Portfolio, Churn & Customer Value

Hugo Cornet, Pierre-Emmanuel Diot, Guillaume Le Halper, Djawed Mancer

2022-03-30

Contents

Abstract

This paper is being realized as part of our last year in master's degree in economics. It aims at studying the firm's most valuable asset: its customers. To that end, we adopt a quantitative approach based on a mix of Econometrics and Data Science techniques with a threefold purpose:

- Model customer *portfolio* as a set of customer segments;
- Predict and analyze customer *attrition*;
- Estimate customer portfolio's overall *value*.

After having defined the subject's key concepts, we apply duration models and machine learning algorithms to a kaggle dataset related to customers of a fictional telecommunications service provider (TSP).

Keywords: *customer portfolio management (CPM), churn, customer value, duration models, machine learning, telecom.*

Chapter 1

Introduction

In a world in which the access to information is almost free or insignificant and where there is a real plurality of offers, churn analysis has become one of the key points a firm needs to focus on. Whoever says plurality of offers needs to introduce the term competition. Thereby, the latter is more and more fierce and cut-throat. Furthermore, switching costs have decreased significantly thanks to market regulation laws. For instance in France, when you switch TSP, the new provider pays you off cancellation fees. All of this being said, it is essential for firms to implement efficient strategies to enhance customer relationships. To that end, the development of both survival models and machine learning algorithms have enabled companies to really push-up their strategies in terms of customer *portfolio* management, monitoring of *attrition* and estimation of customer *value*.

After careful consideration of the issues at stake, the following key steps are focused on:

- Segmentation of customer portfolio as firms generally tend to partition their *portfolio* into multiple segments.
- Estimation of customer lifetime and prediction of *attrition*.
- Measurement of customer *value*.

In the following sections, the concepts of *portfolio*, *attrition* and customer *value* are defined. Then, some pieces of literature review are provided. Before embarking on data analysis and modelling, we present the theoretical basis of the models used in the study. We finally introduce the dataset and implement the methodology with the aim of estimating the overall value related to a fictional customer portfolio of a telecommunications service provider.

1.1 How to define a customer portfolio?

The notion of *portfolio* has greatly evolved before the firms' consumer base was considered as a *portfolio*. In chapter 2 a part of the literature review depicts an evolution of the *portfolio* management notion. A customer *portfolio* can be defined as a set of customers divided into several segments (or clusters) based on similar attributes. These discriminant features can be both economic (willingness to pay, budget constraint, etc.) and sociological (gender, age, socio-professional category, etc.). The underlying objective of this segmentation is to optimally allocate the company's resources.

When dealing with customer *portfolio* management (CPM), two dimensions can be considered. On the one hand, it can be assumed that a customer stays in the same segment throughout their life in the firm's *portfolio*. On the other hand, a dynamic approach can be adopted as suggested by Homburg et al. (2009) on dynamics in customer portfolio. In their article, the authors question the static analysis by assuming that a customer can switch between segments. They explain that one of the firm's objectives is to convert less valuable customers into more valuable ones.

1.2 What is attrition?

Customer *attrition* or churn occurs when a client discontinues using a service or a product offered by a firm. Churn analysis corresponds to both measurement and prediction of the *attrition* rate in the customer base of the company. Evaluating *attrition* depends on the type of relationship between the firm and its clients. When it is defined by a contract, the customer has to inform the firm about their service termination. In the telecom industry, a consumer is required to notify their TSP before going to a competing company. In an opposite way, the firm/client relationship can be non-contractual. In that case, the service termination does not need to be notified. *Attrition* then becomes a latent variable and more advanced models are used to make forecasts.

1.3 What does customer "value" mean?

In customer *portfolio* management, one client's *value* is represented by the **customer lifetime value** (CLV). CLV is the present *value* of all future purchases made by a customer over their lifetime in the firm's portfolio, taking into account the *attrition* risk. CLV depends both on the purchase recency as well as on the purchasing rate and aims at identifying the most valuable customer groups. Formally, Gupta and Lehmann (2003) define CLV for customer i as follows:

$$CLV_i = \sum_{t=0}^{T} \frac{(p_t - c_t) r_{i,t}}{(1+a)^t} - AC_i \tag{1.1}$$

with,

- p_t the price paid by customer i at time t
- c_t the marginal cost at time t
- $r_{i,t}$ the **probability that customer i be active** at time t
- a the discount rate
- AC_i the acquisition cost of customer i
- T the duration of observation

An estimation of the portfolio's overall value can be calculated through **customer equity** (CE) which amounts to the sum of all the CLVs. Since CE appears to be a good proxy of the firm's value, the firm's profit-maximization program can be written as:

$$\begin{aligned} \max_{\mathrm{p}} \quad & \mathrm{CE} = \sum_{i=1}^{N} \sum_{t=0}^{T} \frac{(p_t - c_t) r_{i,t}}{(1+a)^t} - AC_i \\ \text{s.t.} \quad & r_{i,t} \in [0,1] \\ & p_t > c_t \end{aligned} \tag{1.2}$$

where p is the vector of prices over all periods that the firm needs to optimize.

Chapter 2

Literature Review

Now the concepts of *portfolio*, *attrition* and *value* have been defined, it seems relevant to take a look at the literature on these notions. The literature review made in this chapter synthesises and analyses the available articles related to customer *portfolio* modelling, *attrition* analysis as well as customer *value* estimation. The review combines concepts from Economics, Econometrics and Data Science.

2.1 On customer portfolio

Portfolio management methods have been applied to an increasing number of areas over time. This term is originally used in finance by Markowitz (1952) with a view of managing equities. He develops a mathematical framework for assembling a portfolio of assets such that the expected return is maximized for a given level of risk. Markovitz's model is based on diversification which is the idea that owning different kinds of financial assets is less risky than owning only one type. His theory uses the variance of asset prices as a proxy for risk. Later in the 1970-80's, *portfolio* models are incorporated into corporate (Wind and Mahajan, 1981) and marketing (Day, 1977) strategies for profit-maximization via optimal resource allocation. Then, Capon and Glazer (1987) provide insights on efficient management for *portfolios* of technologies and study the complementarity between technological means mobilized by a firm. More recently, in the interest of improving relationships between the firm and its clients, the *portfolio* modelling approach have focused on effective customer relationship management (CRM). The following figure depicts the evolution of *portfolio* analysis through time.

In their article Thakur and Workman (2016) examine how a company can define the value of customers and segment these customers into *portfolios*. They explain how segmentation leads to better understanding of the relative importance of each customer to the company's total profit. The authors consider a

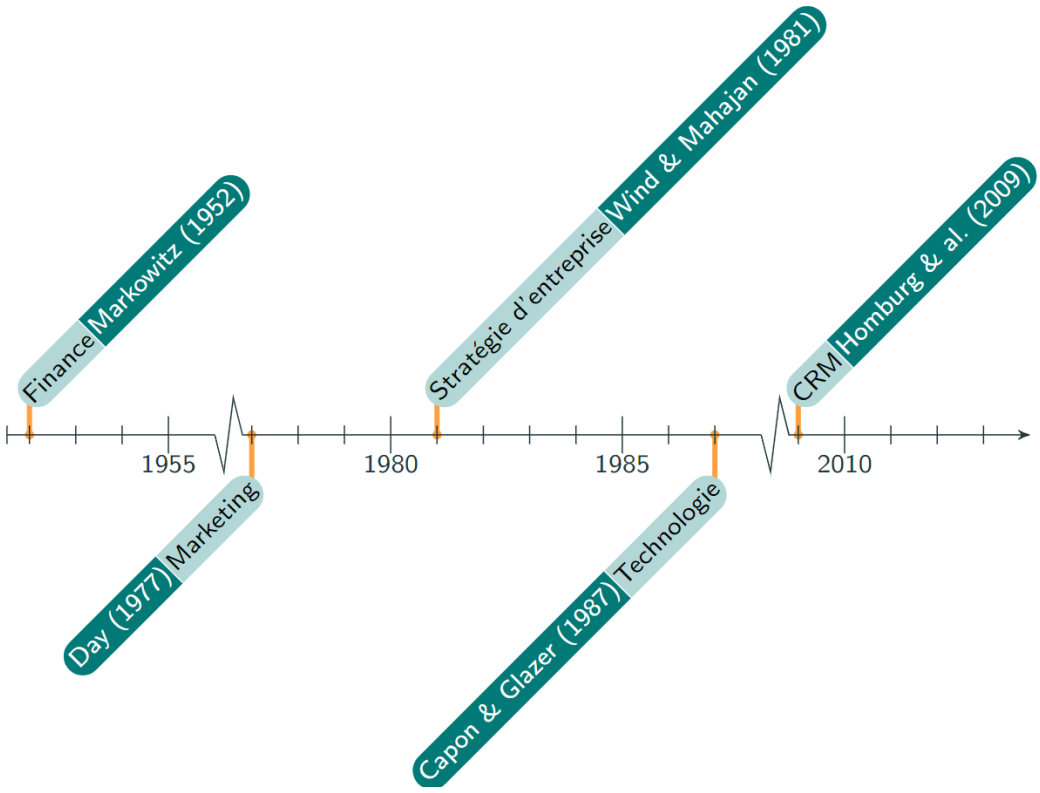

Figure 2.1: A timeline on the concept of portfolio

portfolio segmented into four groups of clients: *platinum*, *gold*, *silver* and *bronze* customers. The *portfolio* segmentation is based both on the cost to serve a client as well as the latter's *value* to the firm, as depicted by figure 2.2.

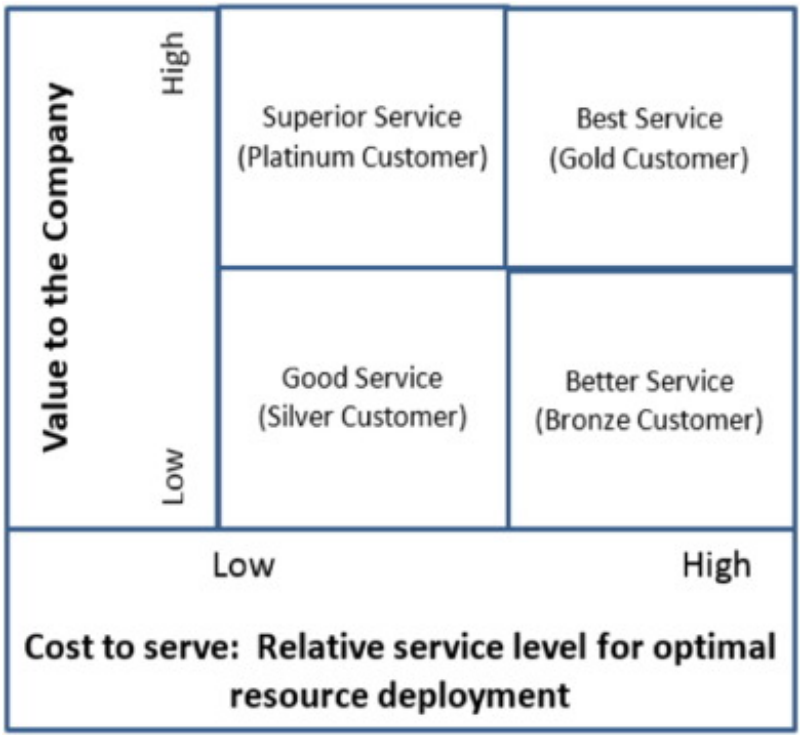

Figure 2.2: Csutomer Portfolio Management (CPM) Matrix

According to this repartition into four main groups, Thakur and Workman highlight three strategies the firm can launch in order to efficiently manage its *portfolio*. **Retention** aims to induce *platinum* customers into repeating their purchases as they have a large contribution to the firm's revenue. **Customer relationship development** can be used to encourage customers to advance and upgrade to a higher segment. Such a strategy can be efficient for customers with high preference for a certain product or those with potential to shift to higher margin products. Conversely, **customer elimination or filtering** strategies are set up by the firm to encourage bottom customers who cost more than they are worth to leave the *portfolio*.

As said in the introduction, an interesting improvement of *portfolio* analysis may be to add a temporal dimension to the models. Homburg et al. (2009) show that the dynamic approach minimizes the current bias of underestimating low-value clients and overestimating high-value ones.

2.2 On attrition

Attrition or churn has become a buzzword these last years. Churn analysis can be seen as an economic problem for three main reasons. Firstly, customers are in some way the firm's more precious asset. Secondly, the firm's resources in terms of customer relationship management are limited, so an efficient allocation needs to be deployed. Thirdly churn being a risk the firm has to cope with, it leads to asymmetric information from the firm's side. With the development of advanced Econometrics and Data Science, several methods can be implemented in order to estimate churn.

On the one hand, survival models are helpful in measuring customer lifetime. In her thesis, Pérez Marín (2006) applies duration analysis to model the behavior of customers from an insurance company with a threefold purpose:

- Identify factors influencing customer loyalty.
- Estimate the remaining lifetime of a client who has subscribed to multiple policies and cancelled one of them.
- Study the influence of covariates changing over time.

Her study is motivated by the importance of the insurer/policyholder relationship in a digitalized environment where the costs of searching for information are loawer and lower and the risk of *attrition* consequently higher. The author develops a two-part methodology to address the study's problematic. She begins by solely selecting insureds with at least two policies. Then, she fits a logit model to predict whether a policyholder will cancel their policies at the same time (type 1) or sequentially (type 2). She finally applies duration models on type 2 clients to determine the remaining time until all their policies are cancelled.

On the other hand, machine learning classification algorithms can be used for churn detection as illustrated by the work of Bellani (2019). Her objective is to develop a predictive model to detect customer churn in an insurance company while highlighting the key drivers of *attrition*. The underlying goals of her research paper are both minimizing revenue loss caused by churn and boosting the firm's competitiveness. Using data on vehicle insurance policies, Bellani incorporates features on the policyholders, the vehicles, the insurance policies as well as marketing data to predict the churn indicator variable. After missing data imputation and dimensionality reduction, the author falls back on undersampling to overcome the issue of unbalanced classes. There are indeed much more active than cancelled policies in the dataset. Her methodology works as follows:

- The set of active policies is divided into 7 groups equal in size to the number of cancelled policies.

- For each group of active policies, classification models (logistic regression, random forest and neural network) are trained on a subset of the original dataset including all the cancelled policies as well as the concerned group of active ones.
- For each model, the predictions are aggregated across the 7 subsets for the final prediction.
- Model selection is made by the means of the Kappa performance metric.

Ultimately when a customer leaves the firm's portfolio, it may worth it to consider all possible outcomes for the reason he churned. For instance, a client might leave their telecom company because of a bad service quality, or because of too high a price. In this context, competing risk analysis can be introduced since its main interest is to determine the reason why the client churned. In their recent article, Slof et al. (2021) try to predict both the likelihood of customer churn and the reasons for *attrition* using customer service data from a Dutch TSP. They estimate duration and competing risk models. In the competing risk model, three possible output states are considered: Controllable risk, Uncontrollable risk and Unknown risk. Each type of risk is assumed independent from another which means a client cannot be at high risk for two risks simultaneously. Besides, the authors implement a Latent Dirichlet Allocation model (see Bley et al. (2003) for more details) to identify the main topics in a set of emails sent by customers to the service center. Six topics are discovered by the algorithm and each of them is then incorporated as explanatory variable into the models. These topic variables increase the performance of both standard duration models and competing risk models for Controllable and Unknown risks. According to Slof, Frasincar, and Matsiiako, *"customers who churn due to the Controllable risk or due to the Unknown risk tend to call the customer service center with a specific problem, while customers who churn due to the Uncontrollable risk do not call the customer service center with a specific problem"*.

2.3 On customer value

In recent years, customer *portfolio* management (CPM) has focused on optimizing clients' *value* to the firm. The company's interest lies in knowing how much net benefit it can expect from a customer today. These expectations are then used to implement efficient marketing strategies to get the highest return on investment. To that end, two key metrics are estimated by firms: customer lifetime value (CLV) and customer equity (CE) (see part 1.3 in the introduction for definitions).

According to Blattberg and Deighton (1996), CLV is a temporal variable defined as the revenue derived from a customer minus the cost to the firm for maintaining the relationship with this very customer. As shown by Reinartz and Kumar. (2003), CLV modelling depends on the type of relationship a firm has with its

clients. In a contractual relationship, customer defections are observed which means that longer lifetime means higher customer value. Conversely, when the relationship is non-contractual, uncertainty arises between the customer's purchase behavior and lifetime.

With the development of data collection tools, companies have lots of customer-level data (or customer transaction data) at their disposal to measure CLV (Fader and al., 2005). Consequently, different modelling approaches can be adopted in order to estimate customer *value.*

Recency Frequency Monetary (RFM) models are considered the simplest strategy to measure CLV and customer loyalty (Gupta et al., 2006). They aim at targeting specific marketing campaigns at specific groups of customers to improve response rates. RFM models consist in creating clusters of clients based on three variables:

- *Recency* which is the time that has elapsed since customers' last activity with the firm.
- *Frequency* that is the number of times customers transacted with the brand in a particular period of time.
- *Monetary* that is to say the value of customers' prior purchases.

However, RFM models have a limited predictive power since they only predict clients' behavior for the next period.

In their article on CLV management, Borle and Singh (2008) draw the review of more advanced modelling techniques that can be implemented to estimate customers' *value.* A popular method to estimate customer lifetime value is the negative binomial distribution (NBD) - Pareto (Fader and al., 2005) which helps solving the lifetime uncertainty issue. The model takes past customer purchase behavior as input such as the number of purchases in a specific time window and the date of last transaction. Then the model outputs a repurchase probability as well as a transaction forecast for each individual. In Borle and Singh's research paper, a hierarchical bayesian model is implemented with a view to jointly predict customer's churn risk and spending pattern. Here, the main advantage of using a bayesian approach is to give priors on CLV's drivers. The study is based on data coming from a membership-based direct marketing company where firm/client relationships are non-contractual. In other words, the times of each customer joining the membership and terminating it are known once these events happen. Thus the implementation of a sophisticated estimation strategy is justified.

