the final CLTV prediction. The first branch relies on a model-based approach, while the second leverages on business expertise. In the model based one, a two-step model provides an estimates of the number of purchases each customer is likely to make. This is achieved sequentially combining:

- a classification random forest (cRF) that aims to discriminate customers that repurchase the product at least once from the ones that do not,

- a regression random forest (rRF) that aims to estimate, for the ones that are predicted as re-purchasers by cRF, how many times they will repurchase.

In parallel, an expert based segmentation approach is employed to divide clients into groups based on their age range, time since last purchase, and the performance level of their first device, indicative of their hearing aid product quality.

Once the number of repurchases is calculated at a customer level, it is multiplied by the average sale amount observed in the first five years of data (i.e., without using any "future" information) for the corresponding segment previously individuated by the expert based segmentation. Then, the result is futher multiplied by an expert based probability representing the likelihood for a customer to remain active in the future, resulting in the final next-five-years prediction.

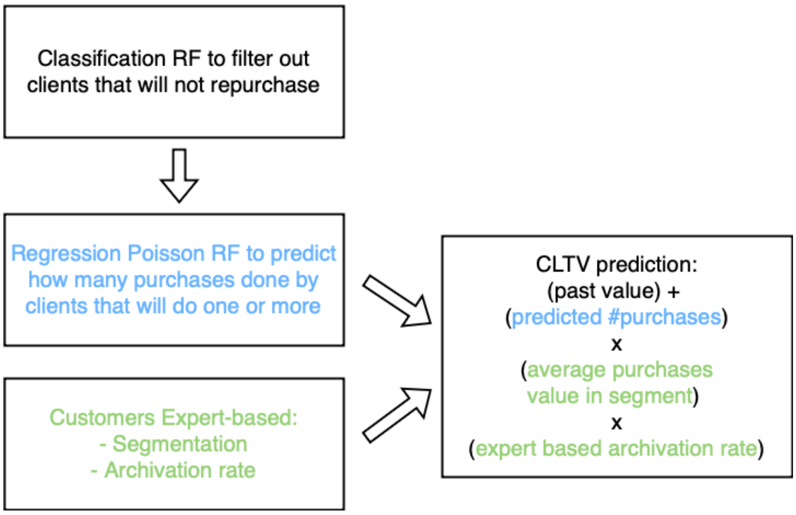

Classification RF to filter out clients that will not repurchase
Regression Poisson RF to predict how many purchases done by clients that will do one or more
Customers Expert-based:
- Segmentation
- Archivation rate
CLTV prediction:
(past value) +
(predicted #purchases)
x
(average purchases value in segment)
x
(expert based archivation rate)

Figure 2.2: Representation of the CLTV calculation pipeline

2.4 Segmentation Approach

The aim of the segmentation is A) to produce homogeneous groups of customers based on the observed variables that B) can be used as explicative variables in the CLTV calculation. Several ways of defining this *homogeneity* could be established. For this reason, it is important to clearly identify the business objective of the project, to make it lead the groups characterization. According to the company guidelines and as highlighted in the literature ([13], [14]), two main dimensions over which to perform the segmentation have been determined: the clients' economic value and their propensity for repeat purchases. The underlying motivation for both of these identified dimensions to be possibly helpful is that the Customer Life Time Value is a multiplication of the number of purchases by their amounts, i.e., the CLTV can be thought of as an combination of how much a client is willing to pay for the products and how reasonable it is to expect them to repurchase.

The ideal result of the segmentation would be a bi-dimensional space in which the clients are well distributed based on two customized variables (dimensions) that represent these characteristics. The example provided as a general guide line for the desired result is presented in Figure 2.3, where we can visualize the economical value and the propensity to repeat repurchase dimensions. It is important to consider that the ideal scenario is the one in which these two dimensions are independent. To achieve that, they should be built using different and possibly independent features.

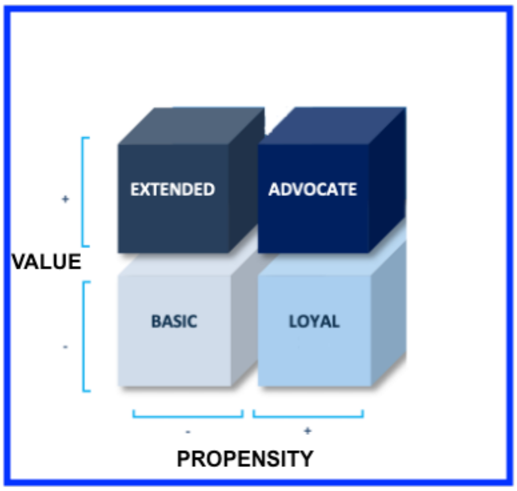


Figure 2.3: Example of ideal segmentation for the customer base of a generic company. Four groups are identified based on the Value and Propensity indexes. Meaningful names are assigned to them to add business sematic to the segmentation.

3
Value and Propensity Indexes

3.1 Value

The value is intended as the economical level of the clients and how much they are willing or can afford to spend on the products. The aim of this section is to create an index to summarize it.

To build an indicator that implements such intuition the information about how much clients spent for the products and the quality/level of the merchandise has to be combined. This has to be done considering only the first five years of history of each client (DF5y). In fact, as previously mentioned, we are working in a context in which we want to model the future using the first 5 years of information.

To build the value index the following variables have been considered [*]:

- *LEVEL DS FIRST PURCHASE* (LIVELLO DS PRIMA VENDITA). Quality level of the product bought in the first purchase. It assumes values between 1 (lowest quality) and 5 (highest quality).

- *N PURCHASES CUSTOMER* (NRO ACQUISTI CLIENTE). How many purchases have been done by the client in their first (0,5] years of history.

[*] The names in the database are almost all in Italian. For a smoother reading they are traslated to english in this report. Note that the Appendix keeps the original version

- *NET AMOUNT FIRST SALE ACTUAL*. Inflation-adjusted first purchase value.
- *avg value purchases actual* (valore medio acquisti actual). Average value of the purchases. Customized variable which is equal to the ratio between *HISTORICAL VALUE CLIENT ACTUAL* (how much clients spent in the first (0,5] years of their history, corrected by the inflation) and *N PURCHASES CUSTOMER* (how many purchases they made).

The selection of these variables for the analysis is based on their semantic relevance, as it appeared reasonable to explore whether they could provide pertinent information about the customers' economic value.

3.1.1 PCA as Value Index generator

The analysis has been guided by the correlation matrix between the four variables (Figure 3.1). It shows that all the variables apart from *avg value purchases actual* have an high positive correlation. Since the objective is to calculate a single number to represent the value index, a principal component analysis (PCA) ([15], pages 60-63) has been considered as a way to summarize the information contained in this set of variables. The motivation is that PCA can be implemented as an index constructor as it provides an aggregation of several features that maintains (as much as possible) their variability, converging it to a smaller number of dimensions. The concept and usage of PCA as indices generator is described in more detail in [16].

The following procedure describes how to apply PCA to extract a single vector summarizing the information contained in the data. Given $X_{N\times 4}$ the matrix containing the 4 variables under analysis and K vectors of weights $w_k = (w_{1k}, w_{2k}, w_{3k}, w_{4k})$, the Principal Component Scores vectors are defined as $t_i = (t_{i1}, ..., t_{iK})$. Each weights vector w_k maps each row x_i of X (in our context, each customer's observations) to its principal component score:

$$t_{ik} = x_i \cdot w_k \tag{3.1}$$

for $i = 1, ..., N$ (customers index) and $k = 1, ..., K$ (principal component scores index). Our purpose is to extract one single Score Index. For this reason, we can focus on t_1 and, consequentially, w_1. The required property of maximizing the variance (i.e, keep as much information as possible from the original X) is verified if and only if the weights w_1 assume a specific

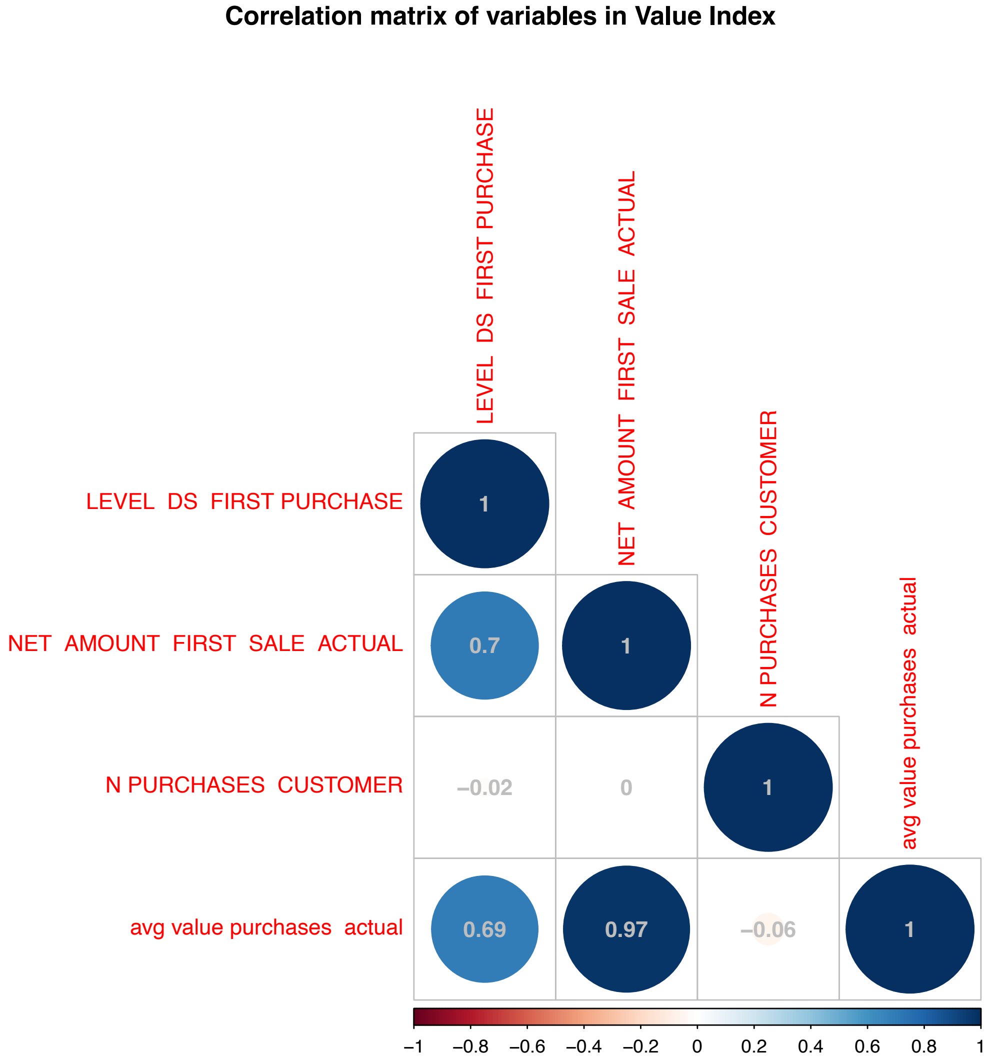


Figure 3.1: Correlation matrix between the variables selected for the calculation of the Value Index. They are all higly positively correlated excluded the number of purchases per customer. This suggests that: A) high levels products are usually sold for higher prices, B) the average value of the purchases is mainly determined by the first one, C) the number of purchases is not correlated to the level and the amount of the purchases.

shape. The required mathematical relation for that property to hold is that:

$$w_1 = \underset{||w_1||=1}{\arg\max}\{\sum_i (t_{i1})^2\} = \underset{||w_1||=1}{\arg\max}\{(x_i \cdot w_1)^2\} \tag{3.2}$$

This is equivalent to say that w_1 is the first eigenvector of $X^T X$. Thank to this relation, the vector of scores t_1 can be calculated as linear transformation of the input data (X) through the

first eigenvector of $X^T X$. The first eigenvector can be calculated in different ways. A possibility is the Singular Value Decomposition [17].
Note that X has been scaled to avoid over-representation of variables due to their unit of measurement.
The results for all the components are shown in Table 3.1. As expected, a significant portion of the total variance (65%) is explained by the first component.
The analysis of the loadings shows that the first component is mainly influenced by three variables: *LEVEL DS FIRST PURCHASE*, *NET AMOUNT FIRST SALE ACTUAL* and *avg value purchases actual*.

	Comp.1	Comp.2	Comp.3	Comp.4
Proportion of variance	0.646	0.250	0.098	0.006
Cumulative Proportion	0.646	0.896	0.993	1.000
LEVEL DS FIRST PURCHASE	0.53	0.02	0.85	0.02
NET AMOUNT FIRST SALE ACTUAL	0.60	0.05	-0.35	-0.71
N PURCHASES CUSTOMER	-0.03	1.00	-0.01	0.05
avg value purchases actual	0.60	-0.02	-0.39	0.70

Table 3.1: The table shows the variance explained by the PCA components and how they are impacted by the 4 variables. First 2 rows: variance explained by each component. The first component explain aroun 65% of the original variance, the first 2 together almost the 90%.
Last 4 rows: PCA loadings. Each row shows how much each component is determined by a certain variable. E.g. the first component is determined almost equally by *LEVEL DS FIRST PURCHASE*, *NET AMOUNT FIRST SALE ACTUAL* and *avg value purchases actual*. The second one is determined by *N PURCHASES CUSTOMER*

Given these evidences, it seems reasonable to consider the scores provided by the PCA as a value index.

However, the dissertation presented so far is limited to a part of client's history. To be considered a valid value index, the PCA must accurately capture the overall value of the client, rather than solely relying on the identified variables. It is essential to evaluate whether the PCA scores effectively reflect the customer's overall value to the company, rather than just summarizing the data from the initial time slot.

To assess the effectiveness of the PCA first dimension's scores in this context, we analyzed how they correlate with the customer's value at the end of the observed period. This joint analysis of the incoming PCA index and the customer's value provides a way to verify its effec-

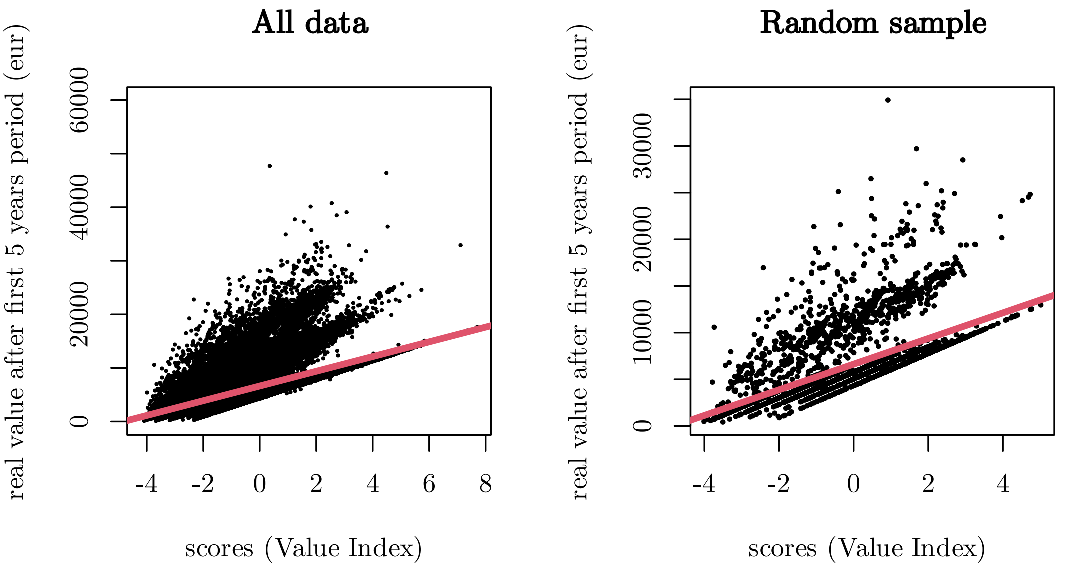


Figure 3.2: The observed value of the client at the end of the first five year period of observation against the scores of the PCA interpreted as Value Index. The **left plot** represents all the customers, while the **right plot** represents a random subset of 5000 customers for a clearer visualization. The red line represents the linear relation between the two features. The PCA scores are significant in explaining the real value with a $pvalue < 2 \cdot 10^{-16}$ for the test against the null model. It is worth noting that a linear relationship is evident in the lowest levels of the real value after the first 5 years. This occurrence is because the value is determined by the cumulative sum of all purchases, including the initial one. Since the scores are influenced positively by the first purchase amount, there cannot be observations with high scores and an overall sum of purchases below a certain threshold.

tiveness as indicator of the customer's overall value.

Figure 3.2 presents the result, showing that, as desired, higher values of the index are related to higher values of the actual variable. Note that the aim is not to have a perfect relation between these two variables, but it seemed reasonable to expect the index to be positively correlated to the real future observed value of a client. A model to assess the statistical significance of the linear relation between the PCA Scores and the real value after 5 years has been estimated, supporting the hypothesis of the scores effectively explaining the distribution of the value which provides an indication of how well the scores explain the targeted feature (revenues after 5 years). The proportionality coefficient is equal to 1374, signifying that to convert from the value index score's space to the real value's space, the scores must be multiplied by 1374. The R^2 index indicates that 39% of the variance of the real observed value is explained by the value index.

In this analysis, we utilized the total amount spent by each client at the end of the first five-year period. Although we had access to the total amount spent over the entire time slot (10 years), we deliberately chose to avoid using future information to minimize overfitting concerns during the final evaluation of overall performance that will be made leveraging on that.

Given these considerations, the first vector of scores t_1 has been considered as an indicator of the value of each customer.

3.2 Propensity

The aim of this Section is to build a dimension representing the propensity to be used for the customer base segmentation. Propensity was intended as the customers' tendency to repurchase the product. Note that a customer is defined as a person that already made one purchase. The ideal propensity index would be a measure of how much a client is willing to buy again one or more times in the future five-year slot given the information observed in the first five. The information contained DF10y is needed to build such an indicator, since this dataset contains the knowledge about the future behaviour in terms of purchases. For this reason, in this section the analysis will be performed on this dataframe.

Propensity can be viewed as the likelihood of a customer repurchasing a product in the future. If a customer is highly probable to make a second purchase, their propensity should be considered high. As the probability increases, so does their propensity.
On the other hand, domain knowledge suggests that the repurchase probability depends on the aging of the customer-company relationship: two people with the exact same features will buy again with different probabilities given how much time passed since their first purchase. For this reason, it seemed reasonable to estimate the probability of repurchase as a function of time from first purchase.
To sum up, for each customer *i* there is the need to calculate their probability of buying a product over time, given the fact that they have already bought it once and having some information x_i about them:

$$propensity(t|x_i) = P(client\,i\,buys\,at\,time\,T+t\,|\,he\,did\,a\,purchase\,at\,time\,T). \quad (3.3)$$

Equation (3.3) suggests that survival models [18] can be used. In order to apply such techniques as described in Section 3.2.1, some preprocessing is needed as detailed in Section 3.2.2.

3.2.1 Survival Models

The need of analyzing time-to-event data has been raised in the most diverse fields, such as medicine, biology, epidemiology, engineering, economics etc. The primary characteristic that stands out in this type of data is its complexity due to the prevalent issue of censoring in real-world applications. Censoring occurs when an individual's lifespan or observation period is restricted to a specific time frame, making the analysis more challenging. For example, in a medical study investigating the survival rates of patients with a certain disease, censoring can occur if some patients are still alive or under observation at the end of the study period. Their exact survival times beyond the study's duration remain unknown, leading to right-censoring. The main difference compared to traditional models (e.g. gamma regressions for continuous positive variables) is that the time under observation variable is properly handled together with the censoring. The situation of our analysis involves only right censoring due to the end of the period of collection of the data. Censured observations are the ones that had not repeated purchases at the date at which the data was extracted from the database. For this reason, given t_i a random variable representing the time until some specified event, two scenarios can raise: the event is actually observed or the event is not observed before the end of the observation period (C). The censoring information is represented by a variable δ_i such that:

$$\delta_i = \begin{cases} 1\, if\; t_i <= C \\ 0\, if\; t_i > C \end{cases} \tag{3.4}$$

This survival information is then modeled as composed by the time-to-event and the censoring information: (t_i, δ_i).

A few basic quantities are needed to understand survival data and models as described in [19]. The following Sections are devoted to addressing this matter.

3.2.1.1 Survival Function

The time t at which the event is observed is modeled by mean of a random variable T. The survival function describes the probability of an individual surviving beyond a time t

$$S(t) = P(T > t) = 1 - F(t) \tag{3.5}$$

where $F(t)$ is the cumulative distribution function of T.
In the context of our analysis, $S(t)$ represents the probability of not buying again before time

t.