In our study, emphasize is placed on estimating the overall value of a customer *portfolio.* The methodology we develop is based on a research paper written by our Econometrics teacher Alain Bousquet, whose goal is to provide tools for an efficient management of patent *portfolios* (Bousquet, 2021). The main idea is to consider each patent as an asset with a related value which can generate income

if this very patent is exploited. The author emphasizes the importance to focus on the *portfolio*'s **variance** on top of its expected value. Specifically, he explains that the variability in the probability of success in the exploitation of patents leads to a decrease in the overall risk to which the *portfolio* is exposed. This modelling approach can be transposed to customer *portfolio* analysis with the customer's *value* corresponding to the CLV and the probability of exploitation being the opposite of the risk of *attrition*. In this context, CLV can be estimated either with techniques mentioned above or regression methods. The customer's risk of churn can be modelled with duration models or machine learning techniques as evoked in 2.2. With this econometric framework, it is expected that customer heterogeneity be a key factor in the total variance of the portfolio's *value*.

Chapter 3

Duration models

This chapter presents theoretical basis of the models that are used to model customer *portfolios.* As customer lifetime in a *portfolio* is usually represented by the time to churn, duration models are adapted to the data we have at our disposal. Thus, this part focuses on introducing standard survival techniques.

3.1 Definition

According to Cameron and Trivedi (2005), duration models (also called survival models) aims at measuring the time spent in a certain state before transitioning to another state. In Econometrics,

- a **state** corresponds to the class in which an individual i is at time t.
- a **transition** is movement from one state to another.
- a **duration** measures the time spent in a certain state and is also called a **spell** length.

Since measuring the time until the event is needed for multiple purposes, duration analysis is used in a variety of economic sectors as depicted by the following table.

Economic sector	Purpose
Macroeconomics	Length of unemployment spells
Insurance	Risk analysis to offer a segmented pricing
Engineering	Time until a device breaks down
Epidemiology	Survival time of a virus
Churn analysis	**Time until a customer leaves the portfolio**

3.2 Censoring and Truncation

When dealing with survival data, some observations are usually **censored** meaning they are related to spells which are not completely observed. Duration data can also suffer from a selection bias which is called **truncation**.

3.2.1 Censoring mechanisms

Left-censoring occurs when the event of interest occurs before the beginning of the observation period. For example, an individual is included in a study of unemployment duration at t_0. At that time he has already been unemployed for a period but he cannot recall exactly the duration of this period. If we observe that he finds a job again at t_1, we can only deduce that the duration of unemployment is bigger than $t_1 - t_0$, this individual is consequently left-censored. Observation 2 on figure 3.1 is associated with a left-censored spell (Liu, 2019).

A spell is considered **right-censored** when it is observed from time t_0 until a censoring time t_c as illustrated by observation 4 on figure 3.1. For instance, the lifetime related to a customer who has not churned at the end of the observation period is right-censored. Let us note X_i the duration of a complete spell and C_i the duration of a right-censored spell. We also note T_i the duration actually observed and δ_i the censoring indicator such that $\delta_i = 1$ if the spell is censored. Then $(t_1, \delta_1), ..., (t_N, \delta_N)$ are the realizations of the following random variables:

$$\begin{aligned} T_i &= \min(X_i, C_i) \\ \delta_i &= \mathbf{1}_{X_i > C_i} \end{aligned} \tag{3.1}$$

3.2.2 Selection bias

Survival data suffers from a **selection bias** (or truncation) when only a subsample of the population of interest is studied. A customer entering the firm's *portfolio* after the end of the study is said to be **right-truncated**, whereas a client who has left the *portfolio* before the beginning of the study is considered **left-truncated**. Mathematically, a random variable X is truncated by a subset $A \in \mathbb{R}^+$ if instead of $\Omega(X)$, we solely observe $\Omega(X) \bigcap A$. On figure 3.1, the first and fifth observations suffers from a selection bias.

3.3 Probabilistic concepts

In survival analysis, the response variable denoted T is a time-to-event variable. Instead of estimating the expected failure time, survival models estimate the **survival** and **hazard rate** functions which depend on the realization of T.

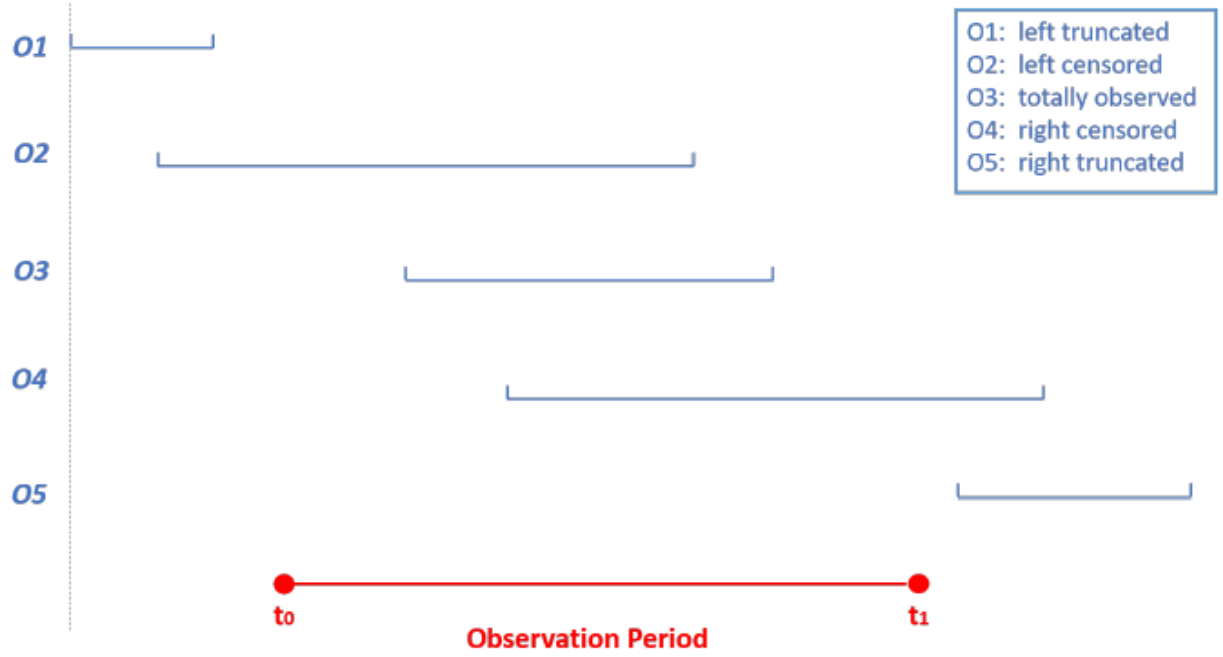

Figure 3.1: Censored and truncated data

3.3.1 Survival function

The survival function $S(t)$ represents the probability that the considered event occurs after time t. For instance, $S(t)$ can measure the probability that a given customer survives in the *portfolio* at least until time t. Mathematically, the survival function is defined as:

$$S(t) = P(T > t) = 1 - F(t) \tag{3.2}$$

where $F(t)$ is the cumulative distribution function.

3.3.2 Hazard and Cumulative Hazard functions

Another key concept in duration analysis is the hazard function $\lambda(t)$ which approximates the probability that the event occurs at time t. For instance, $\lambda(t)$ can measure the probability that a given individual leaves the firm *portfolio* at time t. Formally, it is expressed as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \tag{3.3}$$

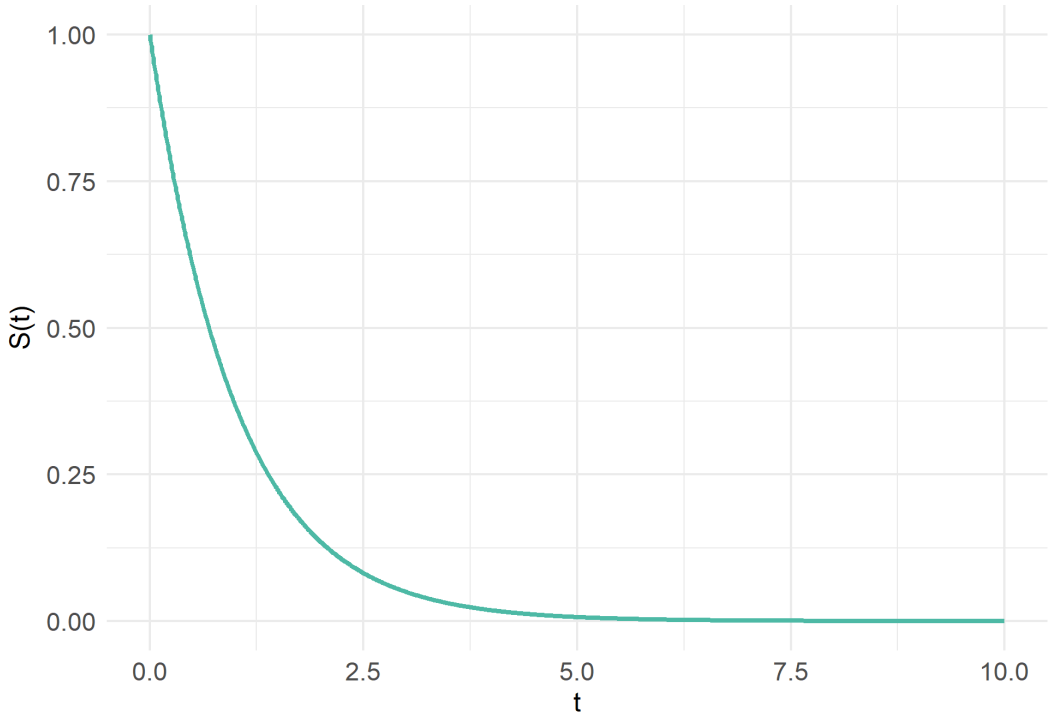

Figure 3.2: Survival function $S_T(t)$ with $T \sim \mathcal{E}(1)$

Using the Bayes formula, equation (3.3) can also be written as (see proof (6.3) in the appendix):

$$\lambda(t) = \frac{-\mathrm{d}\ln\left(S(t)\right)}{\mathrm{d}t} \tag{3.4}$$

Finally, integrating the instantaneous hazard function gives the cumulative hazard function which can be more precisely estimated than the hazard function (Cameron and Trivedi, 2005) and is defined as:

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s = -\ln\left(S(t)\right) \tag{3.5}$$

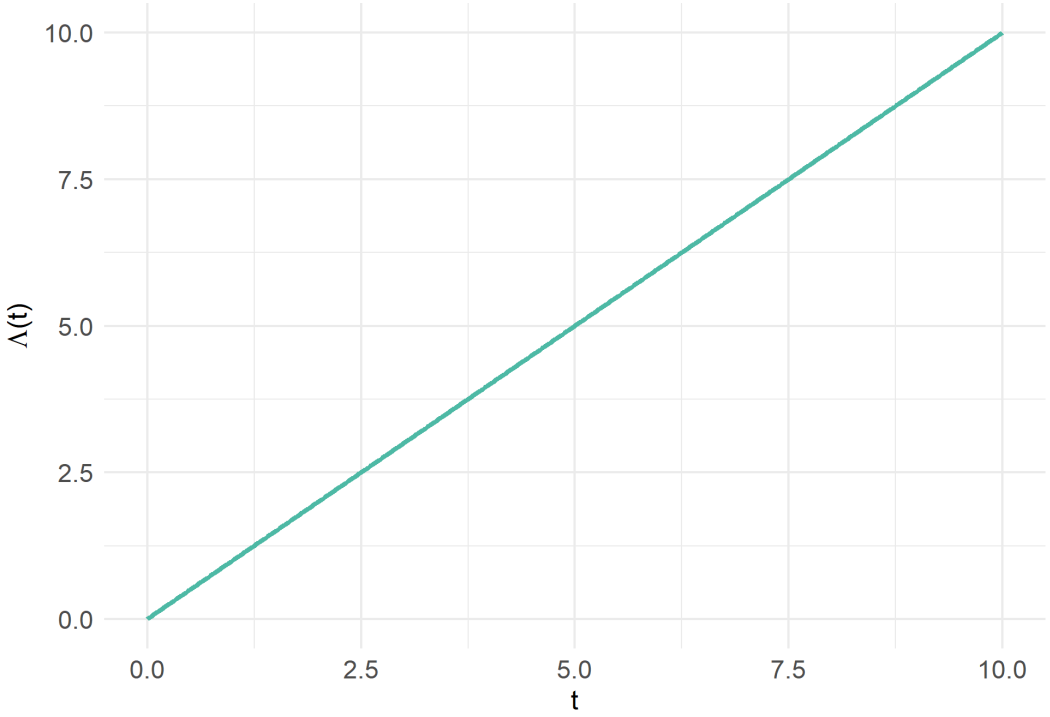

Figure 3.3: Cumulative Hazard function $\Lambda_T(t)$ with $T \sim \mathcal{E}(1)$

Thus, **the hazard, survival and cumulative hazard functions** are three mathematical functions which describe the same distribution.

3.4 Nonparametric models

When dealing with duration data, these methods are helpful to have a general overview of the raw (or unconditional) hazard. Nonparametric models are rather used for data description than prediction. No explanatory variable is included in these models except for treatment variables such as the type of contract a customer has subscribed.

3.4.1 Notations

Let us consider a sample with N observations with k ordered discrete failure times (e.g. a failure can be a churn event), such that $\forall j \in [\![1; k]\!]$:

- t_j the j^{th} discrete failure time,
- d_j the number of spells terminating at t_j,
- m_j the number of right-censored spells in the interval $[t_j, t_{j+1}]$,
- r_j the number of exposed durations right before time t_j i.e. at time t_j^-, such that:

$$r_j = (d_j + m_j) + \cdots + (d_k + m_k) = \sum_{l|l \geq j} (d_l + m_l) \tag{3.6}$$

3.4.2 Hazard function estimator

As the instantaneous hazard at time t_j is defined as $\lambda_j = P[T = t_j | T \geq t_j]$, a trivial estimator of λ_j is obtained by dividing the number of durations for which the event is realized at t_j by the total number of exposed durations at time t_j^-. Formally, it is expressed as:

$$\hat{\lambda}_j = \frac{d_j}{r_j} \tag{3.7}$$

3.4.3 Kaplan-Meier estimator

Once the hazard function estimator computed, the discrete-time survivor function can be estimated using the Kaplan-Meier product-limit estimator. To estimate the survival at time t, this estimator computes the joint probability that a spell stays in the same state until t (e.g. remaining loyal to a firm until a certain time). This method is based on conditional probabilities and the survival function estimate is defined as:

$$\widehat{S}(t) = \Pi_{j|t_j \leq t}(1 - \hat{\lambda}_j) = \Pi_{j|t_j \leq t} \frac{r_j - d_j}{r_j} \tag{3.8}$$

When plotting the survival curve after having performed the Kaplan-Meier estimation, confidence bands are also added to the plot in order to reflect sampling variability (Cameron and Trivedi, 2005). The confidence interval of the survival function $\hat{S}(t)$ is derived from the estimate of the variance of $S(t)$ which is obtained by the Grenwood estimate as in equation (3.9).

$$\widehat{\mathrm{V}}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{j|t_j \leq t} \frac{d_j}{r_j(r_j - d_j)} \tag{3.9}$$

3.4.4 Nelson-Aalen estimator

The cumulative hazard function estimate is given by the Nelson-Aalen estimator which consists in summing up the hazard estimates for each discrete failure time.

$$\hat{\Lambda}(t) = \sum_{j|t_j \leq t} \hat{\lambda}_j = \sum_{j|t_j \leq t} \frac{d_j}{r_j} \tag{3.10}$$

Exponentiating $\hat{\Lambda}(t)$, one can obtain a second estimate of the survival function (see proof (6.4) in the appendix):

$$\tilde{S}(t) = \exp\big(-\hat{\Lambda}(t)\big) \tag{3.11}$$

3.5 Parametric models

The nonparametric estimation is undoubtedly useful when it comes to have a general overview on the survival data. However, one may want to model the hazard and survivor functions with a functional form in which unknown parameters need to be optimized.

Parametric estimation has a twofold purpose that is to implement a robust model to estimate the risk that a specific event occurs while identifying the variables (or covariates) which best explain this risk.

When implementing parametric models, λ, S and Λ are expressed based on the chosen parametric form. The instantaneous hazard function can either be constant or monotone.

In our study we assume that the explanatory variables are time-constant as we do not have dynamic data at our disposal. Thus, solely time-invariant duration models are presented.

3.5.1 Constant hazard (exponential model)

The exponential distribution models the time between events in a Poisson process and has the key property of being *memoryless*. Let us note T a time-to-event variable such that $T \sim \mathcal{E}(\theta)$ where θ is the rate parameter. In this context, *memorylessness* can be defined as follows:

$$\mathbb{P}(T > t + s | T > t) = \mathbb{P}(T > s) \tag{3.12}$$

$\forall t \geq 0\ ,\ \theta > 0$ the density, hazard and survival functions can be expressed as:

$$\begin{aligned} f_\theta(t) &= \theta e^{-\theta t} \\ \lambda_\theta(t) &= \theta \\ S_\theta(t) &= e^{-\theta t} \end{aligned} \tag{3.13}$$

Thus, the exponential distribution is characterized by a **constant** hazard function which is a consequence of the *memorylessness* property.

3.5.2 Monotone hazard

Weibull model

The Weibull distribution is a less restrictive generalization of the exponential distribution defined by a shape parameter ν and a scale parameter θ.

$\forall t \geq 0$ and $\nu,\ \theta > 0$ the density, hazard and survival functions can be expressed as:

$$\begin{aligned} \lambda_{\nu,\theta}(t) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \\ S_{\nu,\theta}(t) &= \exp\left(- \left(\frac{1}{\theta}\right)^{\nu} t \right) \end{aligned} \tag{3.14}$$

The instantaneous hazard function $\lambda_{\nu,\theta}$ is monotonic **decreasing** if $\nu \in [0, 1]$. For instance, the *attrition* risk may decrease as the customer's duration in the *portfolio* increases. In this context, the client gets more and more loyal to the firm. If $\nu = 1$, the hazard rate is constant and $T \sim \mathcal{E}(\theta)$. Converlsely, the hazard function is monotonic **increasing** if $\nu > 1$. This can be the case when

customers tend to continuously search for information on the firm's competitors, thus becoming more likely to churn as time goes by.

Figure 3.4 illustrates the hazard and survivor functions associated to a Weibull-distributed variable T. The two curves' shape depend both on the shape (ν) and scale (θ) parameters. Some remarks can be made looking at the two plots. When $\nu < 1$ the hazard function is decreasing meaning that the risk of the event occurring decreases as time goes by. When $\nu > 1$ the hazard function is convex increasing which indicates that a marginal increase in time leads to an increase of over one unit in the the hazard function. The higher the shape parameter, the more increasing the hazard function. When $\nu = \theta = 1$, it can be noted that the Weibull distribution corresponds to the exponential distribution (see figures 3.2 and 3.3).

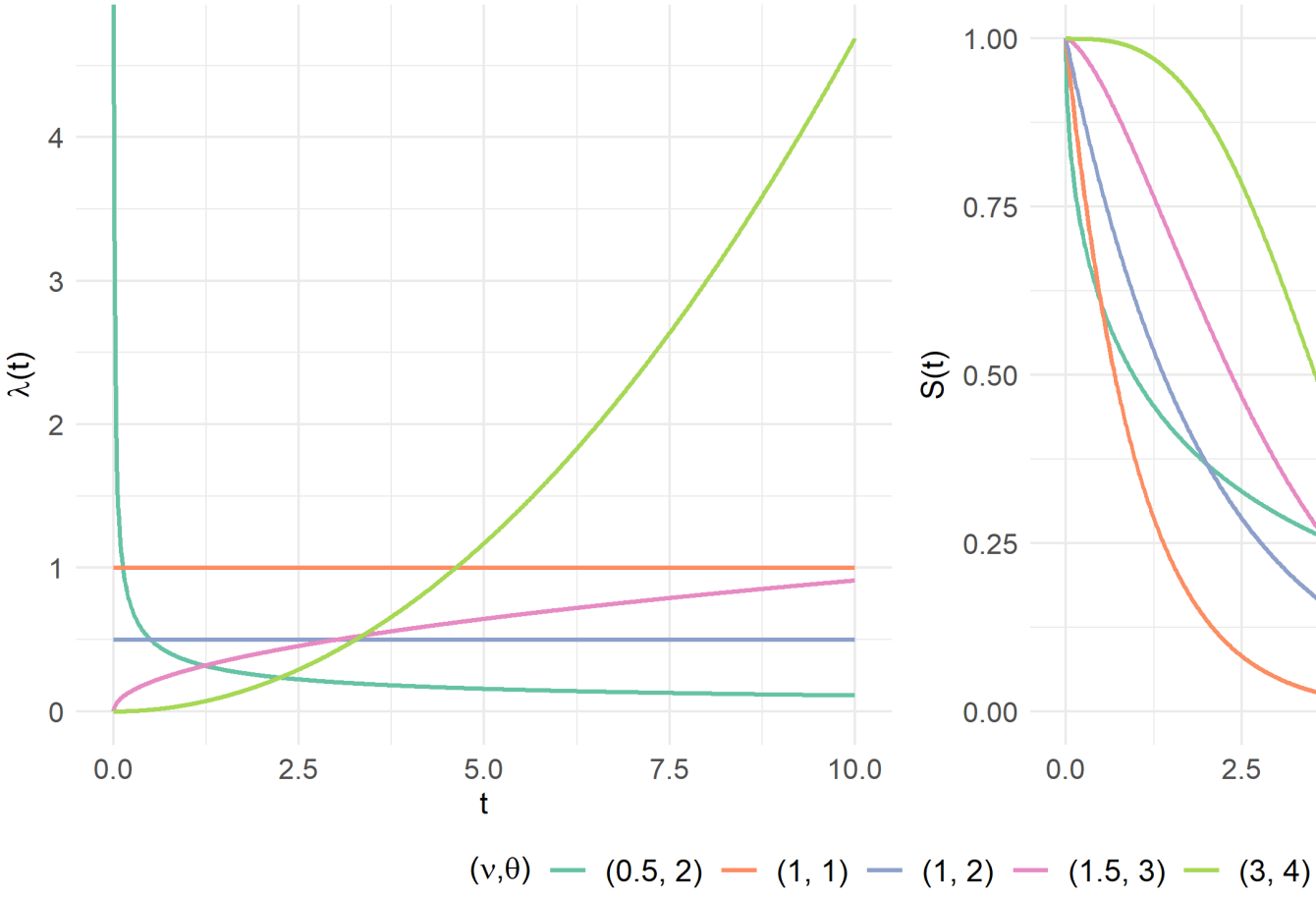

Figure 3.4: Hazard and Survival functions with $T \sim \mathcal{W}(\nu, \theta)$

Other models

Different probabilistic distributions can be chosen to model the hazard and survival functions related to a time-to-event variable with monotone hazard. The Gompertz model is usually used for mortality data in biostatistics. As for the gamma model, it depends both on the gamma and inverse-gamma distributions and is also based on shape and scale parameters.

3.5.3 Concave and convex hazard

When the hazard function does not evolve in a monotonic fashion, the distributions introduced above are limited. The generalized Weibull model appears to be a good choice to estimate phenomena with concave or convex hazards. It is based on three parameters: ν (shape), θ (scale) and γ. When $\gamma = 1$, the generalized Weibull becomes the Weibull distribution $\mathcal{W}(\nu, \theta)$.

3.6 Semi-parametric estimation

3.6.1 *Proportional Hazards* models

Parametric models assume that the baseline (or raw) hazard follows a specific distribution. This assumption can be sometimes too restrictive and semi-parametric models can be more adapted to describe the duration data.

In *proportional hazards* (PH) models, the instantaneous risk function is **proportional** to the baseline hazard $\lambda_0(t, \alpha)$ modulo a **scaling factor** depending on the covariates $\phi(\mathbf{x}, \beta)$. These models allow to generalize the basic survival models to a survival regression model which permits to take individuals' heterogeneity into consideration (Harrell, 1984). The general mathematical formulation is expressed as follows:

$$\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\phi(\mathbf{x}, \beta) \tag{3.15}$$

Note that when the function form of $\lambda_0(t, \alpha)$ is known, we are in the case of parametric estimation. For instance, the exponential, Weibull and Gompertz models are PH models since their respective hazards are function of some covariates.

What does *proportional hazards* mean?

PH models are said to be proportional as the relative hazard ratio between two individuals i and k does not vary over time, such that:

$$\frac{\lambda(t|\mathrm{x_i})}{\lambda(t|\mathrm{x_k})} = \frac{\phi(\mathrm{x_i}, \beta)}{\phi(\mathrm{x_k}, \beta)} \tag{3.16}$$

The formulation stated in equation (3.16) needs to be verified when one wants to fit a PH model to real-life data and is only valid in the case of time-constant covariates.

Marginal effects

In *proportional hazards* models, the marginal effect of covariate x_p on the hazard function can be easily derived since this computation only requires knowledge on β. As shown in Cameron and Trivedi (2005), a one-unit increase in the p^{th} covariate leads to the following variation in the hazard function *ceteris paribus*:

$$\frac{\partial \lambda(t|\mathbf{x}, \beta)}{\partial x_p} = \lambda(t|\mathbf{x}, \beta) \frac{\partial \phi(\mathbf{x}, \beta)/\partial x_p}{\phi(\mathbf{x}, \beta)} \tag{3.17}$$

Thus the new hazard after variation of the p^{th} covariate is the original hazard times the effect of x_p on the model's regression part.

Partial likelihood estimation

The vector of parameters β related to the regression part of the PH model is estimated by partial likelihood maximization. The method's principle consists in only estimating the regression's parameters β by considering the baseline hazard λ_0 as noise. If desired an estimate of the baseline hazard can be recovered after estimation of β using, for instance, the Nelson-Aalen estimator (see part 3.4). Cox's intuition is that no information can be retrieved from the intervals during which no event has occurred and that it is conceivable that λ_0 is null in these intervals. Thus, solely the set of moments when an event occurs are considered in the estimation method.

In order to derive the partial likelihood function, let us note t_j the j^{th} discrete failure time in an N-sample with $j \in [\![1; k]\!]$, such that:

- $t_1 < t_2 < \cdots < t_k$,
- $D(t_j) = \{l : t_l = t_j\}$ is the set of spells completed at t_j with $\#D(t_j) = d_j$,
- $R(t_j) = \{l : t_l \geq t_j\}$ is the set of spells at risk at t_j.

The contribution of a spell in $D(t_j)$ to the likelihood function equals the conditional probability that the spell ends at t_j given it is exposed at that specific time and can be written as (see Cameron and Trivedi (2005) and proof (6.5) for more details):

$$\mathbb{P}[T_j = t_j | R(t_j)] = \frac{\phi(\mathrm{x_j}, \beta)}{\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)} \tag{3.18}$$

Given k discrete failure times are considered and that for each of those there is a set $D(t_j)$ of completed spells, Cox defines the partial likelihood function as the joint product of the probability expressed in (3.18), such that:

$$\mathcal{L}_p = \Pi_{j=1}^{k} \frac{\Pi_{m \in D(tj)} \ \phi(\mathrm{x_j}, \beta)}{\Big[\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)\Big]^{d_j}} \tag{3.19}$$

The latter formulation of the partial likelihood function is explained in more details in proofs (6.6) and (6.7) in the appendix.

3.6.2 Cox PH model

The Cox *proportional hazards* model is the most popular for the analysis of duration data. This model is said to be semi-parametric as it makes no assumption regarding the nature of the baseline hazard function $\lambda_0(t)$. The parametric part only relies in the modelling of the effect of some covariates on the hazard function $\lambda(t)$. The relationship between the vector of covariates and the log hazard is linear and the parameters can be estimated by maximizing the partial likelihood function. The Cox PH model solely assumes that predictors act multiplicatively on the hazard function. The model is formulated as in equation (3.15) with the exponential function as link between the hazard and the covariates i.e. $\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\mathrm{e}^{\mathbf{x}'\beta}$.

3.7 Machine Learning for Survival Data

In the previsous sections, some important models for duration data have been introduced. Here, emphasize is placed on machine learning algorithms that can also be implemented to predict a time-to-event variable such as the time to churn.

3.7.1 Survival Trees

Traditional decision trees, also called CART (Classification And Regression Trees), segment the feature space into multiple rectangles and then fit a simple model to each of these subsets as shown by figure 3.5 (Scholler, 2021b). The algorithm is a recursive partitioning which requires a criterion for choosing the

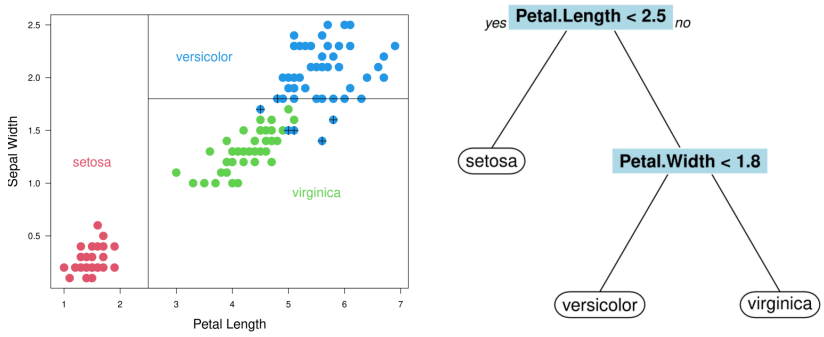

Figure 3.5: Clasification decision tree

best split, another criterion for deciding when to stop the splits and a rule for predicting the class of an observation.

Survival tree (LeBlanc and Crowley, 1993) is the adapted version of CART for duration data. The objective is to use tree based binary splitting algorithm in order to predict hazard rates. To that end, survival time and censoring status are introduced as response variables. The splitting criteria used for survival trees have the same purpose than the criteria used for CART that is to say maximizing between-node heterogeneity or minimizing within-node homogeneity. Nonetheless, node purity is different in the case of survival trees as a node is considered pure if all spells in that node have similar survival duration. The most common criterion is the **logrank test** statistic to compare the two groups formed by the children nodes. For each node, every possible split on each feature is being examined. The best split is the one maximizing the survival difference between two children nodes. The test statistic is χ^2 distributed which means the higher its value, the higher between-node variability so the better the split. Let $t_1, \ldots, t_k$ be the k ordered failure times. At the j^{th} failure time, the logrank statistic is expressed as (Segal, 1988):

$$\chi^2_{\text{logrank}} = \frac{\left[\sum_{j=1}^{k} (d_{0j} - r_{0j} \times d_j / r_j)\right]^2}{\sum_{j=1}^{k} \frac{r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \tag{3.20}$$

3.7.2 Random Survival Forests (RSF)

This algorithm is proposed by Ishwaran and al. (2011) and is an ensemble of decision trees for the analysis of right-censored survival data. As random forests used for regression and classification, RSF are based on **bagging** which implies that B bootstrap samples are drawn from the original data with 63% of them in the bag data and the remaining part in the out-of-bag (OOB) data. For

each bootstrap sample, a survival tree is grown based on p randomly selected features. Then, the parent node is split using the feature among the selected ones that maximizes survival difference between children nodes. Each tree is grown to full size and each terminal node needs to have no less than d_0 unique events. The cumulative hazard function (CHF) is computed for each tree using the Nelson-Aalen estimator such as:

$$\widehat{H_l}(t) = \sum_{t_{j,l}<t} \frac{d_{j,l}}{r_{j,l}} \tag{3.21}$$

where $t_{j,l}$ is the j^{th} distinct event time in leaf l, $d_{j,l}$ the number of events completed at $t_{j,l}$ and $r_{j,l}$ the number of spells at risk at $t_{j,l}$.

All the CHFs are then averaged to obtain the bootstrap ensemble CHF and prediction error is finally computed on the OOB ensemble CHF.

3.7.3 Cox Boosting

Boosting is an ensemble method which combines several weak predictors into a strong predictor. The idea of most boosting methods is to train predictors sequentially, each trying to correct its predecessor. Cox boosting (Binder and Schumacher, 2008) is designed for high dimension survival data and has the purpose of feature selection while improving the performance of the standard Cox model. The key difference with gradient boosting is that Cox boosting does not update all coefficients at each boosting step, but only updates the coefficient that improves the overall fit the most. The loss function is a penalized version of the Cox model's log-likelihood (see equation (3.19) for the likelihood function of the Cox model). Cox boosting helps measuring variable importance as the coefficients associated to more representative variables will be updated in early steps.

3.8 Performance metrics

3.8.1 Concordance index (C-index)

C-index is a goodness of fit measure for models which produce risk scores. It is commonly used to evaluate risk models in survival analysis, where data may be censored.

Consider both the observations and prediction values of two instances $(y_1; \widehat{y}_1)$ and $(y_2; \widehat{y}_2)$. y_i and $\widehat{y}_i$ represent respectively the actual observation time and the predicted time. Mathematically, the C-index is defined as the probability to well predict the order of event occurring time for any pair of instances.

$$c = \mathbb{P}(\hat{y}_1 > \hat{y}_2 | y_1 > y_2) \tag{3.22}$$

Another way to write the C-index metric is to compute the ratio between concordant pairs and the total number of pairs. Consider individual i and let T be the time-to-event variable and η_i the risk score assigned to i by the model. We say that the pair (i, j) is a concordant pair if $\eta_i > \eta_j$ and $T_i < T_j$, and it is a discordant pair if $\eta_i > \eta_j$ and $T_i > T_j$. If both T_i and T_j are censored, then this pair is not taken into account in the computation. If T_j is censored, then:

- If $T_j < T_i$ the pair (i, j) is not considered in the computation since the order cannot be determined.
- If $T_j > T_i$, the order can be determined and (i, j) is concordant if $\eta_i > \eta_j$, discordant otherwise.

Equation (3.22) can then be rewritten as follows:

$$\begin{aligned} c &= \frac{\#\text{concordant pairs}}{\#\text{concordant pairs} + \#\text{discordant pairs}} \\ c &= \frac{\sum_{i \neq j} \mathbf{1}_{\eta_i < \eta_j} \mathbf{1}_{T_i > T_j} d_j}{\sum_{i \neq j} \mathbf{1}_{T_i > T_j} d_j} \end{aligned} \tag{3.23}$$

with d_j the event indicator variable.

The concordance index ranges between 0 and 1. A C-index below 0.5 indicates a very poor model. A C-index of 0.5 means that the model is rather a non-informative model making random predictions. A model with C-index 1 makes perfect prediction. Generally, a C-index higher than 0.7 indicates a good performance.

3.8.2 Brier score

The Brier score is another statistical metric for evaluating duration models' performance and is defined as the mean squared error between the estimated survival probability and the observed survival at time t:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{1}_{\{t_i > t\}} - \hat{S}(t|\mathrm{x}_i) \right)^2 \tag{3.24}$$

The Cox *proportional hazards* model is the most popular for the analysis of duration data. This model is said to be semi-parametric as it makes no assumption regarding the nature of the baseline hazard function $\lambda_0(t)$. The parametric

part only relies in the modelling of the effect of some covariates on the hazard function $\lambda(t)$. The relationship between the vector of covariates and the log hazard is linear and the parameters can be estimated by maximizing the partial likelihood function. The Cox PH model solely assumes that predictors act multiplicatively on the hazard function. The model is formulated as in equation (3.15) with the exponential function as link between the hazard and the covariates i.e. $\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\mathrm{e}^{\mathbf{x}'\beta}$.