3.2.1.2 Hazard Function

The hazard function can be viewed as the probability of an individual at time t to experience the event in the next instant:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{(\Delta t)} \tag{3.6}$$

If T is continuous, then the related cumulative hazard function $H(t)$ is

$$H(t) = \int_0^t h(u)du = -\log[S(t)] \tag{3.7}$$

In the context of our analysis, $h(t)$ represents the hazard ('risk') of buying at time t.

3.2.1.3 Non parametric estimators

Two non parametric estimators are mainly used in the context of survival data.
Kaplan-Meier survival function estimator:

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i} \tag{3.8}$$

Nelson Aalen cumulative hazard estimator:

$$\hat{H}(t) = \sum_{t_i < t} \frac{d_i}{n_i} \tag{3.9}$$

where d_i is the number of events at time t_i, n_i is the number of individuals at risk at time t_i.

The probability of observing the event at time t can be estimated as:

$$\hat{p}(t_i) = \hat{S}(t_{i-1}) - \hat{S}(t_i) \tag{3.10}$$

3.2.1.4 Cox Model

In 1972 Cox [20] proposed a semiparametric Proportional Hazards Model to quantify the relationship between the time to event and p explanatory variables X that became widely used in the domain [19].

The model follows the subsequent structure:

$$h(t|X) = h_0(t) \cdot e^{X\beta} \tag{3.11}$$

where h_0 is an arbitrary baseline hazard rate and can be estimated through non-parametric methods, β is a vector of parameters. Because of its multiplicative shape, hazard functions for any pair of individuals i and j are proportional. Their ratio is equal to $\exp[\sum_{n=1}^{p} \beta_n(x_i - x_j)]$. This formulation of the Cox proportional hazards model does not permit the use of the regular likelihood to estimate the model parameters. The reason lies in the structure of the model, where the baseline hazard function, denoted as h_0, is left unspecified. Since it is not explicitly defined, direct estimation of model parameters using standard likelihood methods becomes infeasible.

To address this challenge, the concept of partial likelihood is introduced. The partial likelihood is an adaptation of the standard likelihood that allows to estimate the regression coefficients while effectively accounting for the unspecified baseline hazard function. It involves calculating the conditional probability of the observed event times given the observed covariates. This probability is then used to construct the likelihood function, which is maximized to obtain the estimates of the regression coefficients.

3.2.1.5 Survival Forest

Random Forests [21] are ensembles of classification or regression trees. They aim to reduce variance in the estimates, which often leads to improved performance compared to non-ensemble models. This family of models has been extended to the domain of survival data [22], introducing the concept of cumulative hazard function at a leaves level. The algorithm, as described in Algorithm 3.2, is based on the aggregation of multiple survival trees. A survival tree is a non parametric tree-based model adapted to censored time-to-event data. Various implementations of survival trees have been proposed, as detailed in [23].

In the domain of survival random forests, the most commonly used survival tree version is the one proposed in [22], which follows a CART-like paradigm [24]. The model fitting process resembles that of a classification CART, but with a different splitting rule (refer to Table 3.3) that aims to maximize the ‘survival difference’ between observations in different nodes. The survival tree structure we used is described in Algorithm 3.1. It is worth noting that in random forests no pruning is needed since the overfitting is mitigated by the aggregation of multiple trees, as pointed out by [25].

Algorithm 3.1 Survival Tree for Survival Forests

```
Data: X
set the root node $N_0$
  set $mtry$ number of variables over which to attempt the split at each tree iteration
  set $min.obs$ the minimum number of observations to maintain in all the leaves
  set $max.depth$ the maximum depth of all the trees
  set a splitting criterion
  while at least 1 active node do
      if $N_i$ has # parents nodes = max.depth - 1 then
          $N_i$ inactive
      end
      randomly select $mtry$ columns from $X$
        select the best predictor variable $X_j$ according to the splitting criterion
        if $X_j$ numerical then
            select a splitting point $s$
              divide the obs in $X$ into $X_{left}$ $(X_{ij} < s)$ and $X_{right}$ $(X_{ij} >= s)$
      end
      if $X_j$ categorical then
          select a splitting level $c$
            divide the obs in $X$ into $X_{left}$ $(X_{ij} \neq s)$ and $X_{right}$ $(X_{ij} = s)$
      end
      if ($X_{right}$ or $X_{left}$ have < $min.obs$) then
          $N_i$ inactive
      end
      Create 2 new nodes for $X_{right}$ and $X_{left}$
end
```

Algorithm 3.2 Survival Random Forest

```
Data: X
set B number of trees
 set the number of variables over which to attempt the split at each tree iteration
 set the minimum number of observations to maintain in all the leaves
 set $p$ proportion of rows to use as training set for each tree
 set the maximum depth of all the trees
 for $b$ in $1 : B$ do
     randomly select a sample $X^{(b)}$ from the data keeping a proportion $p$ of the rows
      set $I_{i,b} = 1$ if observation $i \in X^{(b)}$
      build a survival tree as in Algorithm 3.1, obtaining $L$ terminal nodes
      for $l$ in $L$ do
          get the distinct event times $t_1 < ... < t_{N(l)}$
           get $d_h, Y_h$ number of deaths and individuals at risk at time $t_h$ $\forall h \in 1, ..., N(l)$
           calculate the Cumulative Hazard Function (CHF) $\hat{H}(t) = \sum_{t_{h \leq t}} \frac{d_h}{Y_h}$
     end
     set $\hat{H}_b(t|x_i)$ as the CHF of the final node to which the observation $i$ is assigned
end
for $x_i$ row in X do
     calculate the Out Of Bag ensemble CHF $\hat{H}_e^{(OOB)}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} \hat{H}_b(t|x_i)}{\sum_{b=1}^{B} I_{i,b}}$
      calculate the final CHF prediction $\hat{H}_e(t|x_i) = \frac{1}{B} \sum_{b=1}^{B} \hat{H}_b(t|x_i)$
end
calculate the prediction error (C-index) for the ensemble CHFs $\hat{H}_e^{(OOB)}(t|X)$
```

3.2.1.6 Evaluation Metrics

Researches have proposed several evaluation criteria for survival models. The most relevant are presented in Table3.2. The metric we chose for the model selection is the Concordance Index because of its properties: it does not depend on a single fixed time and can be interpreted as misclassification probability.

metric	reference	description	properties
Cox PLS	[20], [26]	extension of Cox model evaluation to non parametric scenarios	assumes proportional hazards; instable for trees with many leaves
C-index	[22]	comparing all the pairs of survival curves generalizing the ROC curve	interpretation as misclassification probability; not dependent on a single fixed time
BS	[27]	square prediction error adjusted by the inverse probability of censoring	dependent on a single fixed time
IBS	[28]	scaled integral of BS	not dependent on a single fixed time
IAE	[29]	integral of the square distance between the estimated survival curve and the non-parametric estimate of the true one	estimates of the true survival curve tend to be weak
ISE	[30]	as IAE but square distance instead of the absolute distance	estimates of the 'real' survival curve tend to be weak
MAE	[31]	average absolute error between the predicted survival time and the true survival time in the uncensored sample	does not consider censored data; requires survival times

Table 3.2: Evaluation metrics for survival models. Cox Partial-Likelihood Score (Cox PLS), Concordance index (C-index), Brier Score (BS), Integrated Brier Score (IBS), Integrated Absolute Error (IAE), Integrated Square Error (ISE), Mean Absolute Error (MAE). For a comprehensive overview see [1]

The C-index is calculated for each pair of statistical units i and j, with $i \neq j$ and at most one of them is censored. It is based on their hazard scores and times-to-event:

$$C - index = \frac{\sum_{i,j} I(T_j < T_i) \cdot \sum_{i,j} I(\eta_j < \eta_i) \cdot \delta_i}{\sum_{i,j} I_(T_j < T_i)} \tag{3.12}$$

Where T_i is the observed time-to-event for the statistical unit i, δ_i is the censoring as defined in (3.4), $I(x > k) = 1$ if $x > k$, 0 otherwise, η_i is the risk score assigned by the model to unit i.

The underlying idea is that units with the higher risk score should have a shorter time-to-event. For each pair (i, j) two cases can raise:

- i and j are not censored ($\delta_i = \delta_j = 1$), i.e. the event was observed for both statistical units.
  If the risk of i is greater than the risk of j ($\eta_i > \eta_j$) and the event has been observed on i before than on j ($T_i < T_j$), the pair (i, j) is defined as *concordant*.
  The pair is defined as *discordant* if $\eta_i > \eta_j$ and $T_i > T_j$.

- i is not censored, j is censored ($\delta_i = 1, \delta_j = 0$). If $T_j < T_i$, the pair is not considered for the calculation since it cannot be assessed whether i observed the event before or after j. If $T_j > T_i$, (i, j) is *concordant* if $\eta_i > \eta_j$, *discordant* if $\eta_i < \eta_j$.

This index can be considered as the generalization [22] of the traditional classification tasks' ROC curve [32]. In traditional binary classification problems, the ROC curve is used to assess the goodness of a model in distinguishing between two classes. It plots the true positive rate against the false positive rate at various classification thresholds. In the context of survival analysis, the positive class is the event of interest ($\delta = 1$) and the negative class is the censored data ($\delta = 0$). The C-index provides an estimate of how well the model can discriminate between individuals who will experience the event sooner and those who will experience it later or not at all.

3.2.2 Preprocessing for Survival Models training

In order to apply survival models some preprocessing steps are necessary. Table 3.4 provides a summary of this process.

The ideal scenario is one in which the number of repurchases and the time at which they had been performed is available for each customer. Unfortunately, the information about repurchases is not at the desired granularity. The only available data is:

- time past from first purchase,
- total number of purchases,
- time past from last purchase.

This means there is insufficient information about those clients who have purchased more than twice. For example, a customer that has made three purchases in total (the original one

rule	reference	description
L_1 distance	[33]	maximizing area between the Kaplan-Meier survival functions of the child nodes
EL	[34]	minimizing the exponential log-likelihood loss of the hazard at time t
Log-Rank test	[2]	calculate for both incoming nodes the statistic $\frac{observed\,deaths - expected}{\sqrt{Var(expected)}}$, sum them and apply the chi square statistical test, minimizing the adjusted *p value*
2 samples test	[35]	calculate the statistic as in the Log-Rank test with some wights to make it suitable for other parametric tests
deviance	[36]	minimize the node deviance approximating the the full likelihood using the Nelson-Aalen estimator
impurity	[37]	minimizing the weighted sum of $Var(times\,in\,node)$ + $censor's\,impurity\,indicator$
residuals	[38]	minimizing $Deviance(residuals)$ from a null Cox model, using Martingale or Deviance residuals

Table 3.3: Summary of splitting criteria. For a comprehensive overview see [2] and [1].

plus two repurchases) the only available data is about the first and the last purchase: the information about the second is not available. This can be problematic when building a model that explains the dependency between the time and the probability of repurchase. In fact, if the only available information is about the first and the third purchase, a model built based on this would lead to estimates that would consider the second as non-existent. The distribution of the total number of purchases per customer (Table 3.5) shows that only around 7% made more than two. We decided to filter out those with 3 or more to avoid misinterpretations.

An equality check between the age of the first purchase and the age at 2022 minus the time passed between the two has been performed and inconsistent observations have been filtered out.

Survival Models require at least two variables: 1) the censoring flag, equal to 1 if the event of interest has been observed, 0 if not 2) the time-to-event. The former can be formulated as follows:

step	task	% of original data kept	drop rows
1	remove customers with more than 2 purchases	92.67%	✓
2	remove rows with inconsistent age first purchase	92.66%	✓
3	remove rows with time under observation =0	92.65%	✗
4	collapse and clean levels of the APPT_TYPE and SUBTYPE variable	92.65%	✗
5	cut age of first purchase into classes	92.65%	✗
6	infer missing data using median and new class	92.65%	✗

Table 3.4: Preprocessing pipeline.

n purchases	n clients	percentuage
1	27386	65.7%
2	11221	26.9%
3	2641	6.3%
4	360	0.9%
5	42	0.1%
6	8	0.0%
9	1	0.0%
10	1	0.0%
11	1	0.0%

Table 3.5: Distribution of clients given the number of purchases they did. The table includes only customers from the DF10y dataset since this is the dataset on which the analysis will be performed. Around 65% of them never did a second purchase, 27% did exactly 2 purchases and around 7% did more than 2.

$$repurchase\ flag = \begin{cases} 1\ if\ N\ purchases\ customer = 2 \\ 0\ if\ N\ purchases\ customer = 1 \end{cases} \tag{3.13}$$

The time to event cannot be equal to 0 for uncensored data, i.e., the second purchase cannot be done on the same day of the first one. In the context of the business from which this dataset originates, this assumption is reasonable. Typically, a significant amount of time ($>$ 3 years) elapses between two purchases.

The categorical variable *APPT TYPE POST FIRST PURCH* (type of first appointment

booked after the first purchase) has many levels and some of them only have few observations. For this reason, the majority have been cleaned and kept, two of them have been collapsed into a new level and the residual have been mapped together.
The levels ”Adjustment Follow Up”, ”Aftercare”, ”Annual Retest”, ”HA” are kept as they are; ”HAE/Consultation” and ”Online HAE” are collapsed into a single level ”HAE”; all the remaining all collapsed into ”Other”.

The age of first purchase has been divided into the following classes: ’0-45’, ’45-55’, ’55-60’, ’60-65’, ’65-70’, ’70-75’, ’75-80’, ’80-85’, ’85-90’, ’90-95’, ’95+’. This has been done in order to visualize it as a segmentation variable for cases in which non parametric approaches for survival curves estimation were implemented.

The families of model that will be used do not automatically handle missing data. They have been inferred using the median for numerical variables, inserting a new level for qualitative variables.

3.2.3 Survival Models Training and Selection

Once the data has been preprocessed, the actual survival analysis can be performed. The adopted procedure entailed conducting an exploratory data analysis (EDA) of the data and subsequently applying various families of models to evaluate their performances. Before that, a 80/10/10 train/validation/test random split was carried out and all the following analysis were performed on the train and validated on respective set. The test set has been kept for the final assessment of the performances

3.2.3.1 Survival EDA

The initial stage of survival modeling consists in conducting a non-parametric exploratory analysis. There are several reasons for undertaking it:

- estimating survival curves for each group of customers with the same variable level,
- assessing data distribution peculiarities,
- comparing survival curves of different customer groups to understand their behavior,

- identifying variables that could be more effective in explaining the target probability compared to others,

Classical approaches that do not account for censoring can be applied to survival data, but they come at the expense of losing crucial information that could otherwise be derived from this portion of the data. To circumvent this problem, the approach undertaken involved plotting the survival curves obtained from the Kaplan-Meier estimator. This was achieved by stratifying the data into different subsets based on the levels of the variable of interest. The non-parametric estimator described in Section 3.2.1.3 was then used to calculate the estimated survival curve for each group.

To separate the clients into groups, continuous variables needed to be transformed into categorical. For the sake of the exploratory analysis, continuous variables were then grouped into three classes, labeled as *low*, *medium*, and *high*, based on their quartiles. More precisely, each numerical variable was divided into three categories, taking into account the values of the first and third quartiles. The plots resulting from this procedure are presented in Figure 3.3.

The majority of the variables appear to distinguish groups of customers with distinct survival curves. Those variables might be significant in identifying the probability of repurchasing over time. However, relying solely on visual analysis of the plots is inadequate. To verify the statistical significance of each variable defining groups with non-equal survival curves, the log-rank test [39] has been employed, leveraging its asymptotic Chi-square distribution to calculate the *p value*.

We conducted multiple statistical tests on the same data: for each one of the 11 variables, the test has been done on a different combination of the same data, the time-to-event and the censoring variables that together provide the survival curves. This approach led us to encounter the multiple comparisons problem, which arises when considering numerous statistical inferences simultaneously. To address this issue, we applied the Bonferroni Correction [40]. This adjustment involves dividing the desired significance level (α) by the number of tests performed, which, in our case, amounted to 11. To maintain a significance level of 95% ($\alpha = 0.05$), the threshold for comparing observed *p values* was set at $0.05/11 = 0.0045$. Only thanks to this correction we can state that, at a level of significance equal to 95%, the two survival curves identified by the AVERAGE PURCHASES AMOUNT are not different. For all the other cases, the null hypothesis of equality of the curves is rejected. Nevertheless, this result should not be regarded as definitive. Instead, it serves as an indication of the direction to pursue in the modeling phase.

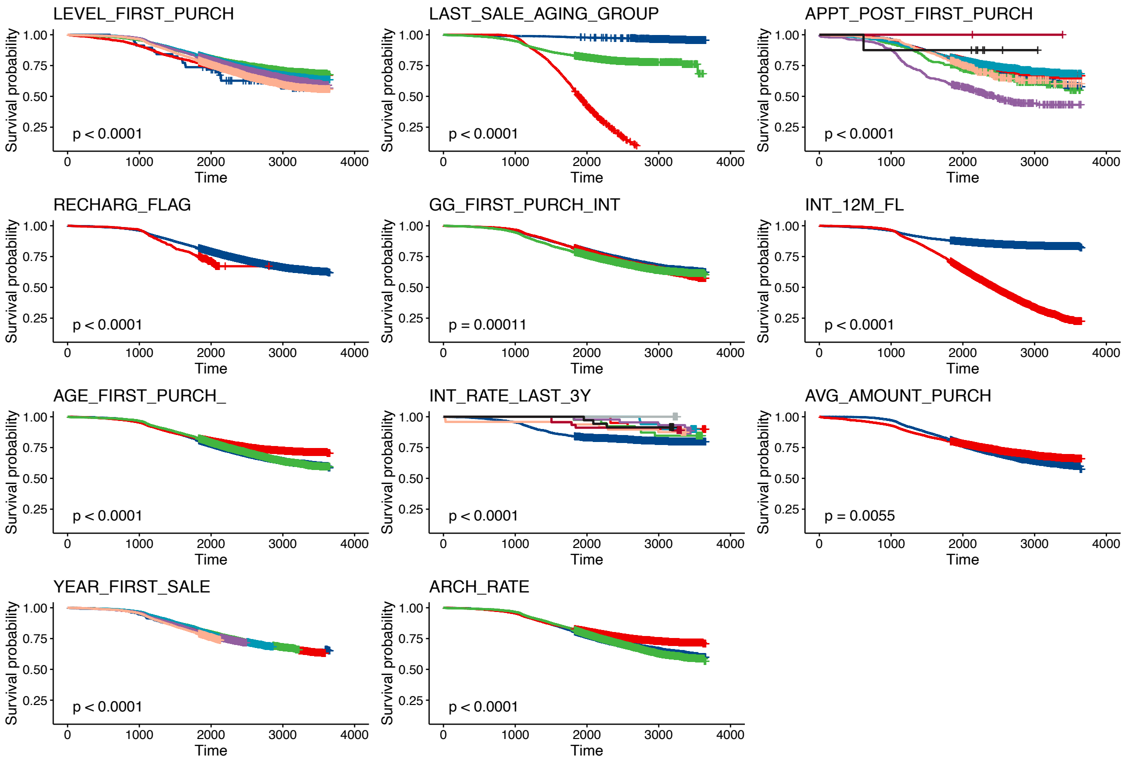


Figure 3.3: Survival curves for the groups defined by the levels of 11 variables. See the Appendix for more details about the variables.

For each plot the *p value* of the log-rank test is printed on the bottom-left for the hypothesis of all the curves being equal. The Bonferroni-adjusted threshold to maintain a confidence interval of 95% is 0.0045. For this reason, AVG_AMOUNT_PURCH is considered as not significant in this regards, since $0.0055 > 0.0045$.

A second relevant finding is that the survival curves tend to overlap. This behaviour is evident, for example, for the groups identified by the red and the blue lines in the RECHARGEABLE FL plot. This has to be taken into account when dealing with Cox models and its assumption of proportional hazards.

Beside the considerations about the non parametric estimates, it would be ideal to include as many possibly useful predictive variables as possible in our models. However, we had to remove all the variables used to calculate the value index to assess the independency assumption between the value and propensity indexes and the redundant variables causing over fitting. For this reason, LEVEL OF FIRST PURCHASE and AVERAGE AMOUNT PURCHASE were dropped in the next steps.