Chapter 4

Data Mining methods

While the previous chapter introduces the fundamentals related to survival analysis, this part focuses on data mining techniques used in our modelling approach. On the on hand, multiple correspondence analysis is introduced as it is used in the following chapter with a view of converting categorical features into continuous principal axes. On the other hand, this section centers on unsupervised classification which is a useful method when it comes to segmentation.

4.1 Mutliple Correspondence Analysis (MCA)

4.1.1 Definition

Multiple Correspondence Analysis is a dimension reducing method which takes multiple categorical variables and seeks to identify associations between levels of those variables. MCA aims at highlighting features that separate classes of individuals, while determining links between variables and categories. To that end, MCA keeps the core information by the means of principal components which are projected axes (Scholler, 2021a).

4.1.2 Complete disjunctive table

MCA can be applied on data stored in a complete disjunctive table which is an indicator matrix.

with,

- I the number of individuals,
- J the number of variables,

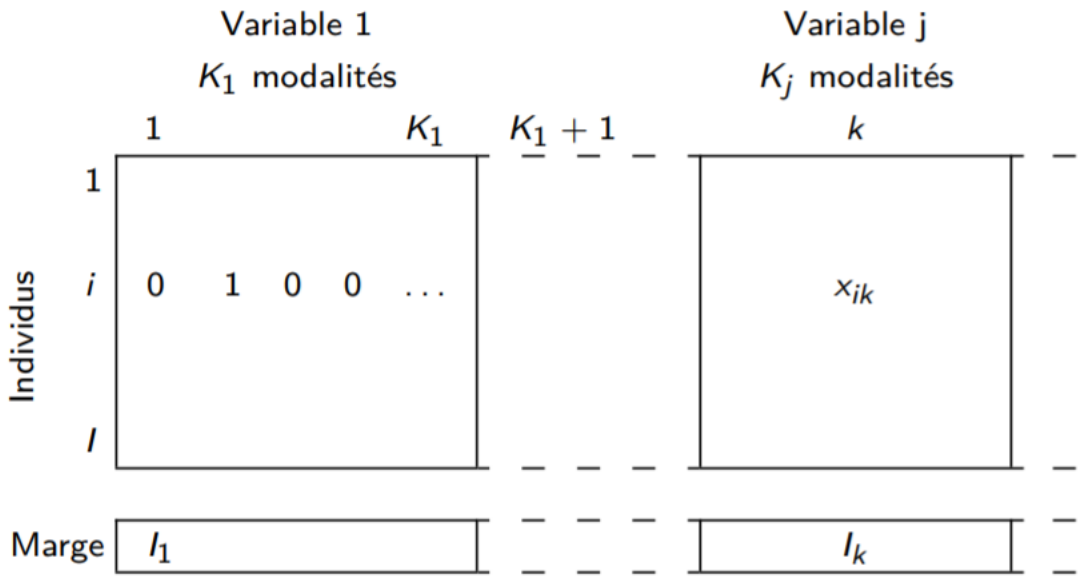

Figure 4.1: Complete disjunctive table

- K_j the number of categories in the j^{th} variable,
- I_k the number of individuals with the k^{th} category.

4.1.3 Distances

The individuals' analysis processed by MCA relies on the distance between individuals which is computed as follows for 2 data points i and l:

$$d^2(i;l) = \frac{1}{J}\sum_{k=1}^{K}\frac{I}{I_k}(x_{ik} - x_{lk})^2 \tag{4.1}$$

The distance between two categories j and k allows to determine how close they are and is calculated as follows:

$$d^2(j;k) = \frac{I}{I_k I_j} \times I_{k\neq j} \tag{4.2}$$

with $I_{k\neq j}$ the number of individuals with one and only one of the j or k categories.

4.1.4 Algorithm

1. The axes' origin is placed at the individuals point cloud's barycenter;
2. A sequence of orthogonal axes is seeked so as to maximize the data's projected inertia;
3. These orthogonal projections are represented onto a plan made up of principal components, (F_1, F_2) being the first projected plan.

4.2 Unsupervised classification

Clustering is an unsupervised learning technique that groups similar data points such that the points in the same group are more similar to each other than the points in the other groups. The group of similar data points is called a cluster. It is said to be unsupervised as the groups are not labeled but discovered by the algorithm. Consequently, it may not be possible to find separate clusters of data points.

4.2.1 Hierarchical Clustering on Principal Components (HCPC)

The HCPC approach combines the three standard methods used in multivariate data analysis:

- Principal Components methods such as MCA,
- Agglomerative hierarchical clustering,
- Consolidation by k-means partitioning.

4.2.2 Agglomerative Hierarchical Clustering (AHC)

Algorithm

1. Each data point is initially considered an individual cluster;
2. Compute the proximity matrix which represents the distances, taken pairwise, between the elements of a set;
3. Merge the two closest clusters and update the proximity matrix until only a single cluster remains.

Distance between two observations

The choice of the distance metric is a critical step in clustering. It defines how the similarity of two elements (x, y) is calculated and it will influence the shape of the clusters. There are several distance measures such as *Euclidean*, *Manhattan*, χ^2, etc. In our study the classical *Euclidean* distance metric is chosen and is defined as follows:

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.3}$$

Similarities between clusters - Ward method

Calculating the similarity between two clusters is determining to merge the clusters. Here, the Ward method (Scholler, 2020) is chosen and it consists in merging groups which drive down as little as possible the within inertia, ie the homogeneity of clusters. Mathematically, Ward criterion is defined as follows:

$$\Delta(A, B) = \frac{1}{n} \times \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B) \tag{4.4}$$

with g_A et g_B the clusters' barycenters (the mean point) et n_A et n_B the clusters' frequencies.

Ward method needs to be minimized and tends to group together clusters which are close to each other in terms of barycenters, as well as small frequency clusters.

Ward method leads to choose the optimal number of clusters. Let Δk be the increase in within inertia when going from $k+1$ groups to k groups. Then, the proper number of classes k^* can be such that $\frac{\Delta k-1}{\Delta k}$ is as little as possible.

4.2.3 The k-means algorithm

k data points are chosen as initial centers.The following steps are repeated until the clusters identified are homogeneous enough or until a fixed number of iterations:

1. The distances between the data points and the centers are computed;
2. Each data point is assigned to the nearest center, ;
3. The k previous centers are replaced by the barycenters of the k classes identified during the previous step.

Chapter 5

Data

In this chapter, we introduce the kaggle dataset related to customers of a fictional telecommunications service provider (TSP). In this duration dataset, the `Churn_Value` status variable indicates whether the customer left the firm's *portfolio* within the last month while the `Tenure_Months` variable stands for the duration actually observed. Besides customer *value* can be approximated by the `CLTV` variable.

5.1 General Overview

The data set used in this study contains 29 variables and 7032 customers from a telecom firm. For each client, the data includes:

- **Demographic** information: `CustomerID`, `City`, `Zip_Code`, `Latitude`, `Longitude`, `Gender`, `Senior_Citizen`, `Partner` and `Dependents`.
- **Customer account** information: `Tenure_Months`, `Contract`, `Paperless_Billing`, `Payment_Method`, `Monthly_Charges`, `Total_Charges`, `Churn_Label`, `Churn_Value`, `Churn_Score`, `CLTV`, `Churn_Reason`.
- **Services** information: `Phone_Service`, `Multiple_Lines`, `Internet_Service`, `Online_Security`, `Online_Backup`, `Device_Protection`, `Tech_Support`, `Streaming_TV`, `Streaming_Movies`.

Table 5.1: Interesting variables for the 5 first customers in the data set

CustomerID	Monthly_Charges	Internet_Service	Tenure_Months	Churn_Value
1	53.85	DSL	2	1
2	70.70	Fiber optic	2	1
3	99.65	Fiber optic	8	1
4	104.80	Fiber optic	28	1
5	103.70	Fiber optic	49	1
6	55.20	DSL	10	1

As shown by table 5.1, the `Churn_Value` status variable indicates whether the customer left the firm's *portfolio* within the last month and `Tenure_Months` is the duration variable.

Since the purpose of our study relies in estimating the overall value of this fictional firm's *portfolio*, two groups of target variables can be considered. On the one hand `Churn_Value` and `Tenure_Months` permit to determine whether a customer is active in the *portfolio*. They are used as response variables in the survival models. On the other hand, `Monthly_Charges` variable indicates the price paid by customers each month and may be used to derive a customer raw value. Even though the `CLTV` variable represents each customer's *value* through measurement of customer lifetime value, we do not have any information on its calculation. Thus, it is not used in the model developed in the next chapter.

5.2 `Churn_Value` and `Tenure_Months`

As the combination of these two features form the response variable in the duration models, a relevant approach to have an overall description of the risk of *attrition* may be to draw the raw survival curves depending on treatment variables. Pearson's χ^2 tests are also performed so as to test the statistical relationships between the churn indicator variable and explanatory features. Pearson's χ^2 test determines whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table. It is thus adapted to test whether two categorical variables are statistically independent. In this context, it appears interesting to use explanatory features related demographic, customer account and services subscribed as treatment variables when fitting the Kaplan-Meier estimator and implementing the tests.

Demographic data

Table 5.2 depicts the χ^2 tests' results performed on demographic variables. Given the *p-values* are ranked in ascending order and given the lower the *p value* the stronger link between two categorical variables, `Dependents` appears to be the most correlated feature with `Churn_Label`. When comparing this result with the corresponding survival plot in figure 5.1, it can be noted that customers with dependants have a longer lifetime in the *portfolio*. Conversely, `Gender` and `Churn_Label` are statistically independent as stated by the high test's *p value* ($\approx .49$).

Table 5.2: Independence χ^2 test between churn and demographic variables

	Statistic	Df	Critical Value	p-value
Dependents	431.65	1	3.84	7.1e-96
Senior_Citizen	158.44	1	3.84	2.5e-36
Partner	157.50	1	3.84	4e-36
Gender	0.48	1	3.84	4.9e-01

In section 3.4, nonparametric estimation has been introduced focusing on two major estimators: Kaplan-Meier for survival function estimation and Nelson-Aalen for estimating the cumulative hazard function. In this part, it is decided to draw survival curves related to customer lifetime in the portfolio depending on different types of treatment variables. In the figure below, four main results can be highlighted *ceteris paribus*:

- There seems to be no difference in terms of lifetime duration between men and women.
- Customers with a partner appear to stay longer in the TSP's *portfolio*.
- Being a senior citizen tends to shorten customer lifetime.
- As said before, customer with children or other dependents seem to be more loyal.

Data on services subscribed

When dealing with data on customers of a TSP, features related to services subscribed may be relevant to explain the estimated survival of these customers in the *portfolio*.

Table 5.3 presents results of χ^2 tests performed between `Churn_Label` and each services information variable. As in the previous table, *p-values* are ranked in ascending order. One can note that `Online_Security` and `Tech_Support` are

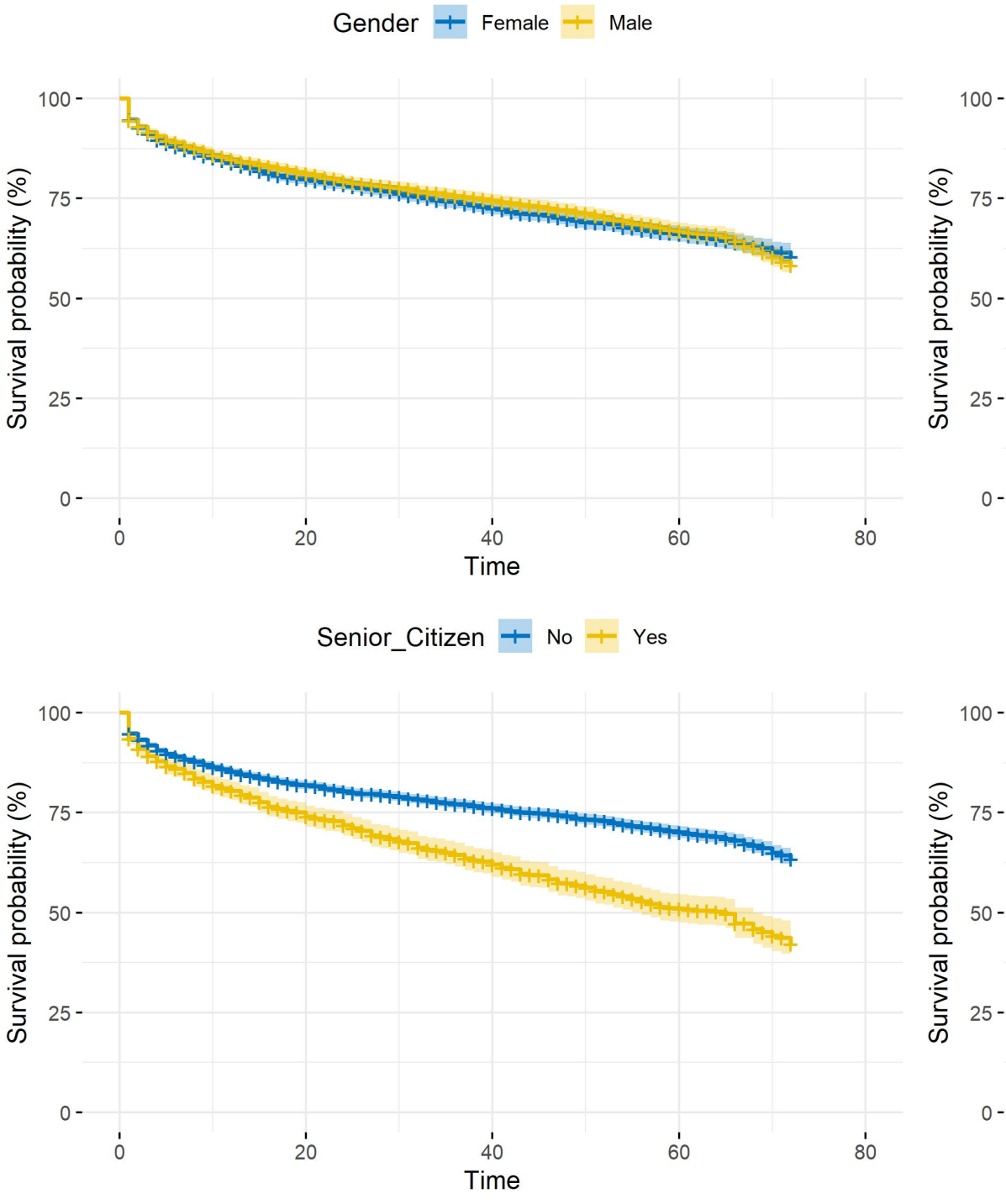

Figure 5.1: Kaplan-Meier survival function depending on demographic information

the most linked to the churn indicator variable. However, variable carrying information on phone services are less correlated to `Churn_Label`.

Table 5.3: Independence χ^2 test between churn and services information variables

	Statistic	Df	Critical Value	p-value
Internet_Service	728.70	2	5.99	5.8e-159
Online_Security	205.42	1	3.84	1.4e-46
Tech_Support	189.97	1	3.84	3.2e-43
Online_Backup	47.25	1	3.84	6.3e-12
Device_Protection	30.50	1	3.84	3.3e-08
Streaming_TV	27.84	1	3.84	1.3e-07
Streaming_Movies	25.76	1	3.84	3.9e-07
Phone_Service	0.87	1	3.84	3.5e-01
Multiple_Lines	0.87	1	3.84	3.5e-01

Figure 5.2 illustrates the χ^2 tests' results by representing the Kaplan-Meier estimated survivor function related to customer lifetime according to treatment variables on services subscribed. On the one hand, there seems to be no significant difference in terms of survival whether the customer uses phone service or not. The same remark can be pointed out based on whether the client has multiple lines as `Phone_Service` and `Multiple_Lines` might be quite correlated. In contrast, huge survival time difference can be noticed between customers with online security and those without, as well as between those having subscribed to technical support and those who have not. Finally, not using Internet service appears to have a positive influence on customer lifetime.

Customer account data

Variables on customer account such as the payment method used and the type of contract between the TSP and the client can be rich in information on customer lifetime. Indeed, table 5.4 shows that churn status strongly depends on the three variables, `Contract` being the most linked to `Churn_Label`.

Table 5.4: Independence χ^2 test between churn and customer account data

	Statistic	Df	Critical Value	p-value
Contract	1179.55	2	5.99	7.3e-257
Payment_Method	645.43	3	7.81	1.4e-139
Paperless_Billing	256.87	1	3.84	8.2e-58

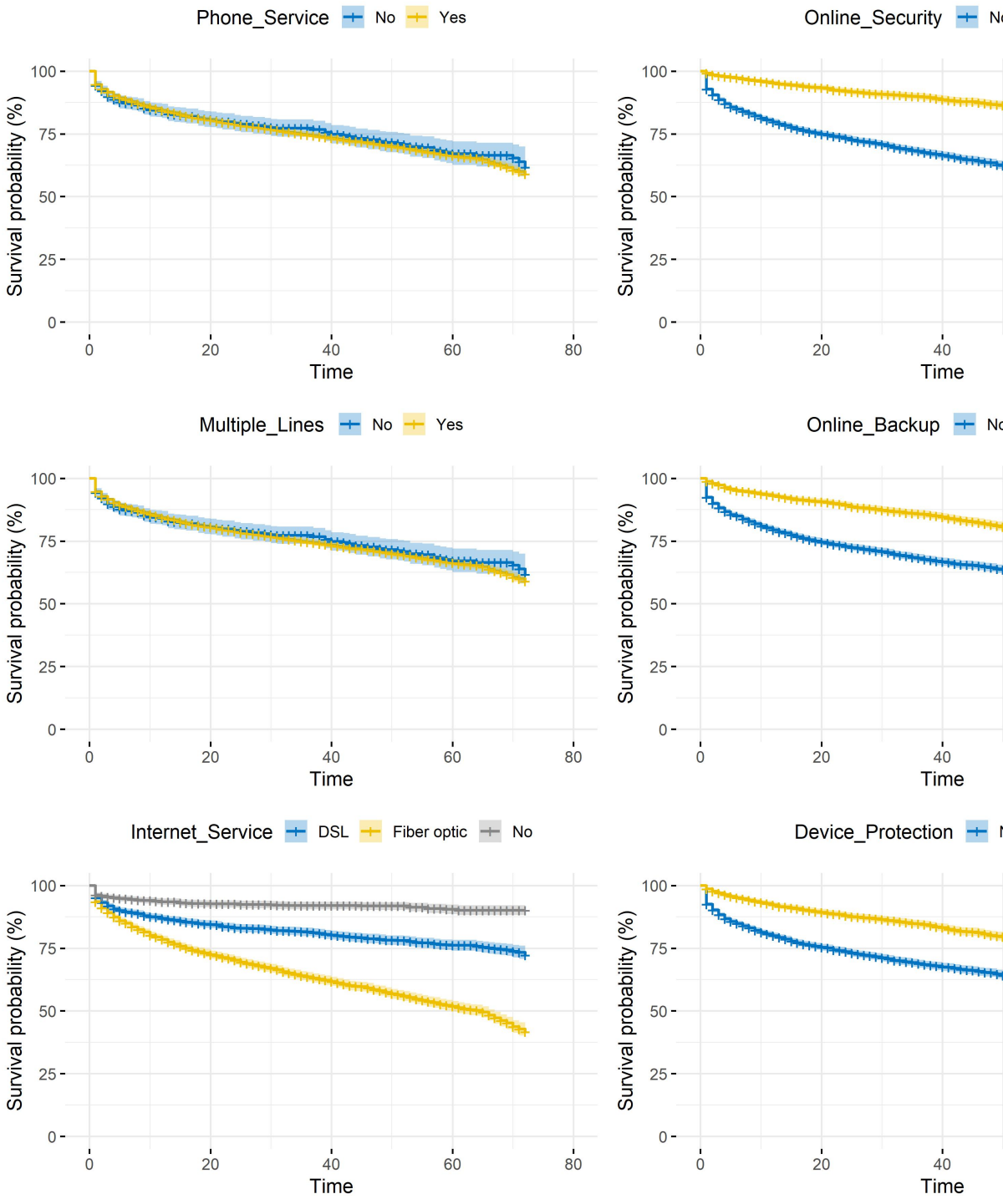

Figure 5.2: Kaplan-Meier survival function depending on services subscribed

Figure 5.3 enriches the χ^2 tests' results as it draws survival curves for each treatment variable's categories. When the firm/client contract is type month-to-month, the estimated survivor function decreases far more than for one-year or two-year contracts. In other words, the churn hazard is higher when the contract is renewed each month. This result makes sense as the customer may decide to leave the *portfolio* once the month has ended as they are not commited for one or two years. Furthermore, clients with paperless billing contracts are more prone to churn, just like those paying by electronic check. It can be deduce that the *attrition* risk is higher when the payment method is simplified.

5.3 Churn, duration and price

The final step in the data exploration consists in analysing the relationship between the `Monthly_Charges`, `Churn_Label` and `Tenure_Months` variables as they play an important role in the modelling strategy we adopt to estimate customer *value*.

Looking at figure 5.4, monthly charges seem to be higher for churners than for retained customers as the density is more right-oriented. High fees might be a driver of customer churn.

Besides, the low *p value* related to the Anova test between `CLTV` and `Churn_Label` indicates that customer lifetime value is statistically different between churner and retained clients.

Table 5.5: Anova test between monthly fees and churn status

	F statistic	Df1	Df2	p-value
Churn_Label	271.58	1	7030	6.8e-60

The following histograms are interesting to the extent that the distribution of `Tenure_Months` depends on the churn status. From figure 5.5, one can note an inflation of low and high values for retained customers. The distribution appears to be more homogeneous for retained clients than for churners. These lasts' tenure months distribution is decreasing and looks like a Poisson distribution with an inflation of low values.

Eventually, figure 5.6 depicts the average monthly charges per number of months in the portfolio. One can notice an increasing evolution between the average monthly fees and the number of months. In other words, it might be assumed that customers with longer lifetimes bring in more money to the firm.

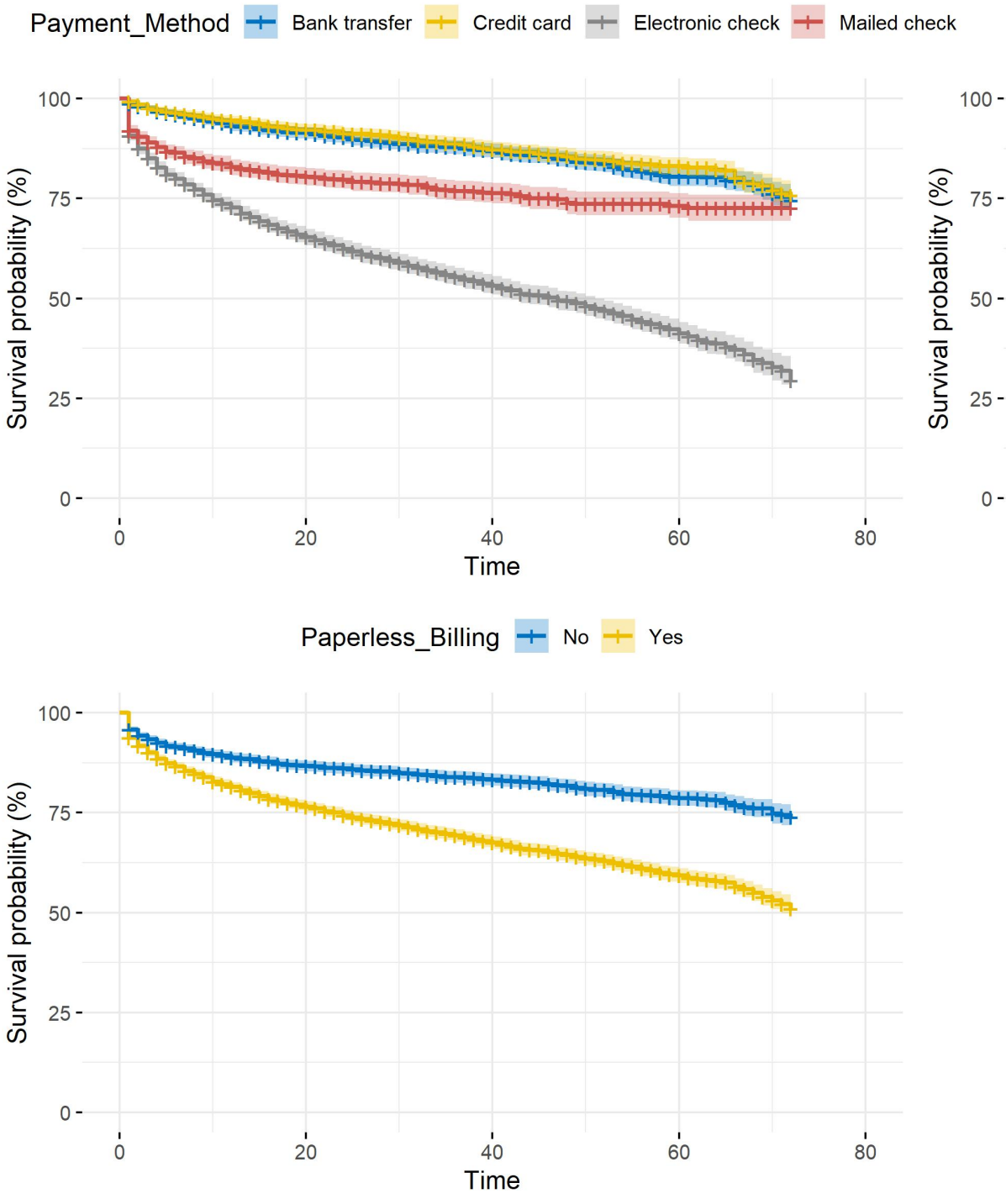

Figure 5.3: Kaplan-Meier survival function depending on customer account information

Figure 5.4: Monthly charges paid by customers depending on churn status

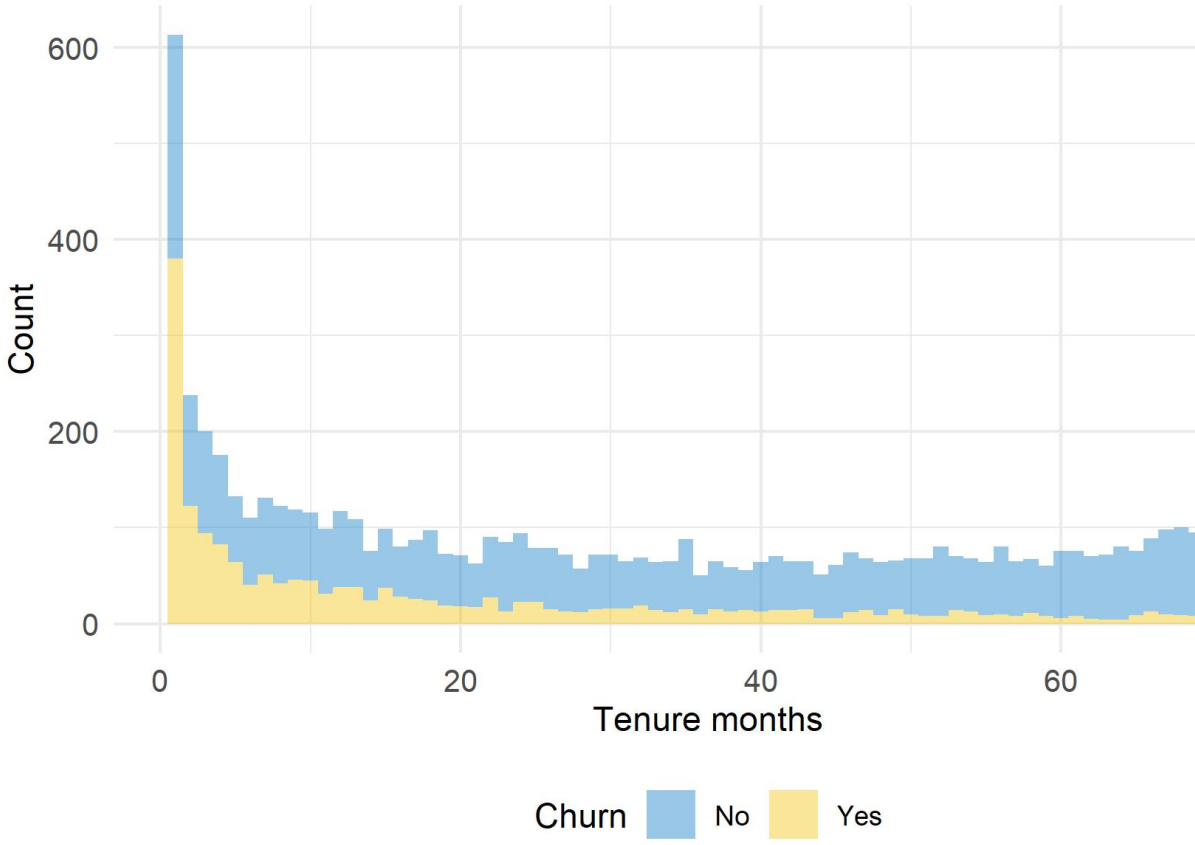

Figure 5.5: Tenure months depending on churn status

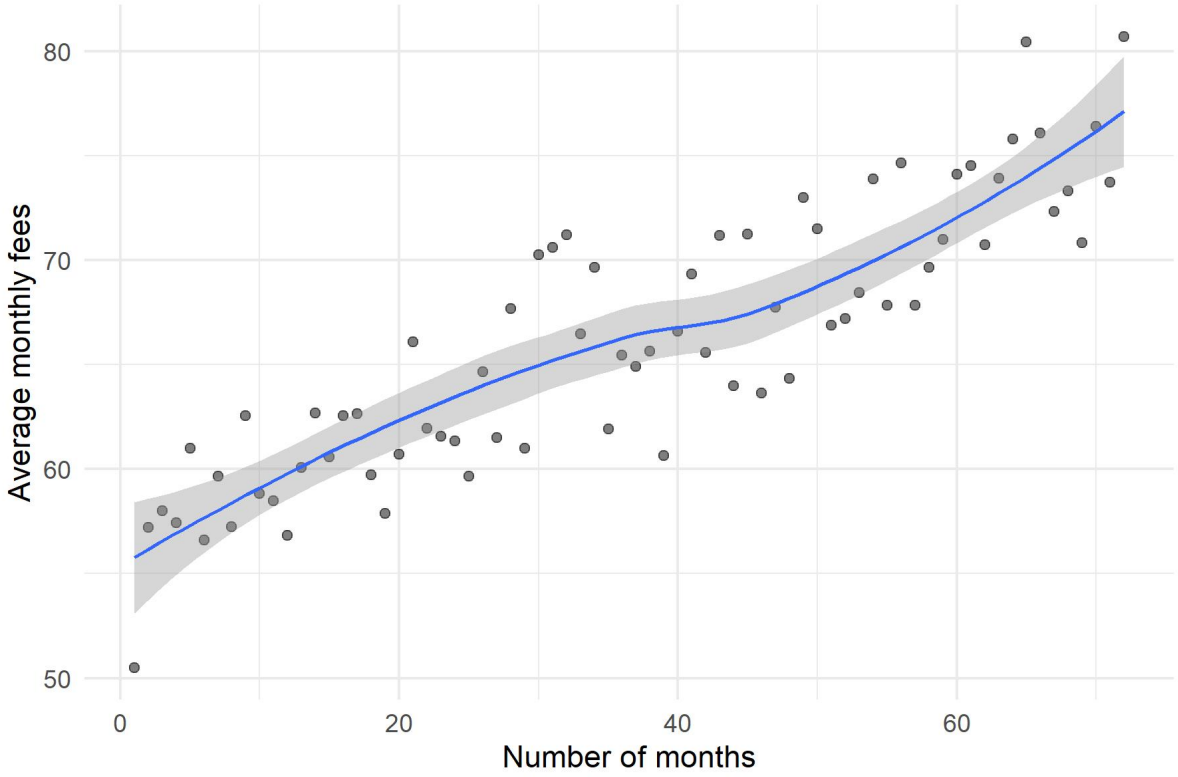

Figure 5.6: Average monthly fees depending on number of months in the portfolio

Chapter 6

Estimation techniques

This chapter explains the different methods used to model a portfolio of customers as well as their related risk of attrition and the overall value of the portfolio. In a first part, clustering techniques are implemented to identify segments of customers based on services and account variables. Then, survival models are fitted to estimate each customer's survival in the firm's portfolio. The selected model can also be used to assess the effects of each variable on the risk of churn. Finally, we answer the study's problematic by computing an estimated value of the portfolio. The latter is calculated using a corporate formula and takes customers' monthly fees and survival probabilities as inputs. The estimated portfolio value is not cost-adjusted there is no information on consumers' costs in the data set.

6.1 Feature selection

Before fitting any survival model or clustering algorithm to the data, the initial step consists in selecting variables that are discriminating in terms of churn hazard. Based on Kaplan-Meier analysis depicted in section 5.2, we have a general overview of features which influence the survival probability. In other words, our feature selection method relies on results obtained with descriptive statistics.

Table 6.1 shows the selected variables for 5 random observations extracted from the data set. It can be noted that these features are related to account or service information, apart from `Dependents` which indicates whether the client lives with any dependents (children, parents, etc) and `Senior_Citizen`. Furthermore, 9 out of the 10 selected variables are categorical which implies that the estimation results could be used to compare different groups of client. `Monthly_Charges` is the only quantitative variable used to fit clustering models and survival regressions.

Table 6.1: Explanatory variables used in survival models and cluster analysis

Senior_Citizen	Dependents	Phone_Service	Internet_Service	Online_Security	Online_Back
Yes	No	Yes	DSL	No	Yes
No	Yes	Yes	No	No	No
No	Yes	Yes	Fiber optic	Yes	Yes
No	No	Yes	No	No	No
No	No	Yes	DSL	Yes	No

6.2 Portfolio segmentation

Customer segmentation helps decision makers having a better understanding of their clients. It can then be used to enhance marketing strategies via personalization. In other words, segmentation can lead to target customers with offers and incentives personalized to their wants, needs and preferences. In order to make segmentation more accurate, it is more appropriate to use cluster analysis than predetermined thresholds or rules, even more when we have several variables at our disposal. In this context, this section focuses on applying clustering methods on features displayed in table 6.1, apart from `Monthly_Charges`.

6.2.1 Transforming qualitative variables into principal axes

The variables selected to perform cluster analysis being categorical, it is needed to transform them into continuous features. To that end, multiple correspondence analysis (MCA) is performed. MCA is a dimension reducing method which takes multiple categorical variables and seeks to identify associations between levels of those variables. MCA aims at highlighting features that separate classes of individuals, while determining links between variables and categories. To that end, MCA keeps the core information by the means of principal components which are projected axes (Scholler, 2021a).

Here, the main objective of applying MCA being to obtain continuous features, it is decided to keep as many axes as it takes to have at least 80% cumulated variance. In other words, we want the principal components to gather enough customer-related information. After having processed the `MCA` function from the R package `FactoMineR` (Lê et al., 2008), 10 principal components are required to keep more than 80% cumulated variance as depicted by figure 6.1.