3.2.3.2 Cox Model Variables Selection

The prevailing model commonly employed in this context is the Cox one, as detailed in Section 3.2.1.4. For this reason, we initiated with the estimation of such model. First, all the variables have been included in the model. These are:

- the interaction rate of the last 3 years, calculated without considering the last 5 years of history(*tasso interaz last 3 Y dopo 5y*)
- how long the person has been a client without considering the last 5 years of history (*AGING CLIENTE AL 2017*)
- how long the person has been a client without considering the last 5 years of history grouped into 3 classes (*LAST SALE AGING GROUP AL 2017*)
- age of first purchase (*eta cliente primo acquisto*)
- age of first purchase grouped by classes (*eta cliente primo acquisto grouped*)
- appointment type after the first purchase (*APPT TYPE POST PRIMO ACQ*)
- days between the first purchase and the next appointment (*GG DISTANZA INTERAZ PRIMO ACQ*)

Once the baseline model had been defined containing as linear terms on the exponential scale the 7 variables previously illustrated, variable selection was performed. Three steps were designed: 1) selection of the continuous versus the categorical versions; 2) selection of non linear transformation of features within the model; 3) selection of the interaction terms of pairs of variables.

CONTINUOUS VERSUS CATEGORICAL

There are two pairs of variables representing the same information in both a continuous and categorical shape, the age and the aging of the client-company relationship. A choice between the two had to be made. The continuous version was given to to model first. Then, the categorical versions were added and the significance of them checked. In regards to the age, the continuous version was selected. For the client-company relationship's aging, the categorical version proved to be more significant.

INTERACTION TERMS

Then, interactions between pairs of variables were included according to the one-step-forward paradigm: at each iteration, an incoming interaction term were added and kept in the model if it improved its performances in terms of C-index. Steps and results are shown in Table 3.6.

QUADRATIC TERMS

Similarly, the one-step-forward paradigm was employed to incorporate quadratic transformations of the numerical features. At each iteration, an incoming variable was squared, and it remained in the model only if its inclusion improved performance in terms of C-index. Steps and results are presented in Table 3.7.
Higher order polynomial terms have not been taken into account to avoid overparametrization.

After completing these steps, a decision needed to be made regarding the seven original variables. It turned out that not all of them were statistically significant for the model. Specifically, the effect of the *aging client-customer relationship* was found not to be significantly different from 0. On one hand, we could have followed the same logic used for interactions and quadratic terms, where we fit the model with and without the variable and retain the version with the highest validation C-index. However, we chose to keep that variable in the model for the sake of explainability. Having the variable in an interaction term but not as an individual predictor could sometimes be challenging to interpret from a business perspective.

3.2.3.3 Selected Cox Model

The selected Cox Model resulted from the previous steps. The variables maintained in the model are presented in Table 3.8 together with their effect on the hazard function.

Once the model was selected, its assumptions needed to be verified. The main assumption underlying the Cox model is that the hazard rates are proportional. It follows that given a pair of hazard functions related to two different statistical units i and j, they are proportional and their ratio is equal to $\exp\{\sum_{n=1}^{p} \beta_n(x_i - x_j)\}$ where p is the number of parameters.
This assumption is verified if and only if the coefficients β are time-independent. The proof of this relationship can be found at [41]. Figure 3.4 shows the estimated relation between time and the coefficients. The assumption underlying the models is not verified: the predictors' impact on hazard ratios are not constant over time. This can significantly impact the interpretation and accuracy of the model's predictions, leading to incorrect inferences and biased parameter estimates.

step	variables	C-index	kept
0	no interaction terms	0.612	
1	int rate last 3 y & age first purchase	0.612	✗
2	int rate last 3 y & appt type post 1st purchase	0.610	✗
3	int rate last 3 y & aging client-customer relationship	0.614	✓
4	int rate last 3 y & days btw 1st purchase and next appt	0.614	✗
5	age first purchase & appt type post 1st purchase	0.613	✗
6	age first purchase & aging client-customer relationship	0.616	✓
7	age first purchase & days btw 1st purchase and next appt	0.616	✗
8	appt type post 1st purchase & aging client-customer relationship	0.616	✗
9	appt type post 1st purchase & days btw 1st purchase and next appt	0.617	✓
10	aging client-customer relationship & days btw 1st purchase and next appt	0.616	✗

Table 3.6: One-step-forward approach for Cox model variables' interaction selection. At each step, the interaction between a pair of variables is given as feature to the model and the performances on the validation set are calculated. If performances increase, the interaction term is kept in the next steps.

step	quadratic term	C-index	kept
0	selected interactions & no squared terms	0.617	
1	int rate last 3 y	0.619	✓
2	age first purchase	0.0.620	✓
3	days btw 1st purchase and next appt	0.621	✓

Table 3.7: One-step-forward approach for Cox model variables' quadratic terms selection. At each step, the squared variable is given as feature to the model and the performances on the validation set are calculated. If performances increase, the interaction term is kept in the next steps.

The selected evaluation metric, C-index, has a value of 0.621. As it can range between 0 and 1, with 0 being the worst and 1 being the best performance, we aim to explore whether we can achieve improvement by employing a different family of models with less restrictive assumptions. To explore potential enhancements, we have opted for the application of Survival Random Forest. These models possess a non-parametric structure that does not rely on limitations such as the proportional hazards assumption.

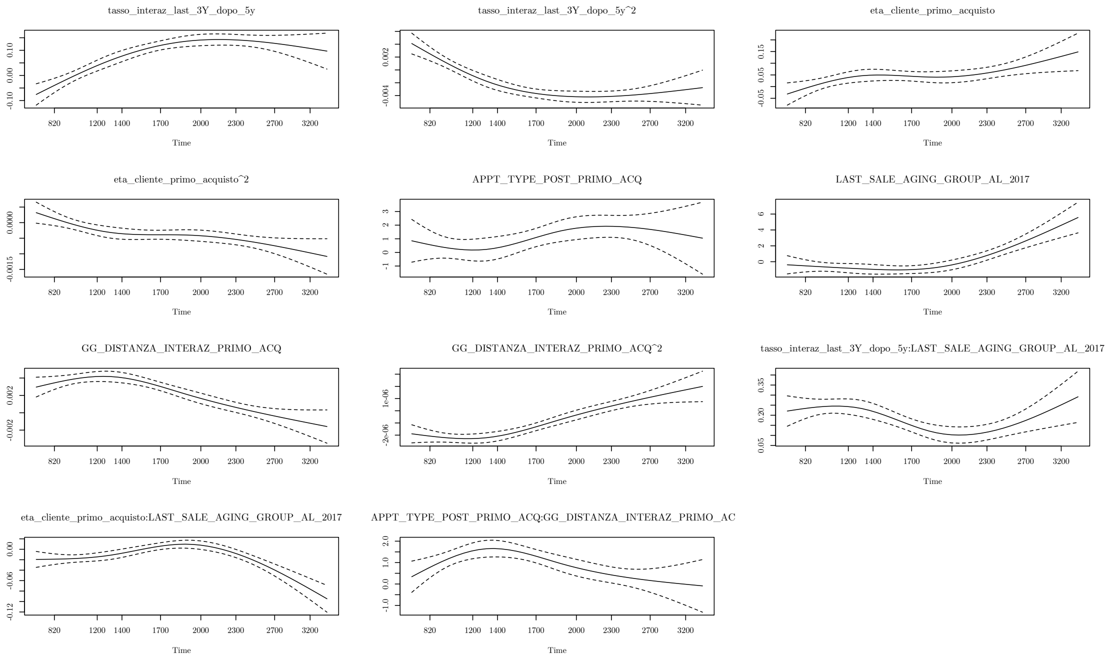


Figure 3.4: Test to assess whether the coefficients β of the Cox model are time-independent. If this test confirms the hypothesis, it verifies the assumption of proportional hazards. Visually, if the hypothesis holds true, the plot should exhibit a horizontal line. The only instances where the test does not reject the hypothesis of independence are for the first (*tasso interaz last 3Y dopo 5y*) and seventh (*GG DISTANZA INTERAZ PRIMO ACQ*) cases, with p-values of 0.21 and 0.19 respectively. However, the global test indicates that the assumption of proportional hazards is rejected.

3.2.3.4 Survival Forest Training and Selection

The training procedure and model selection of a survival forest are more complex and computationally demanding compared to those of the Cox Model. The latter is estimated by maximizing the prime derivative of the partial likelihood function. In contrast, for the survival forest, a fixed number of trees need to be fitted, necessitating decisions regarding various parameters, including branches' maximum depth, the minimum number of observations in each leaf, the

variable	effect on h_0	$pvalue$
int rate last 3y	1.08	< 0.0001
squared int rate last 3 y	0.998	< 0.0001
age first purchase	1.04	< 0.0001
squared age first purchase	0.999	< 0.0001
days btw 1st purch and next appt	1.003	< 0.0001
squared days btw 1st purch and next appt	0.999	< 0.0001
aging client-customer relationship level='HOT'	0.810	0.212
appt type post 1st purchase level='AFERCARE'	0.956	0.088
appt type post 1st purchase level='ANNUAL RETEST'	0.679	0.041
appt type post 1st purchase level='HA'	0.942	0.350
appt type post 1st purchase level='HAE'	1.145	0.1833
appt type post 1st purchase level='Other'	1.074	0.4694
INT(int rate last 3 y; aging client-customer relationship level = 'HOT')	1.199	< 0.0001
INT(age first purchase; aging client-customer relationship level = 'HOT')	0.990	< 0.0001
INT(appt type post 1st purchase level='AFERCARE'; days btw 1st purch and next appt)	0.999	0.008
INT(appt type post 1st purchase level='ANNUAL ; days btw 1st purch and next appt)	1.000	0.977
INT(appt type post 1st purchase level='HA'; days btw 1st purch and next appt)	0.999	0.002
INT(appt type post 1st purchase level='HAE'; days btw 1st purch and next appt)	1.001	0.039
INT(appt type post 1st purchase level='Other'; days btw 1st purch and next appt)	0.998	0.256

Table 3.8: Variables selected in the final Cox model. For continuous variables, the product of $exp(\beta)$ and the observed value of the relative variable can be interpreted as a multiplication factor effect on the baseline hazard function. The same interpretation holds for qualitative variables when the feature itself is replaced by an indicator function equal to 1 when the level of the respective variable is observed. For instance, consider a client i with *aging client-customer relationship* level = 'HOT' and *int rate last 3 y* = 0.3. The estimated hazard function for this client is given by $h_i(t) = h_0(t) \cdot \exp\{1 \cdot 0.810\} \cdot \exp\{0.3 \cdot 1.08\} \cdot \exp\{1 \cdot 0.3 \cdot 1.199\}$. The three terms represent the effects related to *aging client-customer relationship*, *int rate last 3 y*, and their interaction, respectively.

number of variables for possible splitting at each node, the splitting rule, and the sampling strategy.

FINE TUNING

Given the vast number of potential choices for all these parameters, addressing all possibilities is infeasible. Therefore, the selection of values for each hyperparameter was conducted using two different approaches. Initially, the ‘ranger’[†] R package’s documentation suggested the values for all hyperparameters. Subsequently, the number of trees and the minimum number of observations for each leaf were fine-tuned over a grid of possible values, and the results are presented in Table 3.9.

The fine-tuning procedure comprises two steps. First, the Survival Random Forest is trained using a specific set of parameters. Then, its performance is evaluated by calculating the C-index on predictions of the validation set. Note that the Out-Of-Bag C-index has not been used in order to make results comparable with the Cox model’s ones.

Variable selection was not necessarily required because the structure of random forests automatically avoids considering irrelevant variables. In fact, this family of models tends not to select those variables that would not have a significant impact in the tree-building procedure.

Table 3.10 shows a summary of the hyper parameters, along with their range of possible and selected values.

	min	**node**	**size**				
# **trees**	8	15	40	60	175	200	300
75	0.660	0.663	0.666	0.673	0.676	0.6771	0.6753
100	0.660	0.666	0.671	0.669	0.675	0.6756	0.6759
150	0.662	0.664	0.669	0.670	0.674	0.6761	0.6766
200	0.660	0.665	0.668	0.672	0.675	0.6767	0.6768

Table 3.9: Validation set C-index calculated on the prediction given by different values of the number of trees and minimum nodes size. The pair of parameters that results in the best metric is number of trees = 150 and minimum node size = 40. The relative cell is highlighted in green.

[†] https://CRAN.R-project.org/package=ranger

variable	description	range	chosen value
number of trees	number of trees to be fitted	$[1, \infty]$	200 (fine-tuned)
min node size	minimum number of observations for each leaf	$[1, \#\ obs]$	75 (fine-tuned)
mtry	at each step the split is performed on one of the $mtry$ selected variables	$[1, \#\ variables]$	2 (library recommendation)
max depth	maximum number of nodes from root to leaf (including both)	$[0, \#\ obs]$	not constrained
sampling rule	how to randomly select the observations to build a each tree	with or without replacement	without replacement
sampling proportion	proportion of obs to select for each tree	$(0, 1)$	63% sampling
split rule	which rule has to be used for the split	Table 3.3 for more details	logrank

Table 3.10: Survival Random Forest hyperparameters. The *number of trees* and *mtry* were selected through a fine-tuning procedure, while the others were chosen based on literature recommendations.

metric	Cox model	Survival Forest
C-index	0.615	0.677
Brier Score (3 yrs)	0.053	0.048
Brier Score (5 yrs)	0.146	0.130
Brier Score (7 yrs)	0.1792	0.1793
IBS (approx)	0.372	0.358

Table 3.11: Comparison of different evaluation metrics between the selected Cox model and Survival Forest. In almost all the cases, the Survival Forest performs better than the Cox model.
Note that the Integrated Brier Score (IBS) has been approximated by the selection of a random subset of observations to reduce its high computational cost.

3.2.3.5 SRF versus Cox Model

Once the survival forest and the cox model have been estimated and their optimal versions selected, they can be compared. The initial comparison is performed based on the metric chosen before model training, namely, the C-index. The final Cox model exhibited a C-index of 0.621 on the validation set, while the Random Forest achieved a C-index of 0.677. The increase of performances compared to the Cox model is approximately 9%, gaining 0.056. The choice is also based on a broader assessment of the models as well. As elaborated in Section 3.2.3.5, the proportional hazards assumptions were not met in the chosen Cox model. In contrast, the Survival Random Forest, with its non-parametric nature, does not rely on any specific assumptions. Therefore, the final selected model is the survival forest.

3.2.4 Indexes Calculation

The aim of this section is to find a method to condense the information from the estimated survival curve into a single value which will serve as a propensity index.

The survival curves represent the probability of not observing the event as a function of the number of days passed from day 0. In our context, day 0 represents the day of first purchase, and the event occurs when the client performs a second purchase. For clarity, the one's complement of the survival curve is calculated. This modification provides an easier interpretation of the curve as the cumulative probability of repurchasing as a function of the time passed from the first purchase.

The task at hand is to summarize the information embedded in the curve into a propensity in-

dicator, which quantifies how likely a client is to make another purchase.
The aim is to transform discrete functional data into scalar values. Initially, we reasoned from a continuous prospective. Defining $\mathcal{S}$ the space of the survival curves $S \in \mathcal{S},\ S : \mathbb{R} \to [0, 1]$, we need a function f such as:

$$f : \mathcal{S} \to \mathbb{R} \tag{3.14}$$

to map the set of the N survival curves predicted by the model into their scalar representation.

Three possible shape have been identified for f given $S \in \mathcal{S}$ and T upper bound for the time t:

- calculating the integral under the curve:

$$f_1(S(t)) = \int_0^T S(t)dt \tag{3.15}$$

- determining the probability reached after a specific time threshold:

$$f_2(S(t)) = S(k_{time}) \tag{3.16}$$

  where $k_{time} \in (0, T)$ is defined as the time threshold of interest

- identifying the time when a certain probability threshold is attained:

$$f_3(S(t)) = \arg\min_t(I(S(t) >= k_{prob}) = 1) \tag{3.17}$$

  where k_{prob} is defined as the probability threshold of interest and $I(S(t) >= x)$ is the indicator function equal to 1 is $S(t) >= x$, 0 otherwise

It is reasonable to consider that certain customers may never exhibit a high probability of making a second purchase within the ten-year period. Indeed, when calculating the index through a probability threshold on the entire dataset, several such situations emerged, the cardinality of this set depending on the threshold itself. In these cases, an arbitrary large time value (9999) is assigned to represent this information that would be lost otherwise.

type	intuition	selected
area under the curve	higher is the curve, higher is the propensity of the customer	✗
time threshold	high values for the probability of doing a repurchase for a certain customer after a fixed amount of time indicate high propensity	✗
probability threshold	large number of days needed to reach a certain cumulative probability of doing a repurchase indicate low propensity	✓

Table 3.12: Possible choices to calculate the propensity index as a summary of the information contained in 1 - the survival curve. The selected one is the probability threshold that provides a straight forward interpretation and properly divides the observations that actually did a repurchase from the ones that did not.

A summary of the three methods and their underlying intuition is presented in (Table 3.12), where the discrete version of the functions has been implemented without loss of generality. The continuous time t has been substituted with the discrete t_i and an example of the final implementation can be found in Figure 3.5.

3.2.5 Index Selection

Similarly to what was done for the value index, the best method was selected considering the future behavior of the clients. It is reasonable to assume that our propensity index should be effective at distinguishing clients who repurchased from those who did not. While this property is not the primary objective of the index, it serves as a useful heuristic for the selection process. We assumed that the best propensity index would be the one that assigns high propensities to customers that actually made a repurchase, low propensities to the others. Figure 3.6 shows how the 3 methods performed on the test data given a threshold for f_2 and f_3.
First, the integral method was discarded since it did not provide any clear separation of the clients that repurchase from the others and at the same time did not allow a straight forward interpretation of the index. The analysis was further extended to explore different thresholds for the other two methods. This exploratory framework's results are shown in Figure 3.7 and Figure 3.8. The purpose was to investigate the distributions of clients in the test set based on various threshold parameters, dividing them into two groups depending on whether they made a repurchase or not.

The probability threshold (f_3) was selected as the preferred method because it demonstrated superior discrimination between customers who repurchased and those who did not. Specifically, when the threshold value is set to 0.4, approximately 75% of clients with no repurchases are assigned the value of 9999, indicating that they will not reach that cumulative probability within the entire observation period, while only 25% of those that repurchased are assigned such a value. On the other hand, clients which one's complement of the survival curve reaches the 0.4 threshold, are homogeneously distributed over the domain constrained by the considered time slot.

The choice of this threshold holds an intuitive interpretation as the time required to reach a 40% cumulative probability of repurchasing. Low values of this index are related to clients that are 'fast' in reaching that threshold, i.e., they have high propensity.

To achieve a more even distribution of observations along the axis, a logarithmic transformation is applied. Since this transformation is monotonic, the ordering of the data is preserved. Subsequently, the result is subtracted to its upper bound (10), indicating that high values of the index correspond to high propensity. To sum up, the index is calculated as:

$$Prop\,Index \;=\; 10 - \log(f_3(S(t))) \tag{3.18}$$

where f_3 defined in (3.2.4).

In conclusion, the construction of the propensity index has been performed in three steps: fit the survival models on the training set, select the best model on the validation set, select the best summarizing index on the test set.

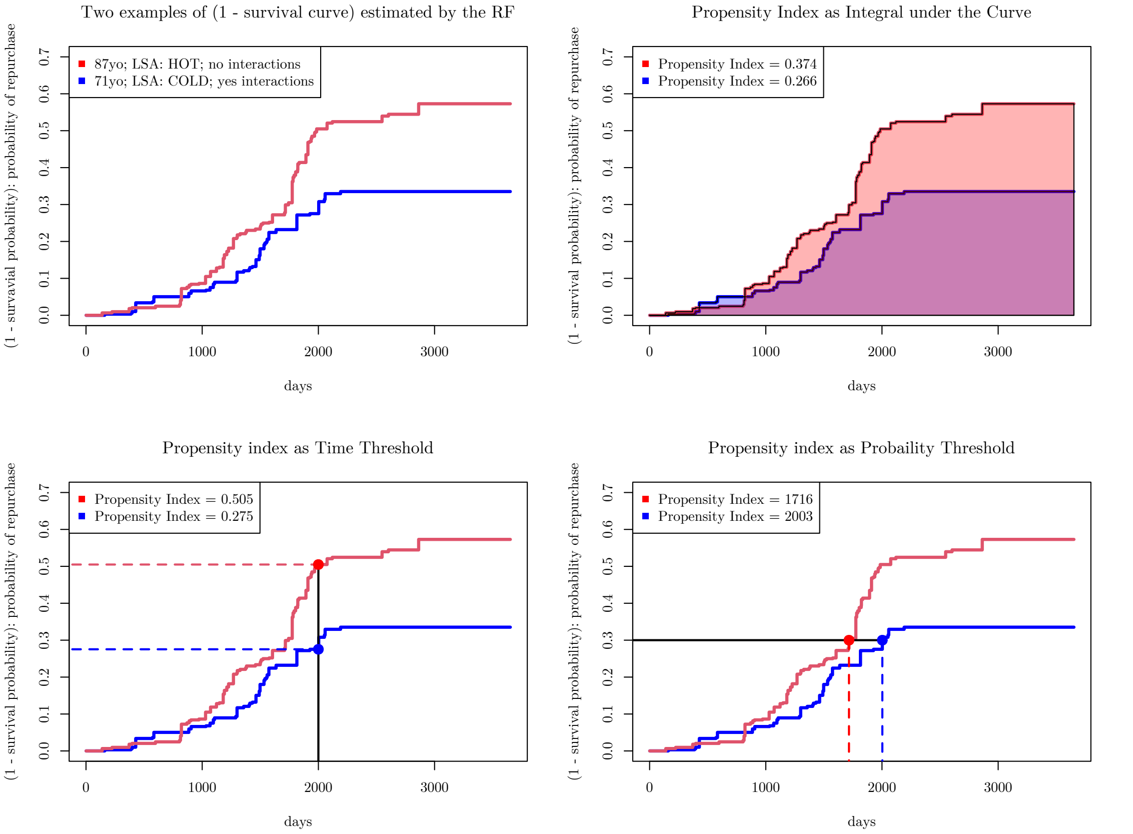


Figure 3.5: Top left: One's complement of a survival curve for two customer with different values of the explicative variables
Top right: propensity index as area under the curve
Bottom left: propensity index as cumulative probability of repurchasing after a time threshold = 2000 days
Bottom right: propensity index as time after which the probability of repurchasing is equal to a probability threshold = 30%

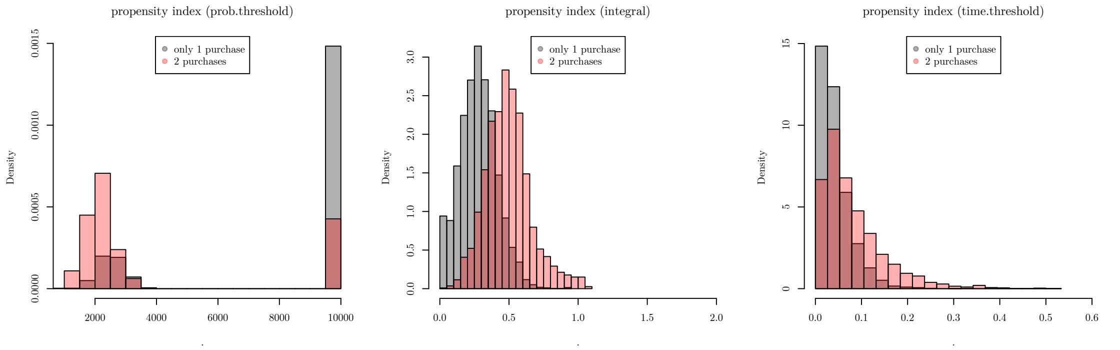


Figure 3.6: Performances of the three methods proposed to summarize the information contained in the survival curve. The x axis represent the value of the index calculated for the given method. The best method is the one that better divides the grey observations from the red ones. The one that seems to do that is the one based on the probability threshold. Note that for the first two a value must be assigned to the threshold; here we used prob.threshold = 0.3 and a time.threshold=2000. The right bar of the probability threshold distribution represent units that never reach a cumulative probability of 0.3.

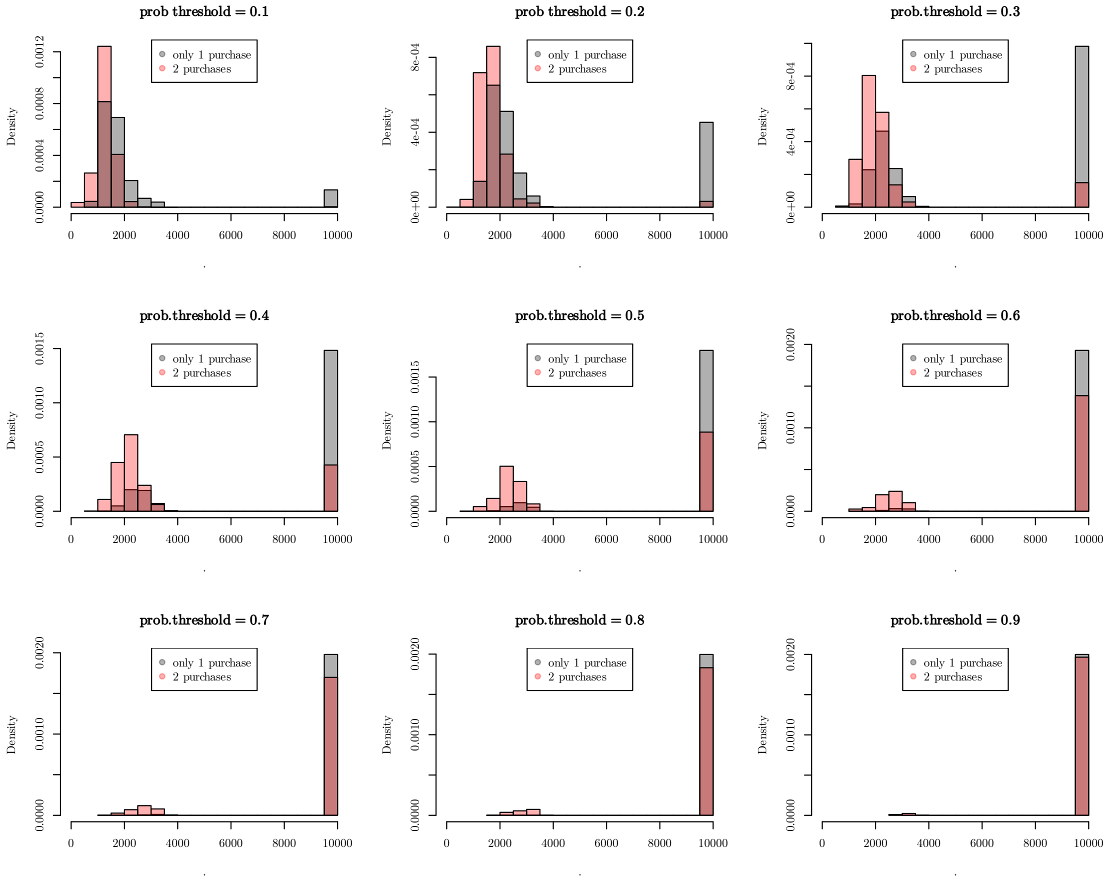


Figure 3.7: Distribution of the propensity index calculated for observations in the test set, and how it varies with different values of the probability threshold. As the threshold increases, the number of observations that do not reach that level of the probability of repurchasing increases significantly, represented by bars with values equal to 9999. We observe a transition from one extreme (top-left plot) where the observations are mainly distributed over low values, to the other extreme (bottom-right) where almost all observations are located at high values. The scope is to identify the best threshold in terms of separation between red and grey observations.

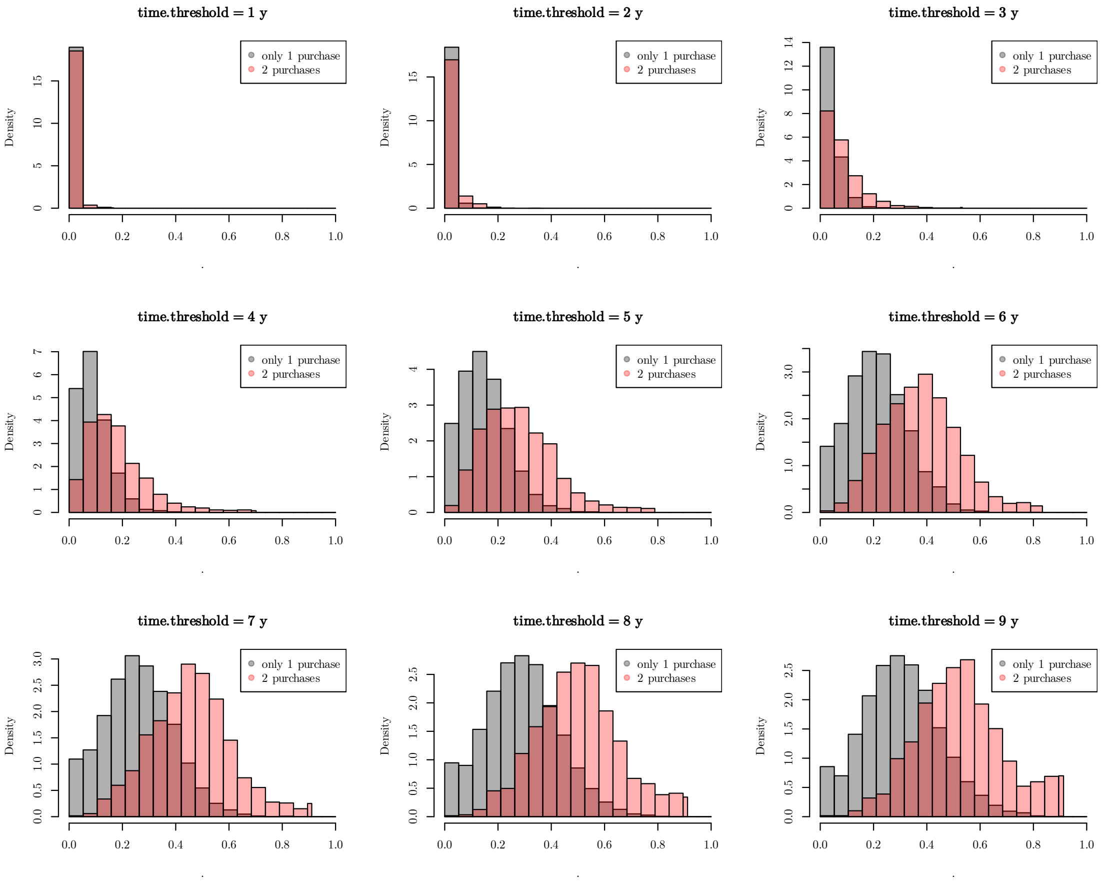


Figure 3.8: Distribution of the propensity index calculated for observations in the test set, and how it varies with different values of the probability threshold. It can be noted that for any threshold there is an overlapping area in which it is not possible to identify clients that did a repurchase from the ones that did not. The scope is to identify the best threshold in terms of separation between red and grey observations.

4
Segmentation Results

4.1 Segmentation based on Value and Propensity Indexes

In Chapter 3 we calculate a value and a propensity index. In this Section, they are used to perform a segmentation of the customer base.
As mentioned in Section 2.1, the objectives of the segmentation are two:

A Establish customers groups that exhibit homogeneity in terms of both propensity and value

B Formulate groups suitable for use in the calculation of Customer Lifetime Value (CLTV).

The ideal scenario is the one that incorporates both these characteristics.

The scope of A) is to group together clients that are similar in terms of value and propensity to repurchase, given the information available about their first 5 years of history as clients of the company. This is a descriptive task than can be performed on the basis of the features of DF5y. The information collected at early stages of customers' lives defines the segmentation variables on which basis each client is assigned to a different group.

On the other hand, to achieve objective B), future information should be taken into account as a reference, given that the groups must capture distinct behaviors concerning the CLTV cal-

culated as a combination of value and propensity over the 10 years time slot (as described in Section 2.3.1). As a result, only clients with more than five years of historical data are considered relevant for this final evaluation.
Based on this reasoning, the design and construction phase of the segmentation approach should be conducted solely on the clients contained in DF10y. In the initial stage, the framework is defined and optimized using this subset of clients. Only after finalizing the segmentation approach and ensuring its effectiveness, it can then be extended to the entire Customer Base for broader application. This approach ensures a focused and efficient development process while maintaining the reliability and generalizability of the segmentation model. Precisely, this choice enables us to assess whether the newly generated dimensions are valuable in terms of creating clusters of clients that exhibit similarity from a Customer Lifetime Value (CLTV) perspective. By first validating the approach with the clients in DF10y, we can ensure that the segmentation captures meaningful patterns and relationships related to CLTV.

To conduct the analysis, the top left plot in Figure 4.1 that illustrates the distribution of customers in the test set based on the two indexes is needed. It can be observed that the points are well spread across the space, indicating a good distribution of customers among the two dimensions.

The next step was to assess whether the two indexes were able to discriminate the units on a Customer Lifetime Value prospective. In fact, the business requirement is that the segments should be composed by clients that behave similarly in terms of CLTV. As a proxy for that we used the total amount of revenues made by the company on each customer, converted to two categories, as presented in the top right plot in Figure 4.1. The two dimensions appeared to distinguish customers who will generate high overall revenues from those who will not. Based on this evidence, it seemed reasonable to seek cuts that optimally divide the space between areas predominantly occupied by clients generating low revenues from those yielding high revenues. This aligns with the objectives defined earlier and is ideal for creating meaningful customer segments with respect to revenue potential.

4.1.1 Cutting planes

An automated method has been implemented to facilitate the process grouping together similar observations, taking advantage of the apparent separation between groups suggested by the two dimensions. This approach ensures that the segmentation process effectively leverages the inherent structure in the data to create meaningful customer groups.

It's important to note that the process of calculating the indexes was fair and avoided over fitting by evaluating their performances on data not used during training. Then, the best indexes were selected, and they were calculated on the entire Customer Base. However, the train-validation-test split used in the index construction is abandoned at this cutting planes step. In fact, the Value Index has been calculated only using the first 5 years of history of each client, so that it is not a problem to use the same observations (clients) since the validation of the cutting planes is based on the 10 years time slot information. The only issue can be raised by the fact that the Propensity Index used as target value the repurchases done in the whole 10 year time slot. This can slightly bias the approach towards over fitting. Nevertheless, this potential bias is not a major concern, as the final validation will partially rely on the performances of the CLTV model predictions on previously unseen data, which will help mitigating any overfitting issues. Thus, the overall approach remains robust and suitable for generating meaningful customer segments for CLTV calculations. The outcome of this reasoning is that the cutting planes can be estimated over a train-test split of DF10y.

Different methods were tested and the details can be found in Table 4.1. The final choice is a classification tree since it provided the best results in terms of accuracy and specificity on the test set and a comes with a business-friendly interpretation. In fact, it results in the orthogonal linear cutting planes that can be seen in Figure 4.1, left bottom plot. Te bottom right plot shows the models predictions, which result in an accuracy of 86% and a specificity of 62% on the test set. Equivalently, we can say that only 14% of the clients are miss-assigned in terms of future revenues given their value and propensity indexes. In this context, a client is miss-assigned if the value and propensity indexes indicate him to be part of a group with an expected future behaviour that he will not follow.

4.1.2 Segmentation

The natural extension of this CLTV-led segmentation to the domain of a more general segmentation of the Customer Base is the identification of groups based on cuts provided by the classification tree, as shown in Figure 4.2. These sets of customers can be labeled as "basic" (low value, low propensity), "extended" (high value, low propensity), "loyal" (low value, high propensity) and "advocate" (high value, high propensity). The groups of "loyals" and "basics" have been further divided to be more consistent with the results of the tree-based cutting planes, resulting in 6 segments as presented in Figure 4.2.

model	parameters	test accuracy	test specificity
tree	min split = 30 / min bucket = 10 / cp = 0.07 / max depth = 30	86%	62%
RF	n trees = 100 / min node size = 3 / max depth = $notconstrained$	84%	60%
SVM	kernel = $polynomial$ / degree = 3 / gamma = $\frac{1}{32000}$ / cost =0.01	84%	56%
FFNN	n layers = 3 / n nodes = $(64, 64, 3)$ / batch norm =$True$ / activation = $relu$	83%	56%

Table 4.1: Classification Models to identify groups of customers similar in terms of overall revenues

Once the segments have been identified, it is important to evaluate the properties and characteristics of the approach to ensure its effectiveness and utility in guiding marketing strategies and business decisions. Here are some key aspects to consider as presented in Chapter 1. These six segments are substantial, since all of them are populated. The profitability of each group (property of the segments identifying the customers that are profitable for the company) is ensured by the structure of the bi-dimensional space, which places the top-performing customers in the top right and the least performing ones in the bottom left. This arrangement also guarantees the differentiability of the segments, as we can reasonably expect that customers categorized as "basic" will respond differently to marketing strategies compared to those classified as "advocate". Moreover, the segments show promising actionability, as marketing strategies could potentially influence "loyal - low value" customers to exhibit behaviors similar to "loyal - middle value" customers. However, it is important to acknowledge that clients' value and propensity may not easily change, as they are based on past characteristics. Hence, the segments can be considered stable over time.
Furthermore, all the segments are reachable, as contact information for all customers is meticulously collected by the company. For the same reason, the main characteristics of each segment are also easily measurable. In conclusion, the process of selecting the different parts that compose the segmentation has been conducted rigorously, ensuring some degree of validity in the final approach. The resulting framework exhibits properties that make it a valuable tool for understanding and targeting various customer groups effectively, with the ultimate goal of enhancing customer relationships and driving business growth.

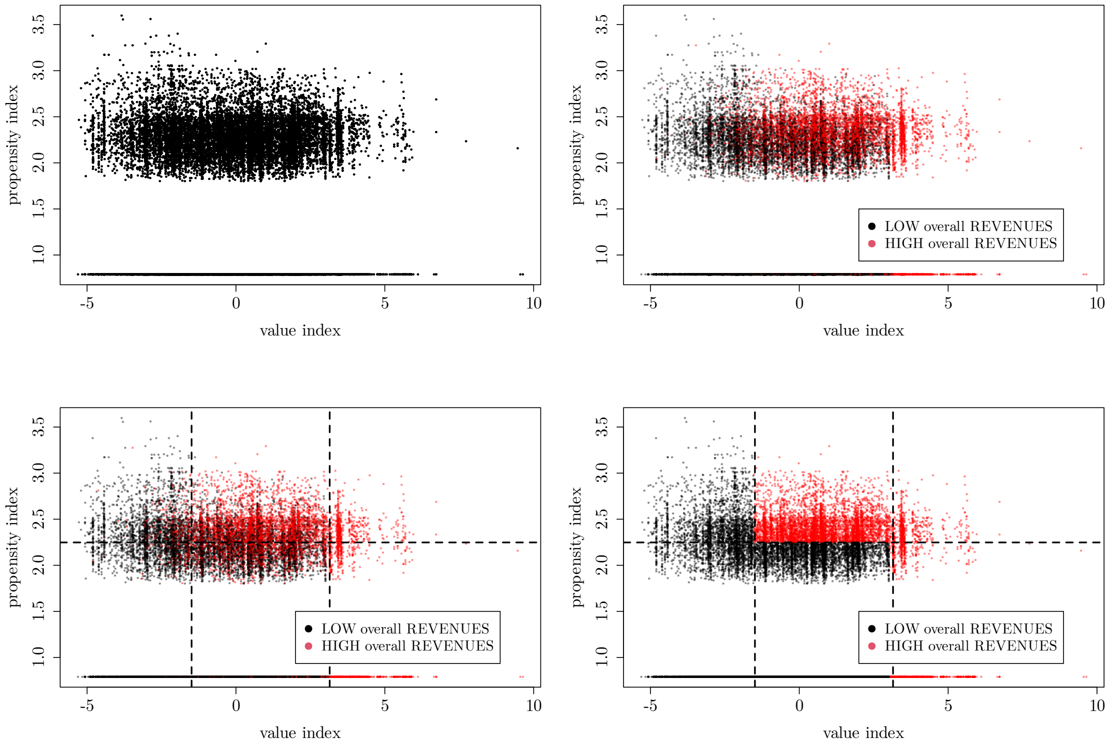


Figure 4.1: Top left: distribution of the clients in DF10y given the two indexes.
Top right: distribution of the clients in DF10y given the two indexes and the observed future revenues each client will provide to the company classified as high or low.
Bottom left: cutting planes calculated by the classification tree having as target the classes identified by the observed future revenues as 'low' and 'high'.
Bottom right: distribution of the clients in DF10y given the two indexes and the predicted future revenues as 'low' and 'high'

4.1.3 CLTV model integration

Once the segmentation approach is defined and validated according to customers grouping requirements, the next step is to assess its potential value as an addition to the Customer Lifetime Value (CLTV) calculation pipeline presented in Section 2.3.1. Figure 4.3 illustrates the existing framework (Framework 1) and three possible ways of incorporating the segmentation information extending the implemented one.
There are two layers at which the newly engineered feature can be impactful.