Now the 10 continuous axes are identified, the next step consists in retrieving the customers' coordinates onto those axes to then perform cluster analysis.

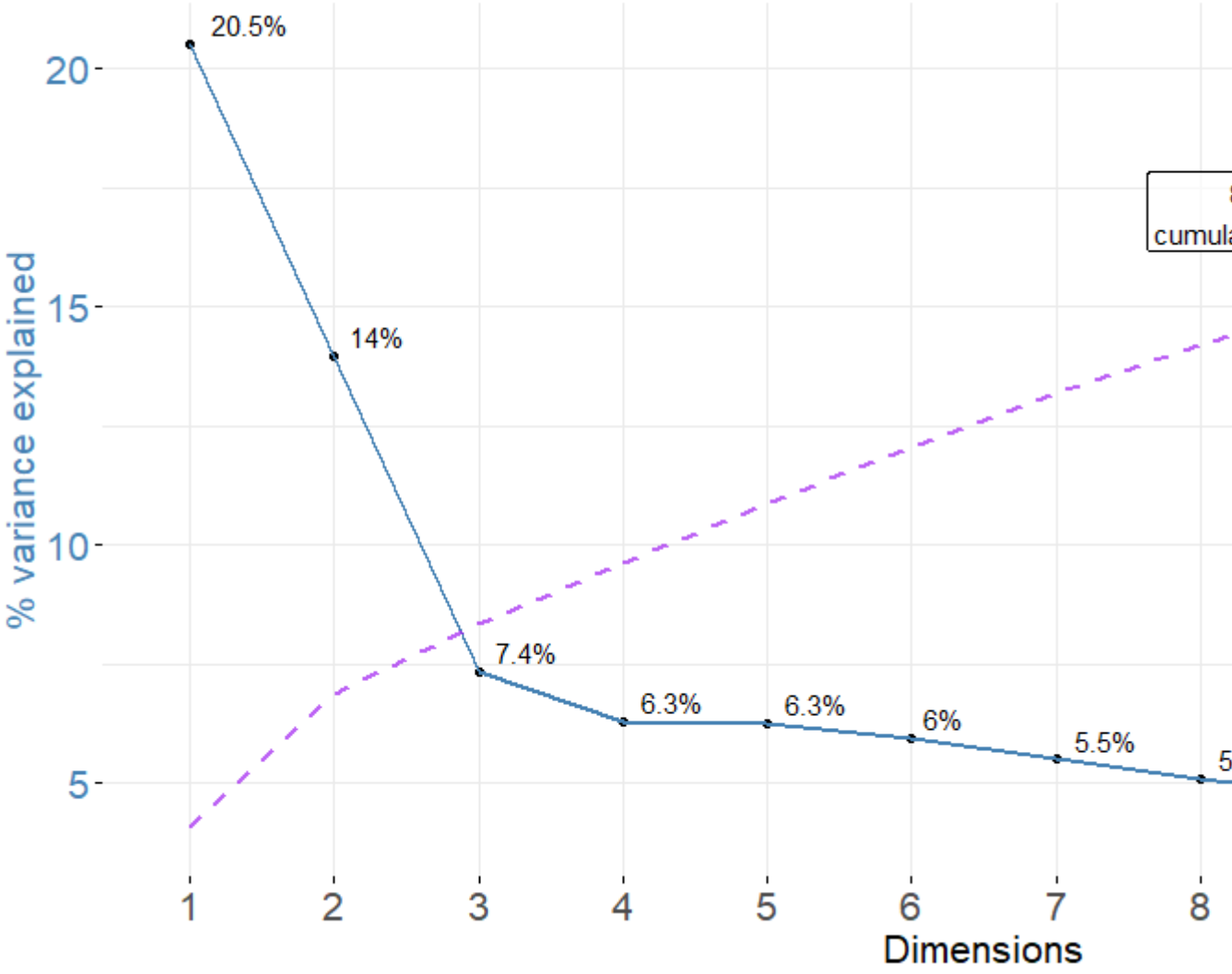

Figure 6.1: Variance explained and cumulated variance after MCA

Table 6.2: The 10 principal axes obtained by MCA

CustomerID	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10
4197	-0.15	-0.02	-0.30	-0.03	0.03	-0.29	-0.61	0.01	-0.04	0.10
3025	0.53	0.04	0.08	0.26	0.09	-0.54	0.55	0.40	-0.02	-0.19
6585	-0.04	-0.24	-0.15	0.40	0.07	-0.20	0.18	-0.25	0.02	0.07
3862	-0.44	-0.20	-0.46	-0.14	-0.01	0.01	-0.05	-0.04	-0.52	0.01
300	0.21	-0.37	0.11	0.00	0.00	-0.41	-0.54	-0.13	0.36	0.06

Note that a more in-depth visualisation of those 10 principal components can be found in figures 6.18 to 6.22 in the appendix. This charts depict the percent contribution of each variable's categories to the principal axes, which is helpful to have a better understanding of MCA results.

6.2.2 Hierarchical clustering on principal components

Multiple correspondence analysis has led us to convert the categorical variables related to account and services information into 10 numerical projected axes. The stake here is to use the customers' projections onto the MCA components in order to identify groups of individuals through clustering techniques. As a reminder, clusters are expected to discriminate between customers based on the services they use and the type of plan they are enrolled into. The method implemented in this part relies on hierarchical clustering on principal components (HCPC).

Optimal number of clusters

The key parameter to optimize when applying clustering methods is the number of clusters k. When using the `HCPC` function from `FactorMineR`, the `nb.clust` parameter is set to -1 so that the tree is automatically cut at the suggested level. More precisely, the function first builds a hierarchical tree. Then the sum of the between-cluster inertia is calculated for each partition. The suggested partition is the one with the higher relative gain in inertia. Intuitively, the underlying objective is to choose a number of clusters leading to k well distinguished groups. Here, the between-cluster inertia is the metric measuring the amount of variability between clusters.

Once the 3 groups identified by the clustering algorithm, one can determine customer repartition within those groups. Looking at table 6.3, cluster 2 represents the largest share of clients. This result is illustrated on figure 6.3.

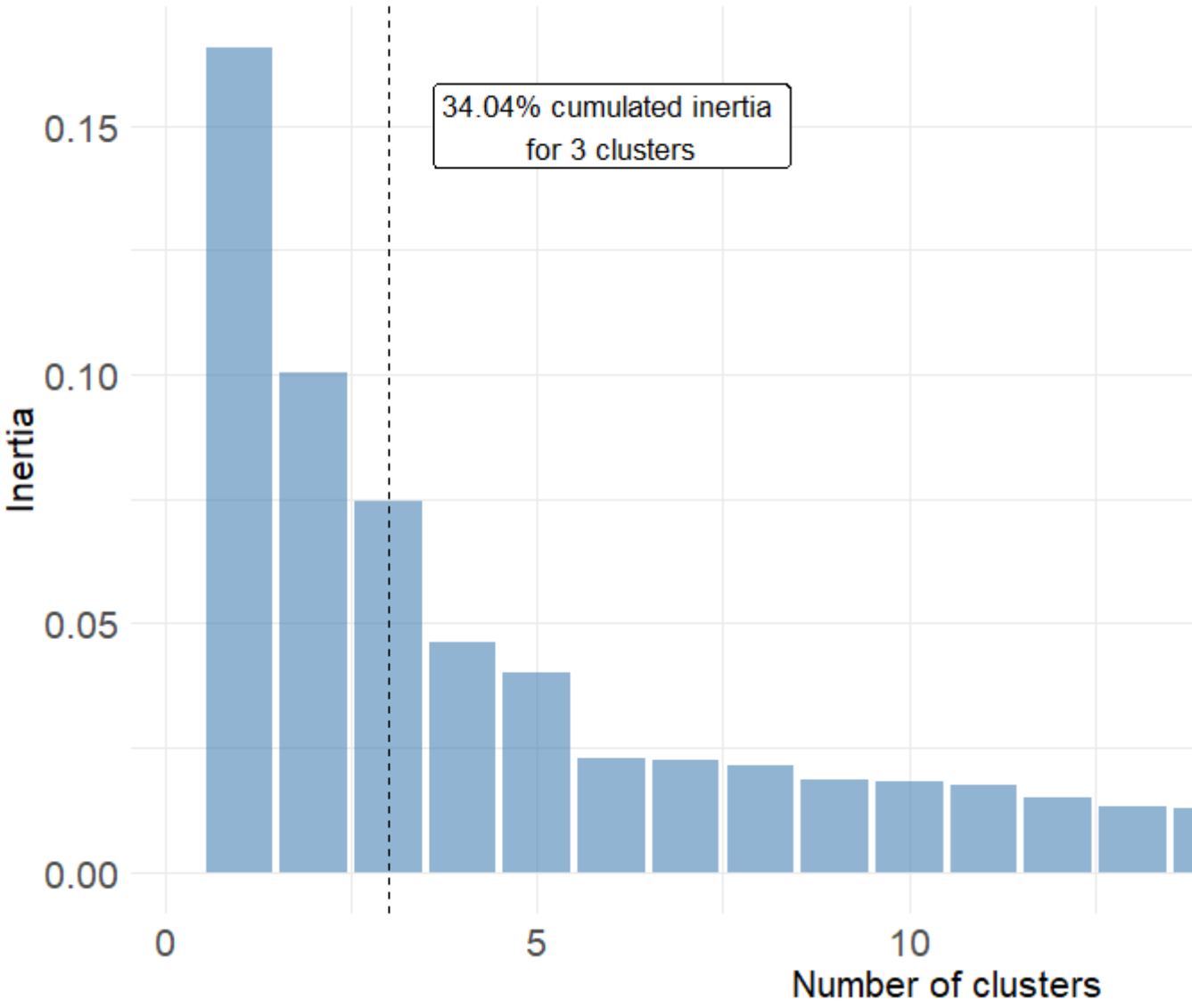

Figure 6.2: Relative gains in between-cluster inertia given the partition

Table 6.3: Customer repartition within the 3 clusters

	Cluster 1	Cluster 2	Cluster 3
%	26.55	41.38	32.07

Cluster visualisation

When performing hierarchical clustering on principal components, visualizing cluster repartition onto MCA axes is relevant to have a first idea on how each cluster if different from the others. Figure 6.3 indicates that the three clusters are well separated in the $(F_1; F_2)$ plan. Cluster 1 is more concentrated than the others and take lower values onto dimension 1. The second group spreads over axis 1 and takes lower values on axis 2. Finally, cluster 3's position shows that its customers are characterized by categories which take positive values on both dimensions 1 and 2.

In the appendix, figure 6.23 aims at visualizing clusters onto the other MCA principal components. Note that none of theses axes manages to separate the three clusters. This may be due to the low amount of variance each of these axes carries.

Cluster description

Once the 3 customer segments are identified on the MCA dimensions, it seems to be interesting to describe them according to the original features. To that end, the repartition of qualitative variables' categories is depicted on the following figure. Comparing figures 6.3 and 6.4, one can derive the following table in order to have a more precise idea on each segment.

Table 6.4: The most representative categories for each cluster

Cluster	Representative Categories
1	Mailed_check, Internet_Service_No
2	Month-to-month, Electronic check, Tech_Support_No
3	Two year, Credit card, Tech_Support_Yes

The first segment may be made up of customers having subscribed to minimum plans whith no internet service whereas segment 3 tends to represent clients with a large variety of services and long-term contracts. On the opposite, cluster 2 clients are enrolled in short-term plans with internet service but not technical options.

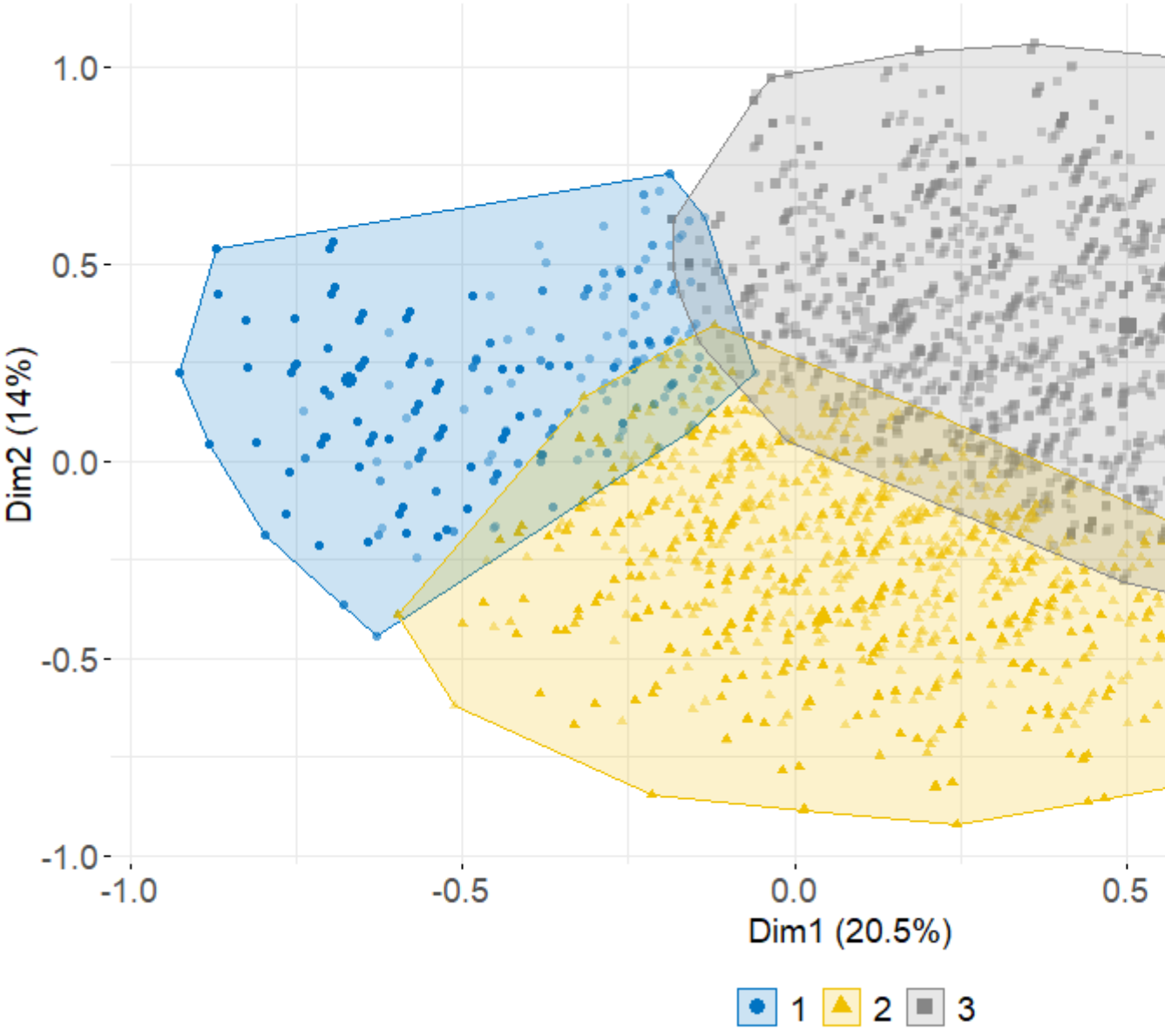

Figure 6.3: Cluster visualisation onto the 2 first MCA axes

Figure 6.4: MCA - Categories plot onto the two first axes

Figure 6.5 depicts the differences in monthly charges paid by each customer segment. It can be observed that most of cluster 1 clients are enrolled in cheap subscriptions. On the contrary, the charges paid by the second and third clusters are higher and less homogeneous. Note that the median price is the same among these two groups.

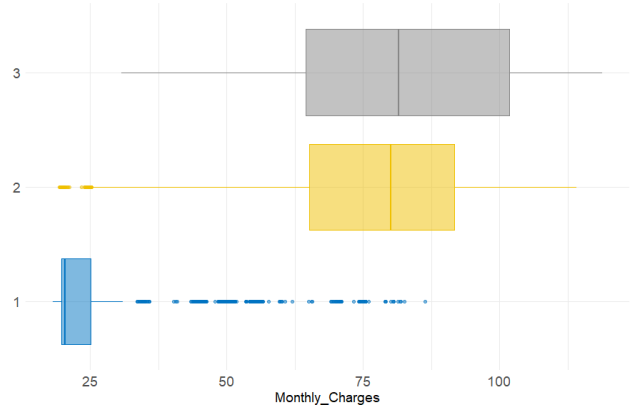

Figure 6.5: Monthly charges repartition across clusters

Since one of the study's purpose is to model customer churn, it is important to compare the churn rate across the 3 groups. Here, the striking point is that cluster 2 clients account for a large proportion of churners. Indeed, more than 75% of the clients having left the portfolio come from the second group. It can also be noted that cluster 1 and 3 are composed of more loyal customers as more than 70% of clients who did not churn are from those 2 groups.

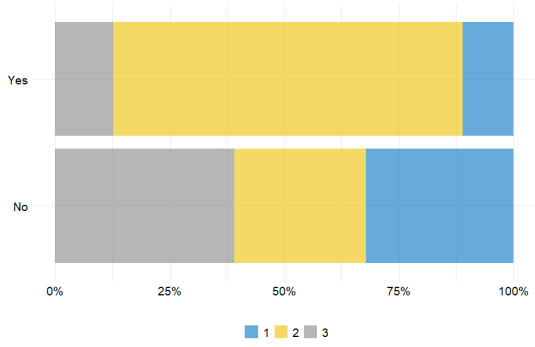

Figure 6.6: Churn repartition per cluster

Concluding remarks on portfolio segmentation

With the aim of divide the portfolio into multiple groups based on categorical features, hierarchical clustering on principal components has led to identify three customer segments which can be broadly described as shown in the following table.