- In the first branch that models the number of repurchases each customer is likely to make. The segmentation can be added as a variable to the 2-step model, as shown in

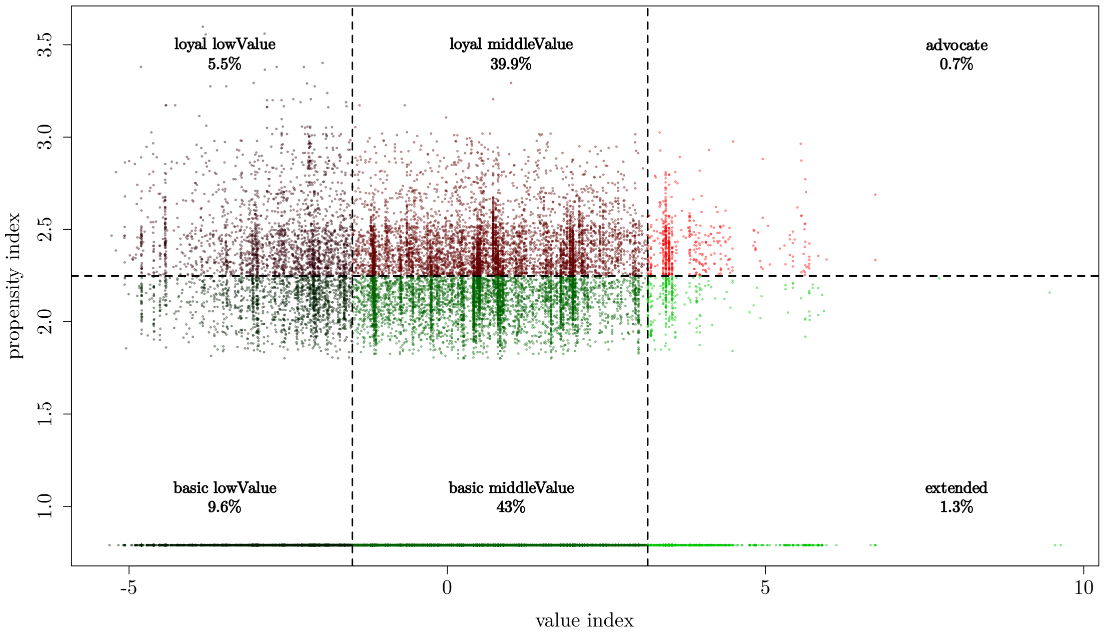


Figure 4.2: Segmentation of the Customer Base into 6 groups. The separation of the space is based on the cuts provided by the classification tree integrated by a semantic-based reasoning.

graphs 2 and 3 of Figure 4.3.

- In the second branch that provides homogeneous groups to average the repurchases' value over time, obtaining a measure of each client's potential spending. The data-driven segmentation can be used instead of the expert-based one, as demonstrated in graphs 3 and 4 of Figure 4.3.

To identify the best-performing framework, the three slightly different approaches were tested, and the evaluation was based on the mean absolute error (MAE) between the predictions and the observed values. This assessment was conducted on a dataset of clients who were not included in any of the previous steps, ensuring an out-of-sample evaluation. The baseline MAE, obtained from the implemented procedure, was found to be equal to 3183$. Among the three approaches, the one that relied on the data-driven segmentation solely for defining groups over which to calculate the average revenues (second branch) provided the worst results, yielding a MAE of 3224$. However, when the data-driven segmentation was used as an explanatory feature in the modeling branch, both Framework 2 and Framework 3 showed improved performances, with MAEs of 3117$ and 3115$, respectively.

These results highlight that the data-driven segmentation positively impacts the modeling of CLTV calculation when the segmentation information is provided to the 2-step modeling branch. On the other hand, the same data-driven segmentation was not useful for identifying groups of customers over which to calculate the average revenues, as it led to an increase in MAE. The key outcome is that the data-driven segmentation becomes valuable when it is integrated as a feature in the first branch, allowing it to influence the modeling of CLTV in a positive way. On the other hand, using the segmentation for grouping purposes in the second branch does not yield the same advantageous outcomes. Furthermore, if the segmentation information is not provided to the models, the performances suffer even more. Therefore, integrating the data-driven segmentation into the first branch is crucial for achieving better results in the CLTV calculation process.

The chosen approach for the CLTV calculation is Framework 3, where the data-driven segmentation has been included in both branches. This decision was based on the fact that Framework 3 demonstrated the lowest mean absolute error (MAE) among the tested approaches, making it the most accurate method for predicting CLTV. Frameworks 2 and 3 resulted in very similar errors, and the former was abandoned since the latter allowed the company to move away from the expert-based segmentation and rely on a more data-driven and objective method, aligning with their goal of enhancing the process. By incorporating the data-driven segmentation as a feature in the 2-step model, the company can now achieve more precise and reliable CLTV predictions for their customer base, getting a decrease of the mean absolute error equal to 68$.

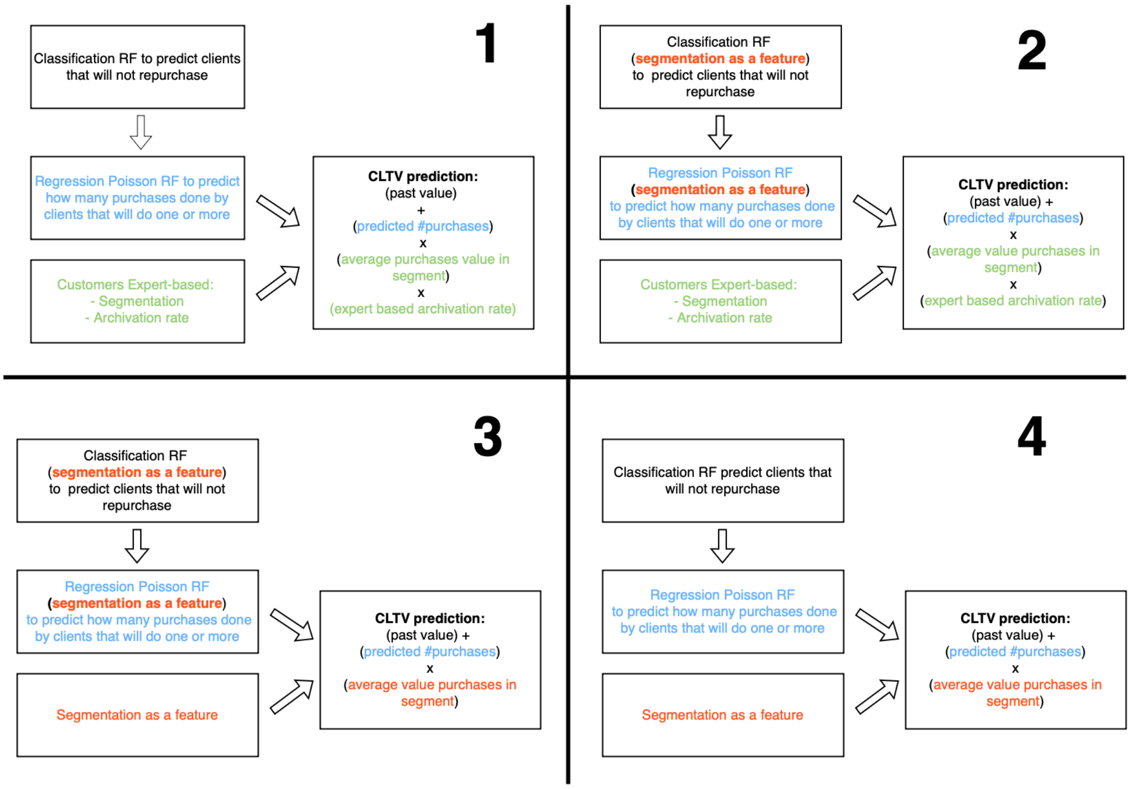

1
Classification RF to predict clients that will not repurchase
Regression Poisson RF to predict how many purchases done by clients that will do one or more
Customers Expert-based:
- Segmentation
- Archivation rate
CLTV prediction:
(past value)
+
(predicted #purchases)
x
(average purchases value in segment)
x
(expert based archivation rate)
2
Classification RF (segmentation as a feature) to predict clients that will not repurchase
Regression Poisson RF (segmentation as a feature) to predict how many purchases done by clients that will do one or more
Customers Expert-based:
- Segmentation
- Archivation rate
CLTV prediction:
(past value) +
(predicted #purchases)
x
(average value purchases in segment)
x
(expert based archivation rate)
3
Classification RF (segmentation as a feature) to predict clients that will not repurchase
Regression Poisson RF (segmentation as a feature) to predict how many purchases done by clients that will do one or more
Segmentation as a feature
CLTV prediction:
(past value) +
(predicted #purchases)
x
(average value purchases in segment)
4
Classification RF predict clients that will not repurchase
Regression Poisson RF to predict how many purchases done by clients that will do one or more
Segmentation as a feature
CLTV prediction:
(past value) +
(predicted #purchases)
x
(average value purchases in segment)

Figure 4.3: Different frameworks for CLTV prediction. **Framework 1** shows the one currently implemented that does not consider the data driven segmentation. **Framework 2** uses the data driven segmentation as a feature in the model that predicts the number of purchases. **Framework 3** as the previous, but also substitutes the expert based segmentation with the data driven one. **Framework 4** uses the segmentation only as a substitute to the expert based one.

5
Mixed Effect Survival Forest

5.1 Introduction to Mixed Effects

The selected survival model used as the basic structure to calculate the propensity index did not yield exceptionally high performance, with a C-index of approximately 0.68. To explore potential ways to enhance this result, we could leverage additional information available in the data, such as the city of residence for each client. This feature was not included in the previous sections since it has 487 distinct levels. The high number of levels poses a challenge in the model's computation and may result in inefficiencies and difficulties in handling such a large categorical variable effectively.

Indeed, it is unfortunate to ignore the potential informative power of the city of residence feature. Understanding if this regressor can identify groups of customers with distinct repurchase probabilities over time is an intriguing research question. The segmentation approach could be improved by determining whether there are unexplained group-specific variations that cannot be adequately accounted for by the other features in the model. By incorporating the city of residence as a component in the survival model, we can explore its impact on the individual customer's behavior while considering the overall variations in repurchase probabilities across different cities. This approach allows us to capture any city-specific patterns that contribute to differences in repurchasing behavior among customers. A way to encapsulate it in the model could be as random effect.

Regrettably, after conducting a literature review, it became evident that the extension of mixed effects to the domain of survival random forests has not been addressed yet.

The aim of Chapter 5 is to propose a starting point for the implementation of mixed effects survival forests.

5.2 Mixed Effects Models Overview

Mixed effects models [42] are statistical models that can be expressed as a composition of fixed and random effects. Fixed effects are represented by parameters associated to the entire population, while random effects are associated with features generated at random at an individual level. The scope of such a combination is to model the relationship between a response variable and features (fixed effects) grouped by other classification factors (random effects), when independence between observations cannot be guaranteed due to the hierarchical structure of the data. The covariance induced by the underlying grouping is handled by assigning common random effects to statistical units that have the same level of classification factor. For example, in the case of our analysis the fixed effects are represented by the features on which the model was trained. The hierarchical structure of the data is given by the fact that customers that share the same city of residence may tend to behave in correlated ways, which would contradict the assumption of independence. The random effect refers to the impact of the city of residence on the target variable. The idea is that there could be a systematic difference between groups that is related to group membership itself. Such difference is modeled through a random variable.

The aim of this Chapter is to extend this approach to the domain of survival forests.
The first step has been a review of how mixed effects are implemented in survival analysis, specifically to the Cox model (Section 5.3).

5.3 Cox Mixed Effects

The extension of the Cox model to one mixed effect [43] is given by:

$$h(t|X) = h_0(t) \cdot e^{X\beta + \mathbf{Zb}}, \;\; b \sim G(0, \Sigma(\theta)) \tag{5.1}$$

where $h_0(t)$ is an unspecified baseline hazard function, $X_{N\times p}$ and $Z_{N\times q}$ are the design matrices for the fixed and random effect, respectively, $\beta_{p\times 1}$ is the vector of fixed effects coefficients and $b_{q\times 1}$ is the vector of random effects coefficients. The random effects distribution G is modeled as Gaussian with mean zero and a variance matrix $\sum$, which in turn depends a vector of parameters θ. From an interpretation perspective, $\exp(X\beta)$ represents the multiplicative effect of the population's features on the hazard ratio, while $\exp(Zb)$ represents the variation of each group in regards to the random effect.

5.3.1 Cox Mixed Effects Model Estimation

The estimation procedure of a Cox mixed effects model' parameters relies on partial likelihood, similar to what was mentioned in 3.2.1.4.

The partial log-likelihood (PLL) is defined as:

$$PLL = \sum_{i=1}^{N} \int_{0}^{\infty} [Y_i(t)\eta_i(t) - \log(\sum_{j}(t)e^{\eta_j(t)})\ dt] \tag{5.2}$$

where $\eta_i(t) = X_i(t)\beta + Z_i(t)b$ is the linear score for subject i at time t, $Y_i(t) = 1$ if subject i is still under observation at time t, $Y_i(t) = 0$ otherwise.

The Maximum Likelihood estimation is calculated through the joint maximization over β and θ. There is not close form of the solution of such maximization problem, so that expected maximization algorithms has to be applied as in [44].

5.4 Mixed Effects Survival Forest

As a starting point, the focus has been limited to the case of one single random feature, drawing inspiration from how the mixed effects Cox model combines survival curves estimation together with the random effects impact. The correlation between groups is modeled as an exponentially transformed multiplicative factor on the hazard function, meaning that various realizations of the random effects proportionally modify the hazard function itself as shown in Figure 5.1.

5.4.1 Model Specification

The same structure of the mixed effects Cox model can be proposed for survival forests. As detailed in Algorithm 3.2, the survival forest provides estimations of the cumulative hazard function $H(t_k|x_i)$ for $k = 0, 1, 2, ..., K$ given the realizations of the explicative variables $x_i =$

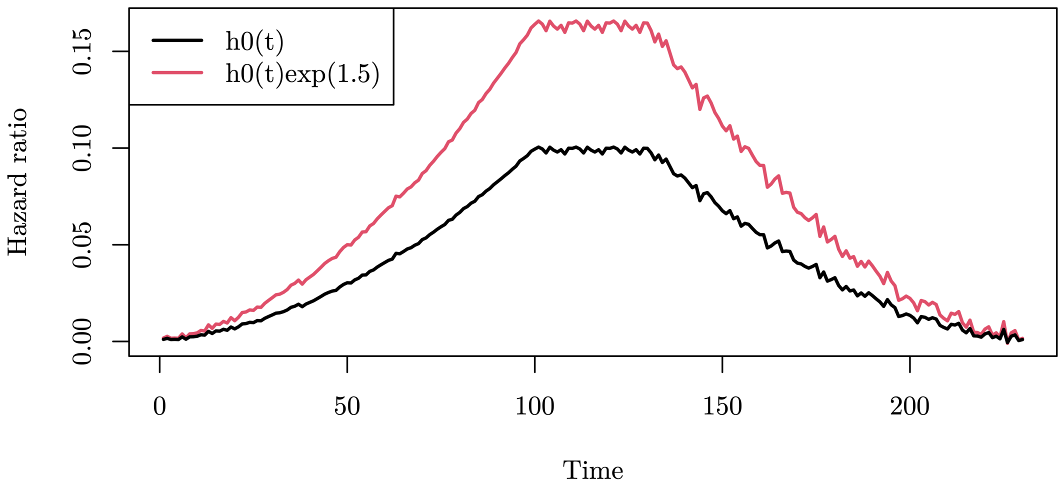


Figure 5.1: Example of the impact of the random effect on the hazard function. The black line represents a baseline hazard function. If we add a random effected which impact is estimated as 1.5, the consequential hazard function is the red one.

$(x_{i1}\ , ...,\ x_{ip})$. The equation that defines the hazard function $h_i(t_k)$ for a client i given their vector of observations x_i at a time $t_k \in T$ is

$$h(t_k|x_i) = f_i(t_j) = \quad H(t_{k+1}|x_i) - H(t_k|x_i) \tag{5.3}$$

where $f_i\ :\ T\ \rightarrow\ [0, 1]$ is a non-parametric function that maps the set of the discrete times $T = \{t_k\ :\ k \in \{0, 1, ..., K\}\}$ to the related hazard rate for the i-th statistical unit.

Similarly to what was proposed for the Cox model, the exponential transformation of the random effect parameter can be inserted in the model as a multiplicative factor for the hazard function. Then, the equation that defines the model becomes:

$$f_i'(t) = f_i(t) \cdot e^{z_i b} \tag{5.4}$$

where $f_i(t)$ is directly referable to the regular survival forest modeling, $b \sim N_q(0, \Sigma(\theta))$ is the random effect modelling parameter, $z_i = (z_{i1},\ z_{i2},\ ...,\ z_{iq})$ is the covariate design vector. Note that by definition the covariance matrix of the random effect parameter is free to assume any shape given θ vector of parameters.

In our context, $q = 487$ (number of distinct cities of residence) and

$$z_{ik} = \begin{cases} 1 \; if \; client \; lives \; in \; city \; k \\ 0 \; otherwise \end{cases} \tag{5.5}$$

It follows that e^{b_k} represents the multiplicative effect of the city k on the hazard function for $k = 1, ...q$.

After defining the model, the next step is to identify a procedure to estimate it.

5.4.2 Model Estimation

The fitting of a survival forest is not trivial. We have identified two components that need to be considered, as indicated by the form of (5.4). On one hand, there is the fixed effect dependent part $f_i(t)$, and on the other hand, the mixed effect dependent part $e^{z_i b}$. At an higher level, the crucial aspect is that these two components should be independent of each other. In other words, they need to model different factors that contribute to the survival behaviour of each statistical unit. The former models the information that can be extracted from the fixed effects variables, while the latter captures the variable that defines the hierarchical structure of the data, which we aim to incorporate as a random effect.

The distinction between the two components indicates the need of a procedure that could involve estimating one component using data that has been refined by the effect modeled by the other component. The natural implementation of such an idea is an iterative framework in which at each step each component is estimated on the residuals deriving from the other one, as presented in Algorithm 5.1.

Algorithm 5.1 Iterative framework on the residuals

```
Data: $[X_{N\times p}|Z_{N\times q}|t_{N\times 1}|\delta_{N\times 1}]$
define the residuals calculated over the $Data$ given a survival model
  $resid(Data \mid surv\_model), resid: \; R^{N\times(p+q+1+1)} \rightarrow R^{N\times(p+q+1+1)}$
  $D^{(0)} \leftarrow [X_{N\times p}|Z_{N\times q}|t_{N\times 1}|\delta_{N\times 1}]$
  estimate the fixed-effect component $\hat{f}(t)$ on $D^{(0)}$
  $D'^{(0)} \leftarrow resid(D^{(0)} \mid \hat{f})$
  estimate the random-effect component $\hat{b}$ on $D'^{(0)}$
  $D^{(1)} = resid(D'^{(0)} \mid e^{Z^{(i)}\hat{b}})$
  set $i \leftarrow 1$
  while stopping criterion do
    update the fixed-effect component $\hat{f}(t)$ on $X^{(i)}, t^{(i)}, \delta^{(i)}$
    $D'^{(i)} \leftarrow resid(D^{(i)} \mid \hat{f})$
    update the random-effect component $\hat{b}$ on $X'^{(i)}, t'^{(i)}, \delta'^{(i)}$
    $D^{(i+1)} \leftarrow resid(D'^{(i)} \mid e^{Z^{(i)}\hat{b}})$
    $i++$
end
```

Given this theoretical setup, prior to thinking about the fitting of two components, we need to identify a way to filter out their effects.

5.4.2.1 Expanded-Baseline Iterative Approach

The first idea has been to define a Cox structure to model the random effects and then start the iterations of the survival forest's fitting on its baseline hazard estimate. This approach did not lead to any result but it has been instructive to deep into the behaviour of the framework. Thanks to the insides got in this Section, we have been able to develop the final approach described in Section 5.4.2.3.

The idea was to start with the following steps:

- Estimate the random effect $\hat{b}$.
- Estimate a non-parametric baseline hazard function $\hat{h}_{baseline}(t)$ corrected for the influence of the random effect.

The random effect can be estimated as in a mixed effects Cox model [45], providing the vector of estimates $\hat{b}$. Two options are available: 1) estimating solely the random effect using the Cox model, temporarily treating $\hat{h}_{baseline}(t) = \hat{h}_0(t)$ as a 'container' for the fixed effect; 2) estimating both the fixed effects and the random effect with the Cox model, effectively removing their influence from $\hat{h}_0(t)$, since $\hat{h}_{baseline}(t) = \hat{h}_0(t)e^{x_i\hat{\beta}}$.

ONLY RANDOM EFFECT BY COX MODEL

If we opt to estimate only the random effect using the Cox model, we obtain:

$$\hat{h}(t|x_i, z_i) = \hat{h}_0(t) \cdot e^{z_i\hat{b}} \tag{5.6}$$

We can derive $\hat{h}_{baseline}(t) = \hat{h}_0(t)$ through the Kaplan-Meier estimator. This is equivalent to use an average curve for each observation.

The fitted hazard function $\hat{h}_0(t)$ is a discrete function that assumes values in all the K unique time stamps at which at least one statistical unit has observed the event:

$$\hat{h}_0 \ : \ \{t_1, t_2, ..., t_K\} \rightarrow [0, 1] \tag{5.7}$$

The censoring information is encapsulated in $\hat{h}_0(t)$ by mean of the fitting procedure that discriminates between the number of events (uncensored observations) and individual at risk (that can be both censored and uncensored).

To obtain the aimed form as in (5.4), $\hat{h}_0(t)$ has to be substituted with the survival forest estimate for i. To do that, we would need to fit the survival forest over the data depurated by the $e^{\hat{b}z_i}$ factor. Considering that $\hat{h}_i(t)$ contains both the fixed and random components effects, it seemed reasonable to start delving into how to model $\hat{h}_0(t)$ with a survival forest.

To accomplish this, a strategy has been adopted that views the Kaplan-Meier estimator as a data-transforming engine. Let's consider one statistical unit i (for $i = 1, ..., N$) with the respective pair of censoring and time-to-event information (δ_i, t_i) and explicative variables (x_i, z_i). The data about i is mapped to K distinct hazard values by $\hat{h}_0(t_k)$. A graph representing this concept is provided in output number 1 of Figure 5.2. This can be defined as the baseline expansion of the data, since it is an augmentation of the data given the baseline hazard function. Note that the censoring information is intrinsically included in the combination of (t_k, h_k), and given the fact that $\hat{h}_0(t)$ is the same for each statistical units, the columns t_k and h_k contain repeated values.

RANDOM AND FIXED EFFECTS BY COX MODEL

Another approach is to extend the Cox model estimating also the fixed effects:

$$\hat{h}_i(t) = \hat{h}_0(t)e^{x_i\hat{\beta}} \cdot e^{z_i\hat{b}} = \hat{h}_{baseline}(t) \cdot e^{z_i\hat{b}} \tag{5.8}$$

where $\hat{h}_{baseline}(t) = \hat{h}_0(t)e^{x_i\hat{\beta}}$. This leads to the structure provided in output number 2 of Figure 5.2, where each statistical unit i has different h_k because their features x_i impact the data transforming engine.

To obtain the aimed form as in (5.4), $\hat{h}_{baseline}$ has to be substituted with the survival forest estimate for i.

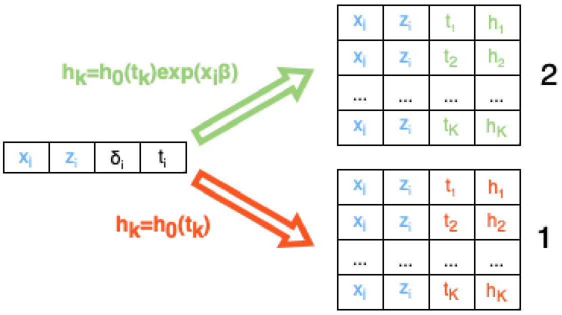


Figure 5.2: Output 1: baseline expansion of a row of the original data through a discrete hazard function $h_0(t)$. **Output 2**: baseline expansion of a row of the original data through a discrete hazard function $h_0(t)$ and the fixed effects of the Cox model $\exp(x_i\beta)$.

It is possible to estimate the survival forest on the two datasets presented so far, with some considerations:

- Each row of both the expanded datasets consists in the explicative variables (x_i, z_i) combined together with all the distinct K times and the respective hazard values. The application of these procedures results in K new rows for each of the original ones, i.e., $N \cdot K$ new rows. Note that if N is large, K tends to be large, since the number of unique time stamps generally tends to increase with the number of observations. For example, in the problem at hand $N = 30881$ and $K = 3124$ unique days at which at least one event has been observed. The result is a dataframe with $30881 \cdot 3124 = 96472244$ rows, and this number scales quickly. To avoid encountering computational issues, the unique times can be aggregated, reducing K. For example, if we consider monthly instead of daily data, $K = 121$. Another approach would be to filter out all time stamps at which the probability of observing the event is almost 0, setting a threshold under which the respective rows (t_k, h_k) are not included in the dataset.

- The observations are highly correlated among themselves. Specifically, the expanded data contains N groups, each comprising K rows that correspond to the behaviour of the same statistical unit at different time stamps.
- For each h_k, a corresponding p_k can be linked, signifying the probability of observing the event at time t_k. These values are employed as weights for assigning sampling probabilities in the random selection of observations to retain within the model during each tree-building process, so that each statistical units have a probability of being selected at each iteration equal to 1 at most. See (3.10) for details about p_k.

These two survival forests represent an initial endeavor to create distinct survival curves for various statistical units. However, their current state is not yet optimal due to being fitted using the time-to-event information derived from the non-parametric estimation of the average survival curve. The censoring details are not explicitly present, since they have been incorporated into the estimation of the h_k column. It's important to note that the random forest doesn't directly utilize this feature, but rather leverages a transformed version of it (p_k) as weights.
This insight sheds light on how these models mimic:

1. the behaviour of the non parametric estimator $\hat{h}_0(t)$, if the survival forest is estimated on the data represented in Figure 5.2, Graph 1.
2. the behaviour of a regular Cox model depurated by the random effect contribution, if the survival forest is estimated on the data represented in Figure 5.2, Graph 2.

Considering the two options, the second approach appears to be more logical. In this scenario, both the random and fixed effects are estimated by the Cox model initially, and the survival forest subsequently models that effect. Then, none of this approaches provides a good estimation yet, as we aim for the random forest to act as an estimator of the inherent information within the data itself, rather than modeling it based on the Cox model's or the Kaplan-Meier estimator's behavior.
The next step would be to re-estimate the random effect on the data from which we would

remove the fixed effect modelled by the survival forest. Regrettably, what done so far has not provided a solution to the problem of removing the fixed effect from the data, due to the lack of a framework to subtract from the target data (that has the shape of a bi-dimensional vector (δ_i, t_i)) the predictions of the survival forest (that have the shape of a function $S : T \to [0, 1]$). This problem will be addressed in the next Section.

To sum up, our current progress involves fitting the components of the model described by (5.4). We were able to estimate the mixed effects part as in a mixed effect Cox model. Additionally, we estimated the fixed effect component using a survival forest, although its potential is not fully realized due to the absence of a method to extract the information modeled by the survival forest itself from the data.
We have to find a practical method to define residuals that can help us to isolate these effects.

5.4.2.2 Survival Residuals

In regression tasks, the concept of residuals is generally reducible to the difference between the model's predictions and the observed target values. In survival analysis, the target value and the prediction have different dimensions. The former is a bi-dimensional vector (t, δ) composed by the time-to-event and the censoring information. The latter is a discrete function $S : \mathbb{R} \to [0, 1]$ that maps distinct time stamps to the probability of event occurrence. Thus, the identification of a definition of residuals is not immediate. In light of this, we conducted a survey of various survival residuals to explore if any could align well with our requirements. The scope is to find a definition of residuals that is able to remove from the fitting data the effect of the fitted model. It must be taken into account that all the residuals types were initially proposed for the Cox model: this implies that not all of them can be extended to survival forests.

COX-SNELL RESIDUALS
The Cox-Snell residuals [46] are defined as the negative of the natural logarithm of the survival probability for each observation at a certain time t:

$$r_i(t) = -\log(\hat{S}_i(t)) \tag{5.9}$$

where $\hat{S}_i(t)$ is the estimated survival function.
When not explicitly indicated, we refer at the Cox-Snell residual r_i for the *i-th* statistical unit as the residuals calculated at the observed time-to-event ($r_i = r_i(t_i)$).

Their main property is that under the model's assumptions they follow an exponential distribution. For this reason they are widely used to assess model's goodness.

MODIFIED COX-SNELL RESIDUALS

A modification of Cox-Snell residuals for censoring information handling was proposed by [47].

Given t'_i the time at which the $i\text{-}th$ observation has been censored and t_i the unknown actual survival time, the Cox-Snell residual at time t'_i for this individual is:

$$r_i(t'_i) = -\log(\hat{S}_i(t'_i)) = \hat{H}_i(t'_i) \tag{5.10}$$

Since $H(t)$ is a cumulative function, it increases over time. It follows that $r_i(t'_i) < r_i(t_i)$, i.e., the residuals at the actual time-to-event would be greater then the residuals at the censoring time, that is inconsistent with any general definition of residuals.

To avoid this, a modification can be performed by adding the censoring information:

$$r'_i = \begin{cases} r_i \;\; if\; i\; uncesored \\ r_i + 1 \;\; otherwise \end{cases} \tag{5.11}$$

that is equivalent to $r'_i = r_i + \delta_i$.

MARTINGALE RESIDUALS

The mean of the modified Cox-Snell residuals as defined in (5.11) is equal to 1 for uncensored observations. To relocate it to 0, preserving a generally desired property of residuals, the following adjustment can be adopted, as presented in [48]:

$$r_{Mi} = \delta_i - r_i \tag{5.12}$$

One way to consider them is as the difference between the observed number of deaths until the observed time-to-event for the $i\text{-}th$ statistical unit and the model prediction ($r_i(t_i) = \hat{H}_i(t_i)$).

DEVIANCE RESIDUALS

The deviance residuals have been proposed by (5.11) as:

$$r_{Di} = sign(r_{Mi})[-2r_{Mi} + \delta_i \log(\delta_i - r_{Mi})]^{\frac{1}{2}} \tag{5.13}$$

These residuals are symmetrically distributed around 0.

OTHERS

There are two other families of residuals that can not be adapted to survival forests because of their link to the likelihood function. These are the Schoenfeld [49] and the score residuals [50].

All these approaches provide a measure that is on the scale of the censoring information.

5.4.2.3 Adding the Martingale residuals information

Our final approach focuses on incorporating the concept of residuals at the censoring level. The prior approach (Section 5.4.2.1) led to a survival forest that predominantly modeled the fixed effects of a Cox model rather than capturing the essence of the actual data. The intuition behind this Section is that Martingale residuals can be used to deprive the data from random and fixed components.

As seen in Section 5.4.2.2, Martingale residuals can be interpreted as the difference between the observed number of events δ_i (that can be equal to 0 or 1) and the estimated number of events r_i (Cox-Snell residuals). As mentioned, Martingale residuals are usually calculated at the observed time-to-event t_i. Nevertheless, it is possible to extend them to other time stamps $t_k \in T$:

$$r_{Mi}(t_k) = \delta_i(t_k) - r_i(t_k) = \delta_i(t_k) - \hat{H}_i(t_k) \tag{5.14}$$

where $\hat{H}_i(t)$ is the cumulative hazard function estimated by the model which effect we aim to extract from the data, $\delta_i(t_k) = 0$ if unit i has not observed the event at time t_k, $\delta_i(t_k) = 1$ if they have. In practice, given the observed time-to-event t_i for unit i, the only available information about censoring is $\delta_i(t_i)$. Censoring at other time stamps can be calculated as:

$$\delta_i(t_k) = \begin{cases} \delta_i(t_i) \; if \; t_k \geq t_i \\ 0 \; if \; t_k < t_i \end{cases} \tag{5.15}$$

This is equivalent to the self-evident statement that the event has not been observed until it has been observed.

An adjustment is necessary for censored units due to the fact that their actual time-to-event remains unknown. To illustrate, let's consider a censored $i\text{-}th$ unit, characterized by the data $(x_i, z_i, \delta_i = 0, t_i)$. Survival models operate under the assumption that an undisclosed time $t_j > t_i$ exists at which the event will be observed. The structure proposed in (5.15) would erroneously imply $\delta_i(t_j) = 0$. This is an incorrect inference. Consequently, it becomes imperative for all time stamps beyond the corresponding time-to-event of censored units to be excluded.

The structure of 5.14 leads to an interpretation of $r_{Mi}(t_k)$ as a metric to say how 'close' models' predictions are to reality for the $i\text{-}th$ unit at time t_k. Let's consider an example to better understand it, knowing that $\delta_i(t_k)$ can only take two possible values, 0 and 1:

i SUCH AS $\delta_i(t_k) = 1$
If $r_{i,model_1}(t_k) = 0.2$ and $r_{i,model_2}(t_k) = 0.7$, then $model_2$ would be considered 'closer' to the observed value than $model_1$, since the Martingale residuals would be $r_{Mi,model_1}(t_k) = 1 - 0.2 = 0.8$ and $r_{Mi,model_2}(t_k) = 1 - 0.7 = 0.3$.

i SUCH AS $\delta_i(t_k) = 0$
If $r_{i,model_1}(t_k) = 0.2$ and $r_{i,model_2}(t_k) = 0.7$, then $model_1$ would be considered 'closer' to the observed value than $model_2$, since the Martingale residuals would be $r_{Mi,model_1}(t_k) = 0 - 0.2 = -0.2$ and $r_{Mi,model_2}(t_k) = 0 - 0.7 = -0.7$.

Given a threshold ϵ that defines if a model's prediction is close enough to the actual value, three scenarios can raise:

1. The model prediction is 'close enough' to the observed value, i.e., $|r_{Mi}(t_k)| \leq \epsilon$.

2. The model prediction is lower than the observed value, i.e., $r_{Mi}(t_k) > \epsilon$.

3. The model prediction is greater than the observed value, i.e., $r_{Mi}(t_k) < -\epsilon$.

Since Martingale residuals can assume values in $(-\infty, 1)$, not being symmetric around 0, cases $2.$ and $3.$ have to be treated differently. It can be handled truncating $r_{Mi}(t_k)$ to -1:

$$r_{Mi}(t_k) = \begin{cases} r_{Mi}(t_k)\ if\ r_{Mi}(t_k) >= -1 \\ -1\ if\ r_{Mi}(t_k) < -1 \end{cases} \tag{5.16}$$

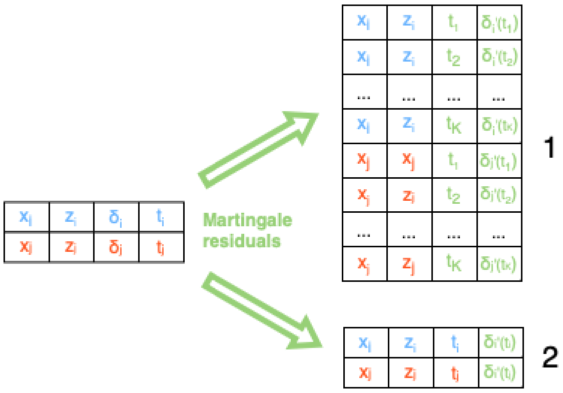
that is equivalent to truncate the prediction to 1 when the model predicts more than 1 event ($r_{Mi} = 0 - r_i = 0 - \hat{H_i(t)}$) for censored observations. For uncensored events the model's predictions are truncated to 2. This approach aligns well with the specific nature of the problem, which involves a maximum of one event occurrence.

Then, the aim is to make use of Martingale residuals to adjust the censoring information by the random effects, by defining a new censoring indicator $\delta_i(t_k)$, to provide depurated data to the random forest. This last requirement implies that we need to obtain values equal to 0 or 1, since the censoring information $\delta_i'(t_k)$ provided to a model for its fitting needs to have this dichotomous shape. A first approach could be:

$$\delta_i'(t_k) = \begin{cases} 0\ if\ |r_{Mi}(t_k)| \leq \epsilon \\ 1\ if\ r_{Mi}(t_k) > \epsilon \end{cases} \tag{5.17}$$

This approach outputs a new dataset over which to fit the survival forest using the newly generated censoring information. Thanks to the broad definition given in (5.17), that does not constrain the choice to a specific time stamps, the final dataset can contain both the rows related to the actual time-to-event t_i and the 'expanded' ones, related to all the time stamps t_k for $k = 1, ..., K$. Consequentially, two options are given: maintain all the time stamps for each observation or only the one referring to the observed time-to-event. The two possibilities are schematized in Figure 5.3. The final choice fell on the latter, to avoid duplication of explicative features generated by the same rows.

To sum up, we have provided a method to filter out from the data the effect modeled by survival forests and mixed effects Cox models. It can be implemented in an iterative flavour as described in Algorithm 5.2, where the convergence criterion is determined by the random effect estimates as proposed by [51]. The goal is to achieve a state where the model is composed of two distinct and independent parts, $\hat{f}_i(t)$ and $e^{z_i\hat{b}}$, both fitted on data that has been fully cleansed of the influences of random and fixed effects, respectively. Nevertheless, ensuring convergence is not a guaranteed outcome of this process. It's conceivable that both components might retain some residual influence of the other, causing them to perpetually filter out each



x_i	z_i	δ_i	t_i
x_j	z_j	δ_j	t_j

x_i	z_i	t_1	$\delta_i'(t_1)$
x_i	z_i	t_2	$\delta_i'(t_2)$
...	...	...	...
x_i	z_i	t_K	$\delta_i'(t_K)$
x_j	x_j	t_1	$\delta_j'(t_1)$
x_j	z_j	t_2	$\delta_j'(t_2)$
...	...	...	...
x_j	z_j	t_K	$\delta_j'(t_K)$

x_i	z_i	t_i	$\delta_i'(t_i)$
x_j	z_j	t_j	$\delta_j'(t_j)$

Figure 5.3: Graph 1: training dataset as expansion of a row of the original data through the Martingale residuals for all the possible time stamps.
Graph 2: training dataset as the original one with a new column representing the new censoring information calculated trough the Martingale residuals.

other's effects from the data. This could lead to an oscillating pattern without achieving a conclusive end point. In other words, the iterative refinement might not always culminate in a perfectly separated state of the two components.

Algorithm 5.2 Mixed effect survival forest

```
Data: $[X_{N\times p}|Z_{N\times q}|t_{N\times 1}|\delta_{N\times 1}]$
$D \leftarrow [X_{N\times p}|Z_{N\times q}|t_{N\times 1}|\delta_{N\times 1}]$
 $\hat{b} \leftarrow 0_{q\times 1}$
 set $threshold$
 while continue do
     $b_{old} \leftarrow \hat{b}$
     estimate $\hat{h}_0(t)$ on $D$ through the Kaplan-Meier method
     estimate a random effect Cox model $\hat{h}(t|z_i) = \hat{h}_0(t)e^{z_i\hat{b}}$ on $D$
     transform $\hat{h}(t|z_i)$ into the cumulative hazard $\hat{H}(t|z_i) = \hat{H}_0(t) \cdot e^{z_i\hat{b}}$
     calculate the Martingale residuals $r_{Mi} = \delta_i - \hat{H}(t_i|z_i), \forall i \in 1, ..., N$
     calculate the depurated censoring information $\delta_i'(t_k)$ as in (5.17) , $\forall i \in 1, ..., N$
     $D' \leftarrow [X_{N\times p}|Z_{N\times q}|t_{N\times 1}|\boldsymbol{\delta'_{N\times 1}}]$
     fit a survival forest on $D'$
     calculate $\hat{H}(t_i|x_i), \forall i \in 1, ..., N$ given by the survival forest
     calculate the Martingale residuals $r_{Mi} = \delta_i - \hat{H}(t_i|x_i), \forall i \in 1, ..., N$
     calculate the depurated censoring information $\delta_i'(t_k)$ as in (5.17) $\forall i \in 1, ..., N$
     $D \leftarrow [X_{N\times p}|Z_{N\times q}|t_{N\times 1}|\boldsymbol{\delta'_{N\times 1}}]$
     $M \leftarrow \max_{j=1,...,q}(|b_{old} - \hat{b}|)$
     $index \leftarrow \arg\max_{j=1,...,q}(|b_{old} - \hat{b}|)$
     $\Delta = \frac{M}{\hat{b}_{index}}$
     if $\Delta < threshold$ then
         continue $\leftarrow FALSE$
     end
end
```

5.4.2.4 Mixed effect Survival Forest: Case Study

The procedure presented in Algorithm 5.2 has been applied to fit a mixed effect survival forest on our data. The implemented R functions to do it can be found in the Appendix.

The results show that Algorithm 5.2 does not converge according to the defined criterion.

Memory usage	Avg time one step	Total time	CPU cores usage %
3.9 GB	425 s	5.8 h	90%

Table 5.1: Training computational cost

CPU	RAM	GPU	R version	Ranger version [52]
Apple M1	16 GB	None	4.3.1	0.15.1

Table 5.2: Environment specifications

Figure 5.4 shows that the measure of convergency is never lower than the set threshold nor presents a descending trend.

The model reached after the maximum number of iterations provides a C-index equals to 0.652, that is worse than the regular survival forest.

Table 5.1 shows the computational cost faced to train the model for 50 iterations on the machine which specifications are given 5.2. It's evident that this procedure consumes an excessive amount of time, rendering it impractical for practical use. Subsequent research efforts should also be directed towards improving efficiency of each step.

In conclusion, this approach is not preferable to the regular random forest for the problem at hand.