Table 6.5: Portfolio segmentation summary

Cluster	Loyalty	Value	Segment name
1	High	Low	*Silver*
2	**Low**	**High**	***Gold***
3	High	High	*Platinum*

Gold customers are valuable to the firm but are not loyal, making them the target segment. The objective would be to encourage them to stay in the portfolio longer.

6.3 Churn analysis

Estimating the risk of attrition related to each customer is an essential step to model the firm's portfolio. In this context, survival models can be implemented with a view of deriving a predicted churn risk and survival function for each client. One the one hand, these predictions can be used to identify loyal consumers and make appropriate decisions. For instance, it might be relevant to offer benefits to a high-value client with a high estimated churn risk. On the other hand, a customer's survival probability at time t represents the chance that this very customer be active in the portfolio at time t. This measure is helpful to compute the estimated value of the portfolio in the last section.

Before presenting the estimation results, it seems important to recall that `Tenure_Months` and `Churn_Value` can be seen as a pair of time and event variables used as target in survival models.

Table 6.6: Time and event variables for survival models

CustomerID	Tenure_Months	Churn_Value
1671	30	1
6705	43	0
3702	12	0
1519	1	1
4882	50	0

6.3.1 The Cox model

When it comes to choose an estimation method on survival data, the Cox PH model appears to be an interesting first choice. As explained in chapter 3, this semi-parametric model makes no assumption regarding the nature of the baseline hazard function $\lambda_0(t)$. The parametric part only relies in the modelling of the effect of some covariates on the hazard function $\lambda(t)$ (see section 3.6.2 for more details).

Fitting the model on the selected features

Using the `coxph` function from the `survival` R library (Terry M. Therneau and Patricia M. Grambsch, 2000), we are able to train a Cox model on the feature vector identified in section 6.1. Once the model fitted, it seems relevant to evaluate its performance on the train data set. Table 6.7 compares the model's log-likelihood to the constrained model's. Given the very low p-value, it can be assumed that the Cox model better fits the data than a model with only the intercept.

Table 6.7: Log-likelihood ratio test

Model	Constrained	pvalue
-9228.77	-10448.08	0

Concordance index c is another metric to assess the performance of models which produces risk scores and is defined as the probability to well predict the order of event occurring time for any pair of instances. For the Cox model, the C-index obtained on the training set is $c \approx 0.865 \pm 0.004$, which is more than satisfying.

Marginal effects

In the Cox model, the relative hazard between two observations is assumed to be constant over time. As a consequence, the relative hazard becomes $\exp \hat{\beta}$ for both dummy and continuous variables. For instance regarding figure 6.7, the relative hazard ratio between customers with a two-year contract and those with a month-to-month contract is 0.046, meaning that the latter group is 22 times more prone to churn than the former. Also, month-to-month clients are about 5 times more likely to churn than customers enrolled in one-year plan. When analysing the relative hazard ratios related to the payment method, it comes that clients who pay by electronic or mailed check are two times riskier to churn than those who pay by bank transfer. Furthermore, being enrolled

in a plan with additional services like `Online_Security`, `Online_Backup` or `Tech_Support` tends to decrease the estimated churn risk. As for the effect of the `Internet_Service` covariate on the risk of attrition, it seems to be mitigated since clients who have a fiber optic internet connection are more than twice as likely to churn as those using a DSL internet connection.

Estimation results

Semi-parametric models aims at estimating the instantaneous hazard function given a baseline hazard and a set of covariates. The model outputs a risk prediction for each individual with a confidence interval. Then, the survival and cumulative hazard functions can be retrieved as explained in section 3.3. When going deeper into the functions depicted in figure 6.8, it can be noticed some inconsistencies between the estimated churn hazard cumulative hazard functions. Given the cumulative churn hazard increases faster when the number of months is high and given the instantaneous hazard is supposed to be the cumulative hazard function's slope, the estimated churn hazard's shape should be convex. Thus, the Cox model does not manage to properly estimate the risk of churn.

Although the fitted model is not flawless, it can be interested to study the estimation differences between the 3 customer segments identified in the previous section. Looking at figure 6.9, one can conclude that *Gold* (cluster 2) clients are more prone to churn than others. The aggregated risk of attrition is indeed higher to such an extent that the cumulative churn hazard is exploding when the number of months is greater than 60. This being said, an efficient portfolio management would be to find strategies to reduce *Gold* customers' churn and increase their duration in the portfolio.

6.3.2 Other survival models

As said in the previous part, the Cox model does not fit the data perfectly as it does not capture the churn hazard's actual shape. In this context, it seems relevant to fit other survival models and test how they perform in predicting the risk of attrition. It is firstly decided to train parametric survival models which consist in assigning a probability distribution to the hazard function (see section 3.5 for more details). To that end, one can use `flexsurv` (Jackson, 2016) which is and R package for parametric survival modelling. After having fitted the exponential, Weibull, gamma, log-logistic and log-normal models, no improvement can be observed with respect to the Cox model. Then, a machine learning approach can be adopted to model duration data using the random survival forest algorithm as explained in section 3.7. The `rfsrc` function from `randomForestSRC` (Ishwaran et al., 2008) is used to train the model after having determined the optimal node size and number of variables to try at each split with the `tune` function. Once the model trained, its performance is compared to

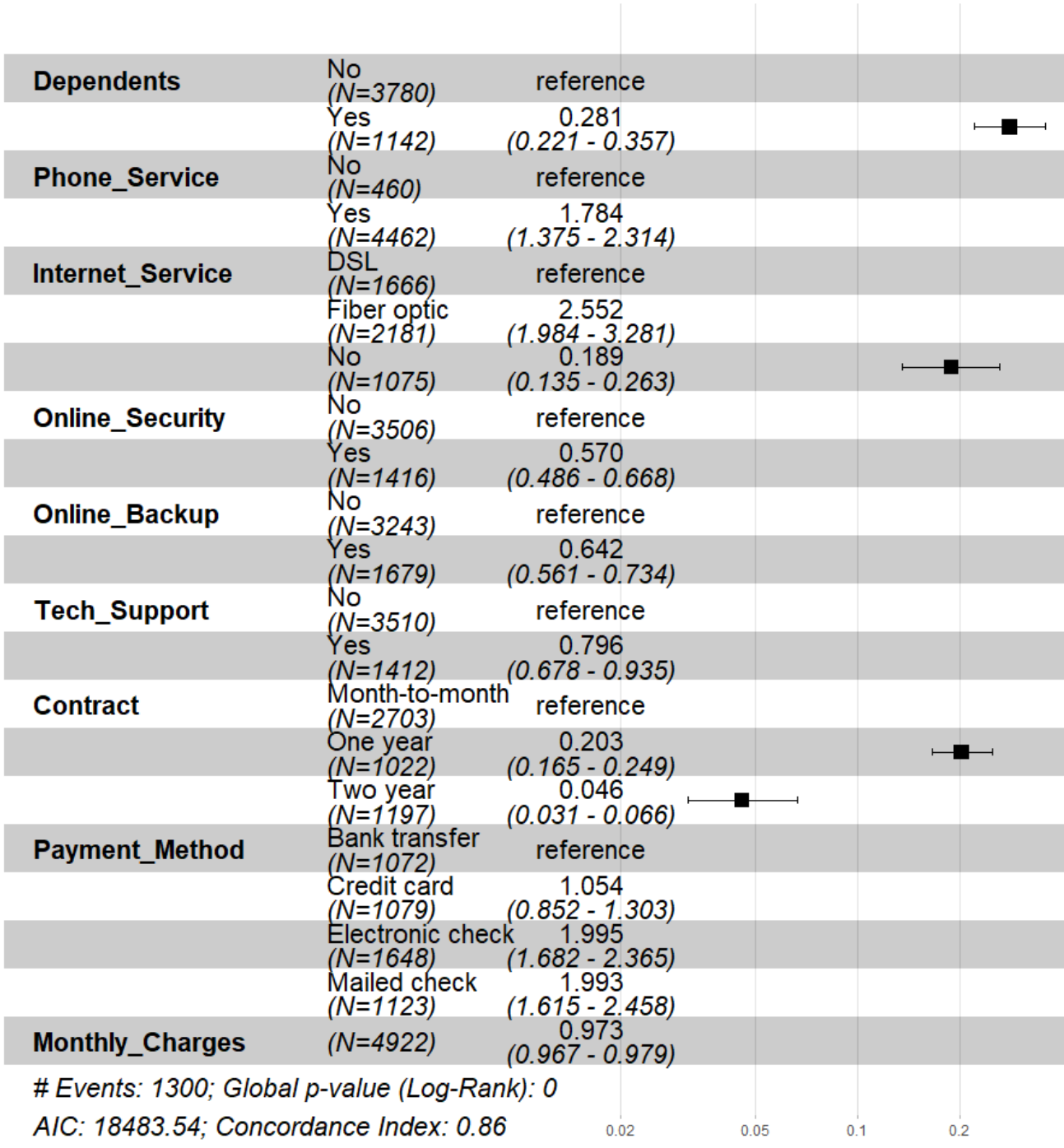

Figure 6.7: Marginal effects obtained with the Cox PH model

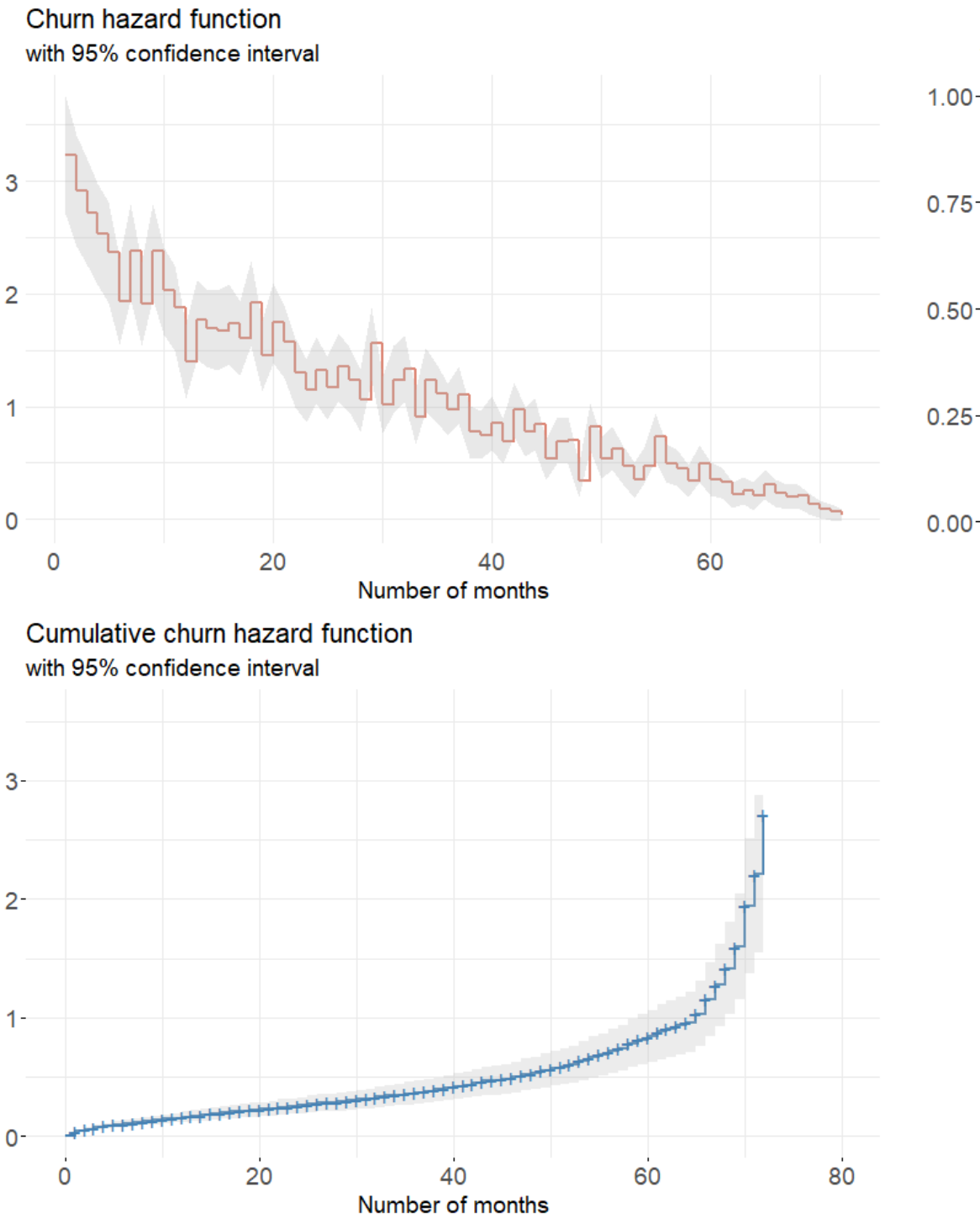

Figure 6.8: Aggregated churn hazard, survival and cumulative hazard functions estimated by Cox model

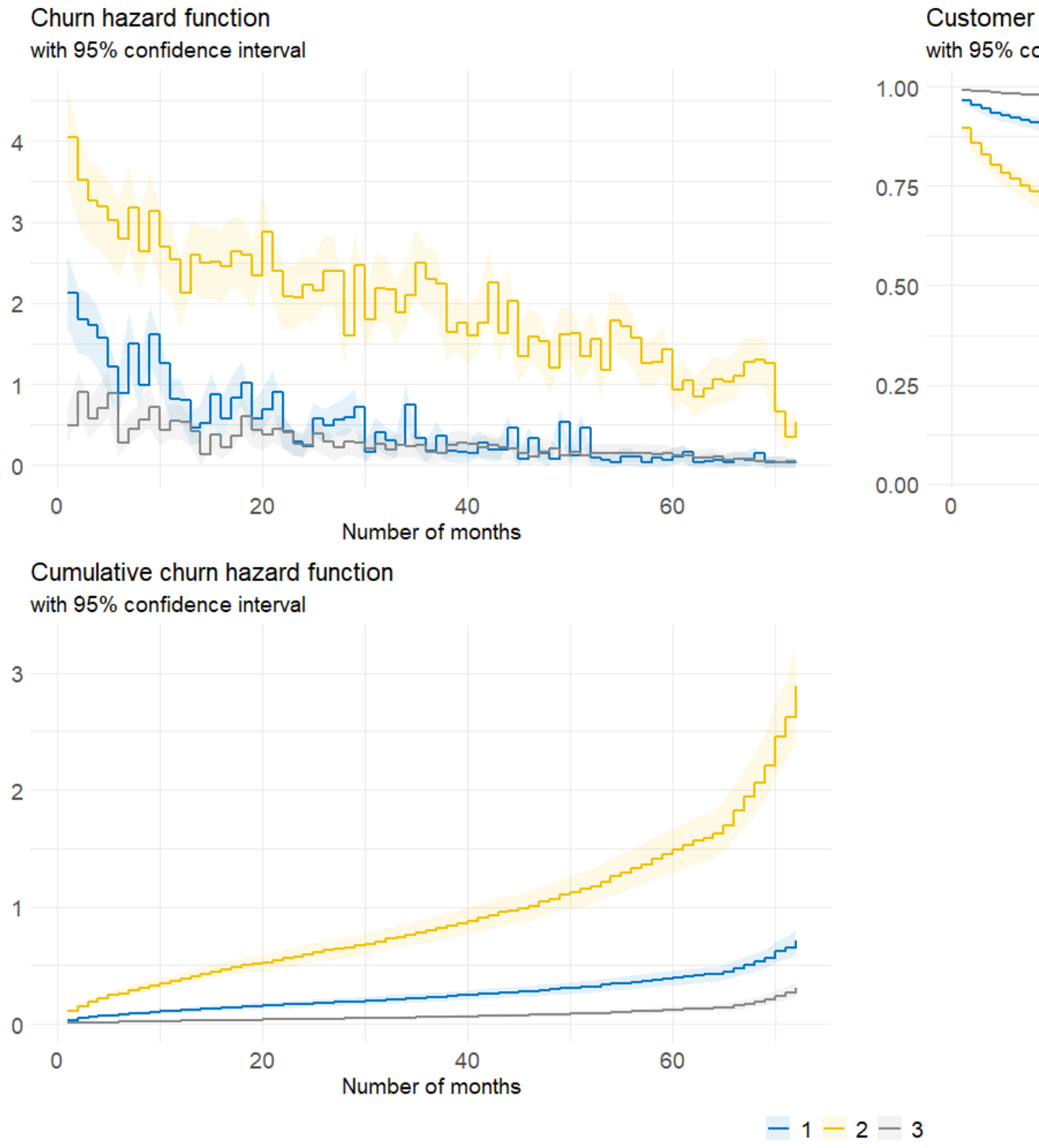

Figure 6.9: Aggregated churn hazard, survival and cumulative hazard functions for each cluster

the Cox model's as shown by table 6.3.2. One cannot but admit that the random survival forest perform poorly in terms of concordance index with respect to the Cox model. The latter manages to output a better risk scoring than the ML algorithm. It is thus decided to select the Cox PH model for the rest of the study.

\begin{table}[H]

\caption{Concordance index (%) obtained with Cox model and Random Survival Forest (RSF)}

	Train	Test
Cox	86.769	86.253
RSF	49.273	49.229

\end{table}

6.4 Portfolio value estimation

In the two previous sections, the portfolio has been partitioned into 3 customer segments and customer lifetime has been estimated by the means of the Cox model. This section's purpose consists in deriving a method to calculate the overall value of the portfolio through the computation of customer lifetime raw value.