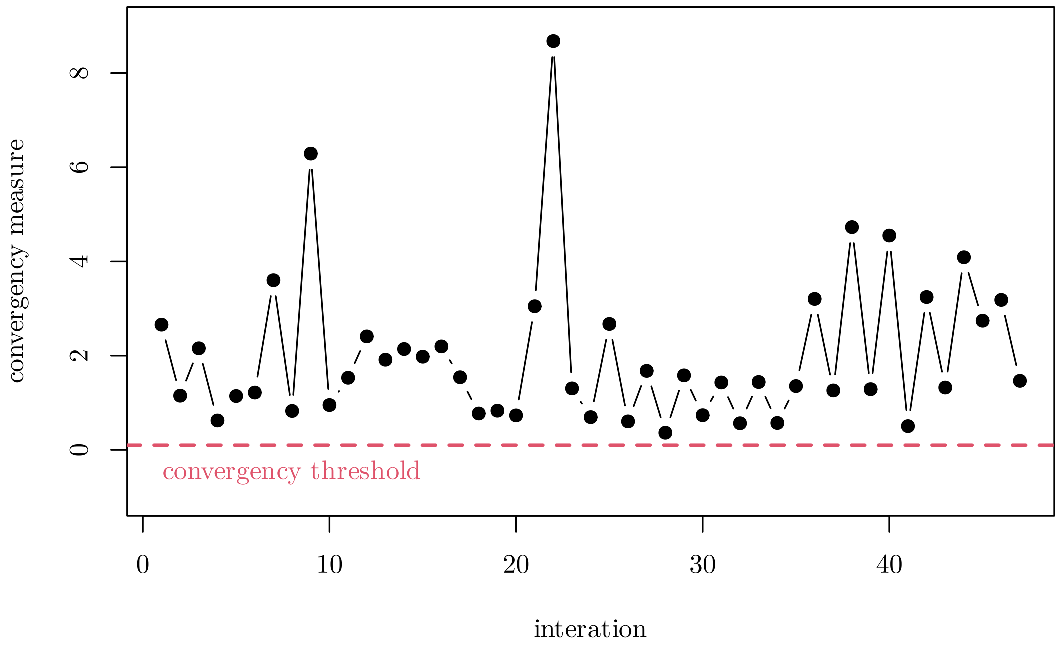


Figure 5.4: Convergence criterion. The mixed effect survival forest is considered to converge when all the elements of the random effect modelling parameter $\hat{b}$ do not increase or decrease of more than 10% in one iteration. The plot shows how the convergence is not reached in the fixed maximum number of iterations, since none of the iterations provided a convergency measure lower than the threshold.

6

Conclusion

6.1 Customer Base Segmentation

In the context of this thesis, an analysis was conducted on a dataset provided by a company in order to devise a robust and effective methodology for segmenting its customer base. This segmentation pursuit was led by a dual-pronged objective: how much economical value does a customer provides and how likely they are to make repeat purchases of the company's products. These two dimensions assumed pronounced significance, especially when viewed within the broader framework of predicting the Customer Lifetime Value (CLTV).

The value index was built as the first vector of scores of a principal component analysis calculated over variables related to the economical value of the purchases. On the other hand, the propensity index resulted being a transformation of the time from first purchase at which the probability of buying the company's product - predicted by a survival forest - was equal to 40%.

The segmentation resulted useful in identifying 6 groups of customers according to the business requirements. Also, when used as a feature in the CLTV calculation pipeline, it provided a decrease of around $80\$$ (3%) of the mean absolute error on predictions.

6.2 Mixed Effects Survival Forest

The motivation behind pursuing a mixed-effects survival forest raised from our goal to enhance the performance of survival models employed for the calculation of the propensity index. In an effort to extend the capabilities of random forests, we endeavored to integrate random effects, similar to those estimated in mixed-effects Cox models. This involved an attempt to create a generalized approach that would iteratively refine the data by removing the effects that the previous model should have already estimated. This has been performed by mean of the Martingale residuals.

However, during the application of this method to our specific problem, the iterative algorithm demonstrated a lack of convergence. It's important to note that a comprehensive evaluation of this approach was outside the scope of the current project. Further exploration and analysis, including simulations, are necessary to assess the efficacy of this approach more extensively. This could potentially provide valuable insights for future work in refining the mixed-effects survival forest methodology.

7 Appendix

7.1 R code for mixed effect survival forest

7.1.1 Functions

```r
#' Function to get the actual name of the time to event variable from a formula
#' of a coxph model
#' @param formula formula to get the time to event variable from
#' @returns a string representing the actual name of the time to event variable
#' @examples
#' formula <- formula(Surv(time_to_ev, event) ~ x1+x2)
#' name <- get_time_to_event_var_name(formula)
#' name
get_time_to_event_var_name <- function(formula){
  tmp_str <- as.character(formula)[2] %>% strsplit(',')
  tmp_str <- gsub('Surv\\(', '', tmp_str[[1]][1])
  tmp_str <- gsub(' ','',tmp_str)
  return(tmp_str)
}

#' Function to get the actual name of the event variable from a formula
#' of a coxph model
#' @param formula formula to get the time to event variable from
#' @returns a string representing the actual name of the event variable
#' @examples
#' formula <- formula(Surv(time_to_ev, event) ~ x1+x2)
#' name <- get_event_var_name(formula)
#' name
get_event_var_name <- function(formula){
  tmp_str <- as.character(formula)[2] %>% strsplit(',')
  tmp_str <- gsub('\\)', '', tmp_str[[1]][2])
  tmp_str <- gsub(' ','',tmp_str)
  return(tmp_str)
}

#' Function to get the actual name of the explicative variables from a formula
#' of a coxph model
#' @param formula formula to get the time to event variable from
#' @returns a string representing the actual names of the explicative variables
#' @examples
#' formula <- formula(Surv(time_to_ev, event) ~ x1+x2)
#' name <- get_explicative_var_names(formula)
#' name
get_explicative_var_names <- function(formula){
  tmp_str <- as.character(formula)[3] %>% strsplit(',')
  tmp_str <- gsub(' ','',tmp_str)
  tmp_str <- tmp_str[[1]][1] %>% strsplit('\\+')
  return(tmp_str[[1]])
}

#' Estimate the baseline hazard, probability, cumulative hazard and survival
#' function given a formula for the standard cox model,
#' without random effects
#' @param formula formula as it is in the coxph model
#' @param data training dataset. Must contain all the variables in formula
#' @returns a dataframe with 2 columns: the time and the baseline hazard
#' for that correspondent
#' @examples
#' data <- lung %>% mutate(status = recode(status, `1` = 0, `2` = 1))
#' formula <- formula(Surv(time, status) ~ sex+ph.karno)
#' out <- estimate_baseline_cox(formula, data)
#' out
estimate_baseline_cox <- function(formula, data){
  cox <- coxph(formula,
             data = data,
             model=TRUE )
  time_var_name         <- get_time_to_event_var_name(formula)
  event_var_name        <- get_event_var_name(formula)
  explicative_var_names <- get_explicative_var_names(formula)
  times <- data[[time_var_name]] %>% unique %>% sort

  new_data <- setNames(data.frame(times, NA), c(time_var_name, event_var_name))
  if(explicative_var_names != '1'){
    for(x in explicative_var_names){
      if(is.character(data[[x]])) data[[x]] <- data[[x]] %>% as.factor
      if(is.factor(data[[x]]))  base_value <- levels(data[[x]])[1]
      else if(is.numeric(data[[x]])) base_value <- 0
      new_data <- cbind(new_data, base_value)
    }
    new_data <- setNames(new_data, c(time_var_name, event_var_name, explicative_var_names))
  }


  St <- predict(cox,
              newdata = new_data,
              type='survival',
              model=T)
  St <- c(1, St) # at time=0 St=1
  risk_factor <- predict(cox,
                      newdata = new_data[1,],
                      type='risk')
  pt <- St[-length(St)] - St[-1] # at time=0 pt doesn't exist
  ht <- pt/St[-c( length(St))] / risk_factor
  ####### assumo tutti dati incensurati! ###########
  event <- rep(1, length(ht))

  # i pt vanno aggiustati perché non posso usare pesi = 0 nel fit dei modelli
  # occhio che non dia problemi di stabilità
  pt[pt==0] <- 1e-27
  output <- data.frame('t'=times, 'event' = event, 'ho'=ht, 'po'=pt, 'Ho'=cumsum(ht), 'St'=St[-1])
  colnames(output)[1:2] <- c(time_var_name, event_var_name)
  return(output)
}


#' Estimate the baseline hazard, probability, cumulative hazard and survival
#' function given a fromula for a non parametric Kaplain-Meier estiate.
#' @param formula formula
#' @param data training dataset. Must contain all the variables in formula
#' @returns a dataframe with some columns: the time and the correspondent
#' baseline hazard,
#' probability, survival function
#' @examples
#' data <- lung %>% mutate(status = recode(status, ‘1‘ = 0, ‘2‘ = 1))
#' formula <- formula(Surv(time, status) ~ sex+ph.karno)
#' out <- estimate_baseline_km(formula, data)
#' out
estimate_baseline_km <- function(formula, data){
  km <- survfit(formula,
                data = data,
                model=TRUE )
  time_var_name         <- get_time_to_event_var_name(formula)
  event_var_name        <- get_event_var_name(formula)
  times <- data[[time_var_name]] %>% unique %>% sort

  St <- km$surv
  tmp <- survival_to_RiskProbabilty(St)
  ht <- tmp$risk
  pt <- tmp$probability

  # i pt vanno aggiustati perché non posso usare pesi = 0 nel fit dei modelli
  # occhio che non dia problemi di stabilità
  pt[pt==0] <- 1e-27
  output <- data.frame('t'=times, 'h0'=ht,
                       'p0'=pt, 'H0'=cumsum(ht),
                       'S0'=St)
  colnames(output)[1] <- time_var_name
  return(output)
}


#' Estimate the baseline hazard, probability, cumulative hazard and survival
#' function. Calls 2 specialized functions for method='cox_model' (parametric
#' estimates) and method='kaplain_meier' (non parametric).
#' @param formula formula as it is in the coxph model if method='cox_model',
#' otherwise same but without elements after '~'. E.g., 'Surv(t,e) <- 1'.
#' @param data training dataset. Must contain all the variables in formula
#' @param time_stamps_reduction_method the unique times are generally too many.
#' Their number has to be reduced. Two methods are implemented.
#' 'threshold' will
#' get rid of all the values of the survival functions that refer to a time at
#' which the probability of observing the event is lower that a set treshold.
#' 'aggregation' will aggregate together times refering to the same time period
#' defined by aggregation_period.
#' @param threshold minimum value the survival probability has to have to be kept
#' in the output. Default is 'expected', i.e., all rows with survival probability
#' greater than the expected one are kept. Otherwise, numeric value in (0,1).
#' If time_stamps_reduction_method != 'threshold' is ignored.
#' @param aggregation_period which period to be used to aggregate the times.
#' @returns $baseline: a dataframe with 2 columns: the time and the baseline hazard
#' for that correspondent.
#' $training_data: original data with aggregated times if used.
#' @examples
#' data <- lung %>% mutate(status = recode(status, ‘1‘ = 0, ‘2‘ = 1))
#' formula <- formula(Surv(time, status) ~ sex+ph.karno)
#' h0 <- estimate_baseline(formula, data,
#'                            time_stamps_reduction_method='threshold')
#' h0
estimate_baseline <- function(formula, data, method = c('cox_model', 'kaplain_meier'),
                              threshold = 'expected',
                              aggregation_days = 30.5,
                              time_stamps_reduction_method = c('threshold', 'aggregation')){
  if(time_stamps_reduction_method == 'aggregation') {
    time_var_name <- get_time_to_event_var_name(formula)
    max_time = max(data[[time_var_name]])
    times = seq(0, max_time, by = aggregation_days)
    times = times
    if(times[length(times)] != max_time) times[length(times)+1] = max_time
    time_tmp <- times[cut(data[[time_var_name]], breaks = c(times, Inf), labels = FALSE)]
    data <- data %>%
      mutate(!!time_var_name := time_tmp + aggregation_days)
    cat(length(times), 'rows have been generated\n\n')

  }

  if(method=='cox_model'){
    out = estimate_baseline_cox(formula, data)
  }else if(method=='kaplain_meier'){
    out = estimate_baseline_km(formula, data)
  }else{cat('ERROR: wrong method. Must be between cox_model and kaplain_meier')}

  if(time_stamps_reduction_method == 'threshold'){
    if(threshold == 'expected') threshold <- 1/length(out$po)
    idxs_to_remove <- out$po < threshold # the info here is not enough, I don't keep it
    cat(sum(!idxs_to_remove), 'rows have been generated\n\n')
    out <- out[!idxs_to_remove,]
  }

  #pt <- out$po/sum(out$po[!idxs_to_remove]) # make sum of them = 1
  #out$po_adj <- pt
  #return(out[!idxs_to_remove,] %>% select(-po))
  out$po_adj <- out$po
  return(list('baseline'=out, 'training_data'=data))
}



#' Function that transform the Cumulative Hazard Function to a Survival Function
#' that contains the mixed effects estimated by a coxme object (Cox model with
#' Random Effects)
#' @param hazard matrix where each row is the cumulative hazard OR the hazard
#' function predicted for one observation. If possible, chose the hazard,
#' it's much faster.
#' @param hazard_type a value between 'cum_haz' or 'haz'. Describing the type of
#' risk function of 'hazard'
#' @param coxme coxme object
#' @param data_CHF the data over which the CHF has been calculated. Must contain
#' the random effect variable
#' @param re_var_name name of the variable (in data_CHF) that is the random effect
add_me_to_survival_curve <- function(hazard, hazard_type, me_model, data_CHF, re_var_name){
  #print(coxme$frail[[re_var_name]])
  coef <- me_model$frail[[re_var_name]]
  lp <- coef[match(data_CHF[[re_var_name]], names(coef))]
  data_CHF <- data_CHF %>%
    mutate(mixed_effect = exp(lp)) %>%
    select(mixed_effect, all_of(re_var_name))
  n <- ncol(hazard)
  if(hazard_type=='cum_haz'){
    hazard <- apply(hazard, 1, function(x) x[-1] - x[-ncol(hazard)]) %>% t
    hazard <- cbind(rep(0, nrow(hazard)), hazard)
  }
  survival_me <- matrix(NA, nrow(hazard), ncol(hazard))
  for(i in 1:(nrow(hazard))){
    if(i%%1000 == 0) cat(i/nrow(hazard)*100, '%\n')
    hazard[i,] <- hazard[i,]*data_CHF$mixed_effect[i]
    survival_me[i, 1:n] <- cumprod(1 - hazard[i, 1:n])
  }
  return(survival_me)
}



#' Function to convert a discrete survival function to the equivalent risk
#' and probility functions
#' @param survival vector of survival values over time
#' @param risk_factor a number that represents the adjustment to be provided to the
#' risk. For example, if you want to remove the effect given by a fixed effect (b_i)
#' related to a variable (x) from a cox model, risk_factor will be equal to
#' exp(b_i * x_i). Default is 1.
#' THE probability IS NOT ADJUSTED BY THE risk-factor!!!!
#' @examples
#'
#' survival <- c(1,1,0.9,0.6,0.4,0.35,0.2,0)
#' out <- survival_to_RiskProbabilty(survival)
#' probability <- out$probability
#' risk <- out$risk
survival_to_RiskProbabilty <- function(survival, risk_factor=1, only_risk=F, only_prob=F){
  St <- c(1, survival) # at time=0 St=1
  pt <- St[-length(St)] - St[-1] # at time=0 pt doesn't exist
  ht <- pt/St[-c( length(St))] / risk_factor
  if(only_risk) return(ht)
  if(only_prob) return(pt)
  return(list('risk'=ht, 'probability'=pt))
}


#' Add the Martingale residuals and the new column of censoring as presented in the thesis
#' based on the martingale residuals.
#' @param expanded_data expanded data to which add the new column(s)
#' @param original_data needed to know the actual times-to-event
#' @param coxme_model model of which to calculate the residuals. It MUST be trained
#' @param baseline_km part of the coxme_model
#' @param epsilon threshold to define if a residual indicates that the next model has to predict a censoring or not.
#' on the original data.
add_martingale_censoring_to_expanded <- function(expanded_data, original_data, coxme_model, baseline_km, epsilon=0.5){
  if(is(coxme_model)=='coxme') {formula <- coxme_model$formula$fixed} else {formula <- coxme_model$formula}

  time_var_name <- get_time_to_event_var_name(formula)
  event_var_name <- get_event_var_name(formula)

  n_times <- nrow(expanded_data) / nrow(original_data)
  repeated_times <- original_data[rep(seq_len(nrow(original_data)), each = n_times), ] %>%
    select(all_of(time_var_name)) %>%
    rename(old_time_btw_first_and_second = time_btw_first_and_second)
  all_data <- cbind(expanded_data, repeated_times)

  # calculate the martingale residuals
  cat('Calculating residuals...')
  multiply_last <- function(row, skip_last_two = T) {
    if(skip_last_two){
      idxs <- c(length(row), length(row)-1)
      r <- as.numeric(row[-idxs]) * as.numeric(row[length(row)])
      r <- c(r, row[idxs])
    }else{
      r <- row * row[length(row)]
    }
    return(r)
  }
  hazards_and_me <- data.frame(me_names = (coxme_model$frail$CS004_LOCATION_CD %>% names),
                               effect = exp((coxme_model$frail$CS004_LOCATION_CD)))
  haz <- data.frame(matrix(rep(baseline_km$H0, nrow(original_data)),
                           ncol = length(baseline_km$H0), byrow = TRUE))
  haz <- haz %>% mutate(mixed_effect_names = original_data$CS004_LOCATION_CD)
  haz <- haz %>% left_join(hazards_and_me, by = join_by(mixed_effect_names==me_names))
  haz <- t(apply(haz, 1, multiply_last)) %>% as.data.frame
  haz[,-ncol(haz)] <- lapply(haz[,-ncol(haz)], as.numeric)

  ri_col=NULL
  for(i in 1:nrow(original_data)){
    ri_all <- haz[i, - c(ncol(haz), ncol(haz)-1)] %>% as.numeric
    ri_col <- c(ri_col, ri_all)
  }
  cat('done\n')

  cat('Creating new censoring info...')
  if(! 'old_censoring_compra_piu_di_una_volta' %in% colnames(all_data) ) all_data <- all_data %>% mutate(old_censoring_compra_piu_di_una_volta
  tmp <- cbind(all_data, ri_col) %>%
    filter(!( (!old_censoring_compra_piu_di_una_volta) & (time_btw_first_and_second>old_time_btw_first_and_second)))


  # Fix censoring column
  tmp <- tmp %>%
    mutate(compra_piu_di_una_volta =
             case_when(old_censoring_compra_piu_di_una_volta & (time_btw_first_and_second >= old_time_btw_first_and_second) ~ TRUE,
                       TRUE ~ FALSE))
  # Martingale res
  tmp <- tmp %>%
    mutate(residuals_martingale = compra_piu_di_una_volta - ri_col) %>%
    mutate(residuals_martingale = ifelse(residuals_martingale>=-1, residuals_martingale, -1))

  # New censoring column from residuals
  tmp <- tmp %>%
    mutate(new_censoring = ifelse( abs(residuals_martingale) <= epsilon, 0, 1))

  cat('done\n')

  return(tmp)
}



#' Add the Martingale residuals and the new column of censoring as presented in the thesis
#' based on the martingale residuals.
#' @param expanded_data expanded data to which add the new column(s)
#' @param original_data needed to know the actual times-to-event
#' @param model model of which to calculate the residuals. It MUST be trained
#' @param baseline_km part of the coxme_model
#' @param epsilon threshold to define if a residual indicates that the next model has to predict a censoring or not.
#' on the original data.
add_martingale_censoring_to_original <- function(original_data, model, baseline_km=NULL, epsilon=0.5, formula_srf=NULL){
  if(is(model)=='coxme') {

    cat('Calculating residuals...')

    hazards_and_me <- data.frame(me_names = (model$frail$CS004_LOCATION_CD %>% names),
                                 effect = exp((model$frail$CS004_LOCATION_CD)))
    ri_col=NULL
    fun <- function(row){
      H_hat <- baseline_km[baseline_km$time_btw_first_and_second == as.numeric(row[1]),]$H0
      re <- hazards_and_me[hazards_and_me$me_names == row[10],]$effect
      return(H_hat*re)
    }

    ri_col <- apply(original_data, 1, fun)
    cat('done\n')

    cat('Creating new censoring info...')
    tmp <- cbind(original_data, ri_col)

    # Martingale res
    tmp <- tmp %>%
      mutate(residuals_martingale = compra_piu_di_una_volta - ri_col) %>%
      mutate(residuals_martingale = ifelse(residuals_martingale>=-1, residuals_martingale, -1)) %>%
      rename(compra_piu_di_una_volta_old = compra_piu_di_una_volta)

    # New censoring column from residuals
    tmp <- tmp %>%
      mutate(compra_piu_di_una_volta = ifelse( abs(residuals_martingale) <= epsilon, 0, 1))

    cat('done\n')
    }
  if(is(model)=='ranger'){

    cat('Calculating residuals...')
    data_for_predict_srf <- expand_dataset_interactions(formula_srf, original_data)
    predictions <- predict(model, data_for_predict_srf)

    ri_col <- NULL
    for(i in 1:nrow(original_data)){
      idx_time <- which(model$unique.death.times == original_data$time_btw_first_and_second[i])
      H_hat <-predictions$chf[i, idx_time]
      ri_col <- c(ri_col, H_hat)
    }
    cat('done\n')

    cat('Creating new censoring info...')
    if('ri_col' %in% colnames(original_data)) original_data <- original_data %>% select(-ri_col)
    tmp <- cbind(original_data, ri_col)

    # Martingale res
    tmp <- tmp %>%
      mutate(residuals_martingale = compra_piu_di_una_volta - ri_col) %>%
      mutate(residuals_martingale = ifelse(residuals_martingale >=-1, residuals_martingale, -1)) %>%
      mutate(compra_piu_di_una_volta_old = compra_piu_di_una_volta) %>%
      select(-compra_piu_di_una_volta)

    # New censoring column from residuals
    tmp <- tmp %>%
      mutate(compra_piu_di_una_volta = ifelse( abs(residuals_martingale) <= epsilon, 0, 1))

    cat('done\n')
  }
  return(tmp)
}
```

7.1.2 Fitting

```
library(coxme)
library(survival)
library(dplyr)
library(SurvMetrics)
library(ranger)

explicative_variables <- c('eta_cliente_primo_acquisto',
                           'GG_DISTANZA_INTERAZ_PRIMO_ACQ',
                           'tasso_interaz_last_3Y_dopo_5y',
                           'ETA_PRIMO_ACQUISTO',
                           'LAST_SALE_AGING_GROUP_AL_2017',
                           'APPT_TYPE_POST_PRIMO_ACQ',
                           'AGING_CLIENTE_AL_2017',
                           'CS004_LOCATION_CD',
                           'compra_piu_di_una_volta')


one_step <- function(input_data, explicative_variables, epsilon=0.5){
  require(coxme)
  require(survival)
  require(dplyr)
  require(SurvMetrics)
  require(ranger)

  input_data <- input_data %>% select(all_of(c('time_btw_first_and_second', 'compra_piu_di_una_volta', explicative_variables)))

  # Estimate the Kaplan-Meier curve
  formula <- formula('Surv(time_btw_first_and_second,compra_piu_di_una_volta)~1')
  tmp <- estimate_baseline(formula,
                           data = input_data,
                           method = 'kaplain_meier',
                           aggregation_days = 1,
                           time_stamps_reduction_method = 'aggregation')


  baseline_km_not_aggr <- tmp$baseline
  original_data <- input_data
  rm(input_data)
  me_model_after_RF_from_KM <- coxme(Surv(time_btw_first_and_second, compra_piu_di_una_volta) ~
                                        (1|CS004_LOCATION_CD),
                                     data=original_data)

  expanded_data_Martingale_censoring <- add_martingale_censoring_to_original(original_data,
                                                                              me_model_after_RF_from_KM,
                                                                              baseline_km_not_aggr,
                                                                              epsilon = epsilon)

  rf_without_me_from_KM <- ranger(Surv(time_btw_first_and_second,compra_piu_di_una_volta)~
                                    tasso_interaz_last_3Y_dopo_5y +
                                    ETA_PRIMO_ACQUISTO +
                                    APPT_TYPE_POST_PRIMO_ACQ +
                                    LAST_SALE_AGING_GROUP_AL_2017+
                                    GG_DISTANZA_INTERAZ_PRIMO_ACQ +
                                    tasso_interaz_last_3Y_dopo_5y:AGING_CLIENTE_AL_2017 +
                                    ETA_PRIMO_ACQUISTO:AGING_CLIENTE_AL_2017,
                                  data = expanded_data_Martingale_censoring,
                                  num.trees = 200,
                                  min.node.size=75,
                                  oob.error=F,
                                  num.threads = 8,
                                  case.weights = expanded_data_Martingale_censoring$po
  )

  formula_srf <- as.formula('Surv(time_btw_first_and_second,compra_piu_di_una_volta)~
                                  tasso_interaz_last_3Y_dopo_5y +
                                  ETA_PRIMO_ACQUISTO +
                                  APPT_TYPE_POST_PRIMO_ACQ +
                                  LAST_SALE_AGING_GROUP_AL_2017+
                                  GG_DISTANZA_INTERAZ_PRIMO_ACQ +
                                  tasso_interaz_last_3Y_dopo_5y:AGING_CLIENTE_AL_2017 +
                                  ETA_PRIMO_ACQUISTO:AGING_CLIENTE_AL_2017')


  data_without_SRF_impact <- add_martingale_censoring_to_original(original_data = original_data,
                                                                   model = rf_without_me_from_KM,
                                                                   formula_srf = formula_srf,
                                                                   epsilon = epsilon)
  data_without_both_impact <- add_martingale_censoring_to_original(original_data = expanded_data_Martingale_censoring,
                                                                   model = rf_without_me_from_KM,
                                                                   formula_srf = formula_srf,
                                                                   epsilon = epsilon)


  data_for_predict_srf <- expand_dataset_interactions(formula_srf, original_data)
  srf_predictions <- predict(rf_without_me_from_KM, data_for_predict_srf)

  s_pred <- srf_predictions$chf %>% add_me_to_survival_curve(hazard_type = 'cum_haz',
                                                              me_model = me_model_after_RF_from_KM,
                                                              data_CHF = original_data %>% select('CS004_LOCATION_CD'),
                                                              re_var_name = 'CS004_LOCATION_CD')
  half_times <- rf_without_me_from_KM$unique.death.times[seq(1,length(rf_without_me_from_KM$unique.death.times), by=4)]
  med_index = median(1:length(half_times))
  mat_rsf =s_pred[,seq(1,length(rf_without_me_from_KM$unique.death.times), by=4)]
  surv_obj = Surv(original_data$time_btw_first_and_second, original_data$compra_piu_di_una_volta)
  c_index_val = Cindex(surv_obj, predicted = mat_rsf[, med_index])

  return(list('coxme'=me_model_after_RF_from_KM,
              'srf'=rf_without_me_from_KM,
              'combined_surv_predictions' = s_pred,
                  'data_without_SRF_effects' = data_without_SRF_impact,
                  'data_without_both_effects' = data_without_both_impact,
                  'c_index' = c_index_val))
  }

  stopping_criterion <- function(b_new, b_old, threshold){
    M <- max(abs(b_new-b_old))
    idx_M <- which.max(abs(b_new-b_old))
    tr <- M/b_old[idx_M]
    cond <- abs(tr) < threshold
    print(tr)
    return(list('cond'=cond, 'tr'=tr))
  }


  n <- 5000
  df_survival_train_rf_sample <- sample_with_10obs_per_level(df_survival_train_rf,
                                                              variable = 'CS004_LOCATION_CD',
                                                              n_rows = n,
                                                              seed = 12345)

  # df_survival_train_rf_sample
  tmp1 <- one_step(df_survival_train_rf_sample, explicative_variables)
  conds <- NULL
  i <- 1
  max_iter <- 50
  while(TRUE){
    tmp <- one_step(tmp1$data_without_SRF_effects, explicative_variables)
    b_new <- tmp$coxme$frail$CS004_LOCATION_CD
    b_old <- tmp1$coxme$frail$CS004_LOCATION_CD
    cond <- stopping_criterion(b_new, b_old, threshold = 0.1)
    conds <- c(conds, cond$tr)
    if(cond$cond) break
    tmp1 <- tmp
    i <- i+1
    if(i == 50) break
  }
```

7.2 Tables

The table about the features in the original data is provided here.

Variable name	Only in DF2022	Description
CONSOL_CUSTOMER_KY	FALSE	unique customer id
LIVELLO_DS_PRIMA_VENDITA	FALSE	device level from 1 (low quality) to 5 (high quality)
NET_AMOUNT_FIRST_SALE_ACTUAL	FALSE	inflection-adjusted value of the first sale
VALORE_STORICO_CLIENTE_ACTUAL	FALSE	inflection-adjusted historical value of the client as sum of all the purchases
NRO_ACQUISTI_CLIENTE	FALSE	total number of purchases
NRO_UNITA_ACQUISTATE_CLIENTE	FALSE	total number of purchased units
VALORE_MEDIO_ACQUISTI_CLIENTE	FALSE	average value of 1 purchase
VALORE_MEDIO_UNITA_CLIENTE	FALSE	average value of 1 unit
LAST_SALE_AGING_GROUP	FALSE	time passed from last purchase
ETA_PRIMO_ACQUISTO	FALSE	age first purchase
ETA_OGGI_CLIENTE	FALSE	age at 2017-10-31
tasso_interaz_last_3Y	FALSE	interaction rate (visits to the shops, calls, emails...) between 2014-10-31 and 2017-10-31
AGING_CLIENTE	FALSE	time passed from first purchase
SEGMENTAZIONE_3_DESCR	FALSE	preexisting expert segmentation
ETA_PRIMO_ACQUISTO_CL	TRUE	class of age first purchase
CU015_BIRTHDATE_DT	TRUE	birth date
PRODUCT_KY	TRUE	product key
LAST_SALE_AGING_GROUP_detailed	TRUE	class of time passed from last purchase
DT_PRIMO_ACQUISTO	TRUE	device type first purchase
DT_ULTIMO_ACQUISTO	TRUE	device type last purchase
gg_distanza_primo_acquisto	TRUE	days from first purchase
gg_distanza_ultimo_acquisto	TRUE	days from last purchase
eta_ultimo_acquisto	TRUE	age last purchase
VALORE_STORICO_CLIENTE_BASE	TRUE	preexisting expert segmentation
CS003_LOCATION_ID	TRUE	customer's location's id
CS004_LOCATION_CD	TRUE	customer's city's id
CS005_LOCATION_NAME	TRUE	customer's location's name
CS017_REGION_DS	TRUE	shop's region's name
CS014_CORE_ACCOUNT_NAME_DS	TRUE	shop of reference
CS013_CORE_ACCOUNT_NUMBER_DS	TRUE	shop's code
NET_AMOUNT_FIRST_SALE	TRUE	amount first purchase
PTA_UDITIVA_PRIMO_ACQ	TRUE	result of the hearing test first purchase
TEST_DT_PRIMO_ACQ	TRUE	date of the hearing test first purchase
HA022_TECHNOLOGY_DS	TRUE	technology of the device
HA021_PRODUCT_BRAND	TRUE	brand of the device
HA040_RECHARGEABLE_FLAG	TRUE	Y if rechargeable device
HA008_PRODUCT_TYPE_DS	TRUE	device type
VALORE_STORICO_CLIENTE	TRUE	historical value client
LIV_DS_x_PTA_UDIT	TRUE	device level for hearing test
NET_AMOUNT_FIRST_SALE_CL	TRUE	class of net amount first purchase
APPT_TYPE_POST_PRIMO_ACQ	TRUE	type of first appointment after first purchase
APPT_SUBTYPE_POST_PRIMO_ACQ	TRUE	sub type of first appointment after first purchase type of first appuntament after first purchase
REFERRAL_SOURCE_POST_PRIMO_ACQ	TRUE	type of first appointment after first purchase type of first appuntament after first purchase
REFERRAL_SUBSOURCE_POST_PRIMO_ACQ	TRUE	sub type of first appointment after first purchase type of first appuntament after first purchase
DATA_INTERAZIONE_POST_PRIMO_ACQ	TRUE	date first interaction after first purchase
GG_DISTANZA_INTERAZ_PRIMO_ACQ	TRUE	days between first purchase and first interaction
BANDS_DISTANZA_INTERAZIONE_PRIMO_ACQ	TRUE	class days between first purchase and first interaction days between first purchase and first interaction
YEAR_FIRST_SALE	TRUE	year of first purchase
LIVELLO_DS_PRIMA_VENDITA_FASCIA	TRUE	class of level device first p
nro_appt_post_ultimo_acq_last_3y	TRUE	number of appointments after last purchase last 3 years

Table 7.1: Description of the varibles in the two datasets. DF2022 contains all the variables, while DF2017 only some of them

References

[1] A. Shimokawa, Y. Kawasaki, and E. Miyaoka, "Comparison of splitting methods on survival tree," *The international journal of biostatistics*, vol. 11, no. 1, pp. 175–188, 2015.

[2] M. Radespiel-Tröger, T. Rabenstein, H. T. Schneider, and B. Lausen, "Comparison of tree-based methods for prognostic stratification of survival data," *Artificial Intelligence in Medicine*, vol. 28, no. 3, pp. 323–341, 2003.

[3] P. Kotler and G. Armstrong, *Marketing management, analysis, planning, implementation, and control*. London: Prentice-Hall International, 1994.

[4] A. Nairn and P. Berthon, "Creating the customer: The influence of advertising on consumer market segments–evidence and ethics," *Journal of Business Ethics*, vol. 42, pp. 83–100, 2003.

[5] S.-Y. Kim, T.-S. Jung, E.-H. Suh, and H.-S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study," *Expert systems with applications*, vol. 31, no. 1, pp. 101–107, 2006.

[6] P. C. Verhoef and B. Donkers, "Predicting customer potential value an application in the insurance industry," *Decision support systems*, vol. 32, no. 2, pp. 189–199, 2001.

[7] J. C. Hoekstra and E. K. Huizingh, "The lifetime value concept in customer-based marketing," *Journal of Market-Focused Management*, vol. 3, pp. 257–274, 1999.

[8] H. Hwang, T. Jung, and E. Suh, "An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry," *Expert systems with applications*, vol. 26, no. 2, pp. 181–188, 2004.

[9] S. Dibb, "Market segmentation: strategies for success," *Marketing Intelligence & Planning*, vol. 16, no. 7, pp. 394–406, 1998.

[10] P. Kotler, *Marketing management: A south Asian perspective*. Pearson Education India, 2009.

[11] P. D. Berger and N. I. Nasr, "Customer lifetime value: Marketing models and applications," *Journal of Interactive Marketing*, vol. 12, no. 1, pp. 17–30, 1998.

[12] W. Chang, C. Chang, and Q. Li, "Customer lifetime value: A review," *Social Behavior and Personality: an international journal*, vol. 40, 08 2012.

[13] M. S. Ziafat H, "Using data mining techniques in customer segmentation," *Int. Journal of Engineering Research and Applications*, vol. 4, pp. 70–79, 09 2014.

[14] G. Zhang, "Customer segmentation based on survival character," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, 2007, pp. 3391–3396.

[15] A. Azzalini and B. Scarpa, *Data Analysis and Data Mining*. Oxford University Press, 2012.

[16] S. Vyas and L. Kumaranayake, "Constructing socio-economic status indices: how to use principal components analysis," *Health Policy and Planning*, vol. 21, no. 6, pp. 459–468, 10 2006.

[17] E. Gregorio and L. Salce, *Algebra lineare*, nuova edizione ed. Padova: Libreria Progetto, 2012.

[18] R. C. Elandt-Johnson and N. L. Johnson, *Survival models and data analysis*. John Wiley & Sons, 1980, vol. 110.

[19] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed. Springer, 2003.

[20] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[21] L. Breiman, "Random forest," *Machine Learning*, vol. 45, 10 2001.

[22] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841 – 860, 2008.

[23] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur, "A review of survival trees," *Statistics Surveys*, vol. 5, 01 2011.

[24] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen, *Classification and regression trees*, ser. «The »Wadsworth statistics/probability series. Monterey (CA): Wadsworth and Brooks/Cole Advanced Books and Software, c1984.

[25] H. Wang and G. Li, "A selective review on random survival forests for high dimensional data," *Quantitative Bio-Science*, vol. 36, 12 2017.

[26] D. J. Bertsimas Dimitris, "Optimal survival trees," *Machine Learning*, vol. 111, 12 2022.

[27] S. M. Schmoor C, Sauerbrei W, "Assessment and comparison of prognostic classification schemes for survival data." *Stat Med*, 09 1999.

[28] H. Zhou, X. Cheng, S. Wang, Y. Zou, and H. Wang, *SurvMetrics: Predictive Evaluation Metrics in Survival Analysis*, 2022, r package version 0.5.0. [Online]. Available: https://CRAN.R-project.org/package=SurvMetrics

[29] H. Moradian and D. Larocque, "l_1 splitting rules in survival forests," *Lifetime Data Analysis*, 10 2017.

[30] Y. Zou, G. Fan, and R. Zhang, "Integrated square error of hazard rate estimation for survival data with missing censoring indicators," *Journal of Systems Science and Complexity*, 04 2021.

[31] M. Scehmper, "The explained variation in proportional hazards regression," *Biometrika*, vol. 77, no. 1, pp. 216–218, 03 1990.

[32] Z. H. Hoo, J. Candlish, and D. Teare, "What is an roc curve?" pp. 357–359, 2017.

[33] O. R. Gordon L, "Tree-structured survival analysis." *Cancer Treat Rep.*, vol. 69, 10 1985.

[34] A. J. Davis RB, "Exponential survival trees." *Stat Med*, vol. 8, 08 1989.

[35] M. LeBlanc, "Survival trees by goodness of split." *Journal of the American Statistical Association*, vol. 88, no. 422, 1993-06-01.

[36] M. LeBlanc and J. Crowley, "Relative risk trees for censored survival data," *Biometrics*, pp. 411–425, 1992.

[37] H. Zhang, "Splitting criteria in survival trees," 1995.

[38] S. Keleş and M. R. Segal, "Residual-based tree-structured survival analysis," *Statistics in medicine*, vol. 21, no. 2, pp. 313–326, 2002.

[39] D. P. Harrington and T. R. Fleming, "A class of rank test procedures for censored survival data," *Biometrika*, vol. 69, no. 3, pp. 553–566, 12 1982.

[40] H. Abdi, "Bonferroni and sidák corrections for multiple comparisons," *Encyclopedia of measurement and statistics*, vol. 3, no. 01, p. 2007, 2007.

[41] P. M. Grambsch and T. Thernau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, vol. 81, no. 3, pp. 515–526, 09 1994.

[42] J. C. Pinheiro and D. M. Bates, "Linear mixed-effects models: basic concepts and examples," *Mixed-effects models in S and S-Plus*, pp. 3–56, 2000.

[43] T. Therneau *et al.*, "Mixed effects cox models," *CRAN repository*, 2015.

[44] T. A. Balan and H. Putter, "A tutorial on frailty models," *Statistical methods in medical research*, vol. 29, no. 11, pp. 3424–3454, 2020.

[45] T. Therneau, *coxme: Mixed Effects Cox Models*, 2022, r package version 2.2-18.1. [Online]. Available: https://CRAN.R-project.org/package=coxme

[46] D. R. Cox and E. J. Snell, "A general definition of residuals," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 248–265, 1968.

[47] J. Crowley and M. Hu, "Covariance analysis of heart transplant survival data," *Journal of the American Statistical Association*, vol. 72, no. 357, pp. 27–36, 1977.

[48] T. M. Therneau, P. M. Grambsch, and T. R. Fleming, "Martingale-based residuals for survival models," *Biometrika*, vol. 77, no. 1, pp. 147–160, 1990.

[49] D. Schoenfeld, "Partial residuals for the proportional hazards regression model," *Biometrika*, vol. 69, no. 1, pp. 239–241, 1982.

[50] S. Halabi, S. Dutta, Y. Wu, and A. Liu, "Score and deviance residuals based on the full likelihood approach in survival analysis," *Pharmaceutical statistics*, vol. 19, no. 6, pp. 940–954, 2020.

[51] M. Pellagatti, C. Masci, F. Ieva, and A. M. Paganoni, "Generalized mixed-effects random forest: A flexible approach to predict university student dropout," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 14, no. 3, pp. 241–257, 2021.

[52] M. N. Wright and A. Ziegler, "ranger: A fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, vol. 77, no. 1, pp. 1–17, 2017.

Acknowledgments

I would like to express my gratitude to my advisor, Prof. Bruno Scarpa of the University of Padova, for his guidance throughout the course of this research.

I am also thankful to Angelo Basile, my tutor from Alkemy, who not only provided me with the necessary data but also created the ideal environment for me to work on this project.

I would like thank my colleagues Andres, Ismail, Nicole and Sergio, with whom I had the opportunity to share numerous discussions and experiences throughout the thesis writing semester.

Finally, I want to thank my family for always being there with their unwavering support.