6.4.1 The model

As said in chapter 5, it is decided not to use the `CLTV` variable to estimate customer lifetime value since no information has been provided on the computation method. Instead, another method is adopted to predict each client's lifetime value. For this purpose, two inputs are needed namely each customer's survival function as well as the monthly fees represented by the `Monthly_Charges` feature. Based on the CLV formulation proposed by Gupta and Lehmann (Gupta and Lehmann, 2003), we derive a model aiming at calculating customer lifetime raw value (CLRV) as depicted in equation (6.1).

$$\text{CLRV}_i = \sum_{t=0}^{T} \frac{p_i \; r_{i,t}}{(1+a)^t} \tag{6.1}$$

with,

- p_i the monthly fee paid by customer i,

- $r_{i,t}$ the probability that customer i be in the portfolio at time t,
- a the discount factor,
- T the time horizon.

It may be interesting to note that p_i is time-invariant and corresponds to `Monthly_Charges`. $r_{i,t}$ is computed using the survival function estimated with the Cox model. a is fixed at 8% and T equals 72 months which is the longest lifetime in the data.

Summing over the N CLRVs, one can derive customer raw equity which is the portfolio raw value (V) and may be a good proxy for the firm's overall revenues.

$$\mathrm{V} = \sum_{i=1}^{N} \mathrm{CLRV}_i = \sum_{i=1}^{N} \sum_{t=0}^{T} \frac{p_i \; r_{i,t}}{(1+a)^t} \tag{6.2}$$

6.4.2 Customer Lifetime Raw Value

In the literature, it is common place to measure customer lifetime value that is the overall profit brought by a client over her entire lifetime in the portfolio. In our study, the data set used does not provide any information on the costs related to each customer. Consequently, it is decided to evaluate customer value by computing the overall revenues the firm might earn during the relationship with their clients. We name this metric customer lifetime raw value (CLRV) and it is defined in equation (6.1).

The process implemented to compute client i's CLRV works as follows:

1. Estimate client i's survival function over 72 months using the Cox model fitted in section 6.3.1.
2. For each month, multiply the value of the survival function by the monthly fee p_i paid by client i. Then, divide the product by $(1+a)^t$ where t is the month's index and a the discount factor.
3. Take the sum of the T ratios.

Since the data is made up of more than 7,000 customers and given the CLRV calculation is time consuming, parallelization has been used to calculate every CLRV. Once all the CLRVs are computed, one may take the sum in order to retrieve the portfolio's global value as well as a 95% confidence interval as shown by table 6.8.

Table 6.8: Portfolio estimated value with 95% confidence interval

V lower	V	V upper
17,604,144	18,270,000	19,031,648

Going deeper into the statistical analysis of customer lifetime raw value, it can be noticed from table 6.9 and figure 6.10 that the clients are quite heterogeneous. The distribution of CLRV is left-skewed with a median lower than the mean. In addition, the CLRV ranges from 324 to 6815 indicating that each client does not have the same value to the firm.

Table 6.9: CLRV Statistical summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
323.99	1125.9	2104.66	2598.12	3850.41	6815.42

Eventually, it seems interesting to visualize the contribution of each month to customer lifetime raw value. Based on figure 6.11, the montlhy contribution decreases as the number of months increases which leads to a concave cumulative value. This result may be explained by the increase in $(1+a)^t$ and the decrease in the survival probability $r_{i,t}$ as t increases (see equation (6.1)). In other words, customers bring in more value to the firm at the beginning of the relationship.

6.4.3 Cluster contribution to the portfolio value

In section 6.2, 3 clusters of customers have been identified using hierarchical clustering. This being said, it is undoubtedly necessary to calculate the value of each group using the method presented above. The following table depicts each cluster's CLRV with 95% confidence interval as well as the % proportion in terms of number of clients and % contribution in terms of value. One can notice that the *Platinum* cluster accounts for more than half the portfolio total value even if it represents barely a third of the number of clients. Besides, *Silver* customers are the less valuable to the firm as their contribution barely amounts to a 12%. Recall that this group is characterized by a minimum subscription with no internet connection or additional services. Finally, the *Gold* segment's contribution is $\approx$ 40% lower than cluster 3's. It may be explained by cluster 2 client's higher propensity to churn.

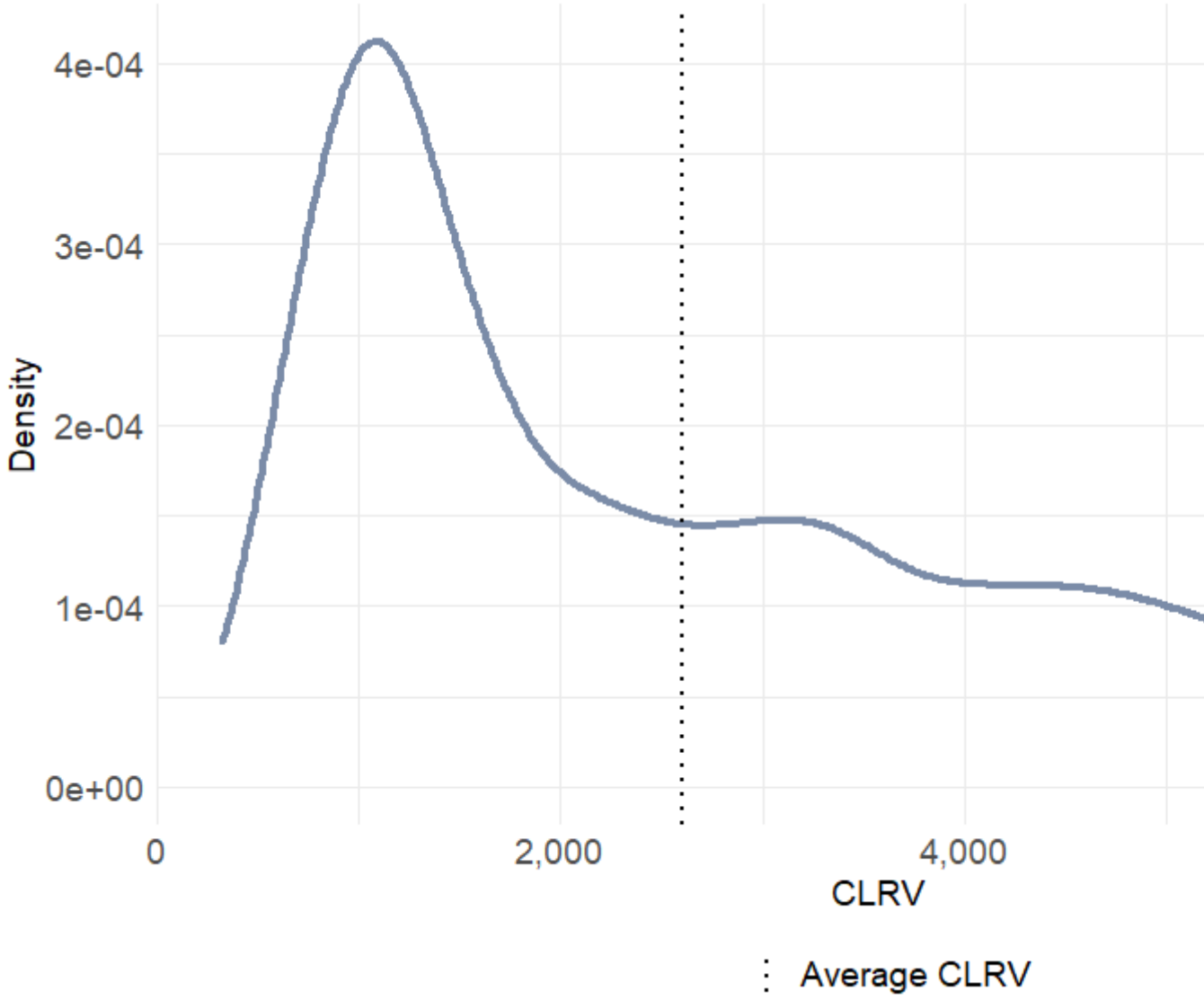

Figure 6.10: Distribution of Customer Lifetime Raw Value

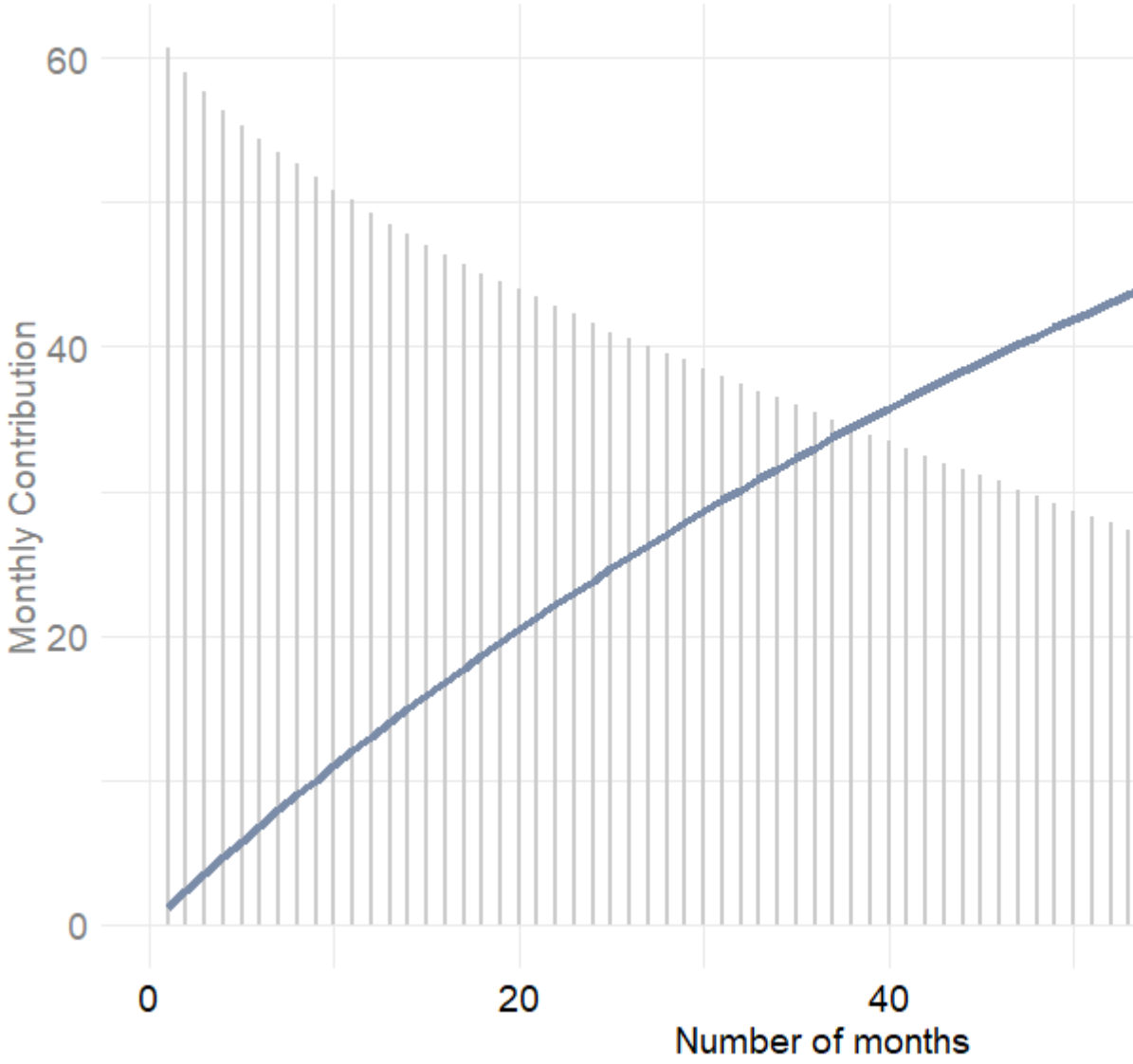

Figure 6.11: Monthly contribution and cumulative value depending on the number of months in the portfolio

Table 6.10: Total CLRV per cluster

Cluster	Proportion (%)	V lower	V	V upper	Contribution (%)
Silver (1)	26.6	2,035,928	2,117,932	2,213,746	11.59
Gold (2)	41.4	5,760,579	6,211,027	6,734,243	34.00
Platinum (3)	32.1	9,807,637	9,941,041	10,083,660	54.41

Once the value differences across clusters identified, let us dig into a more detailed analysis of customer value based on the group they belong to. On figure 6.12, the 3 CLRV distributions indicate large disparities between each client segment. While cluster 1 customers are defined an inflation of values located before 2,000, the two other groups are more homogeneous. The CLRV of cluster 2 (respectively 3) clients seems evenly distributed between 0 (respectively 1,500) and 6,000 (respectively 7,000).

Table 6.11: CLRV statistical summary for each customer segment

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Silver (1)	333.82	837.50	1094.51	1134.40	1264.28	4679.80
Gold (2)	323.99	1194.59	1862.16	2134.37	2886.02	6009.78
Platinum (3)	1268.74	3393.93	4489.62	4408.44	5454.71	6815.42

When looking at table 6.11, large variations appear between each cluster's average CLRV. With a view of testing whether these differences are significant, the ANOVA test is performed between the `CLRV` and `cluster` variables. The very low p-value depicted in the table below indicates significant different means across the 3 groups. Hence, it can be affirmed that the contribution of each client segment to the portfolio value is statistically different.

Table 6.12: Results of the ANOVA test between CLRV and cluster

	Df	Sum of squares	MSE	F value	P value
Cluster	2	12,016,055,579	6,008,027,789	5,205.39	0
Residuals	7029	8,112,831,034	1,154,194		

With the objective of determining how each group's representative customer differ one from another in terms of CLRV, it seems a good idea to visualize the evolution of their respective cumulative value as the number of months in the portfolio increases. Figure 6.13 shows that the lower value to the firm, the flatter curve. To put it another way, *Silver* clients' monthly contribution to

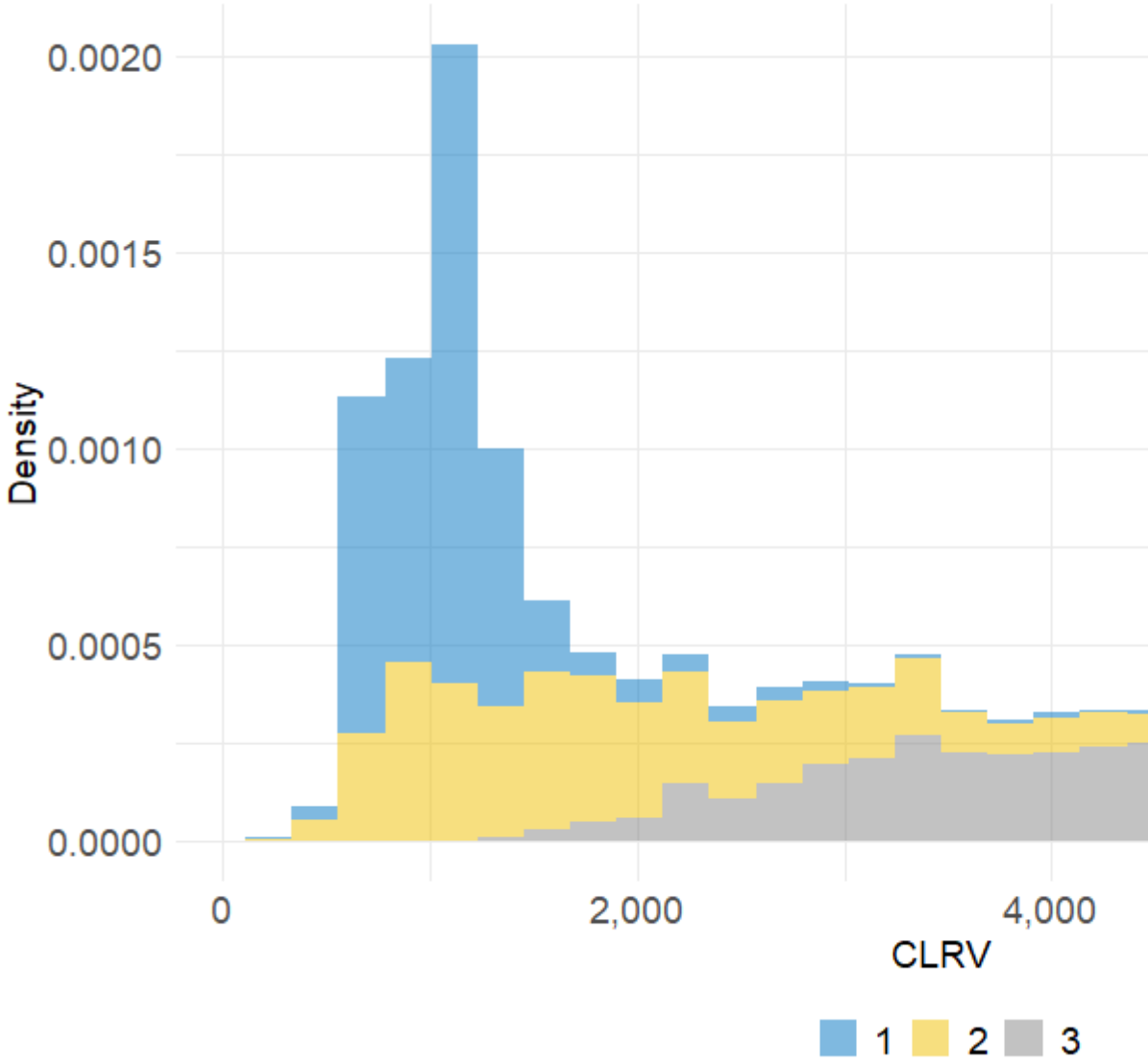

Figure 6.12: Distribution of Customer Lifetime Raw Value per cluster

CLRV decreases rapidly and becomes constantly low after 30 months. On the contrary, *Platinum* customers bring in a larger monthly value and the latter decreases to a lesser extent. As for *Gold* customers, their cumulative value becomes nearly constant from month 40 which may be due to their lower loyalty to the firm.

6.4.4 Simulations

The final step in our study aims at simulating different multiple scenarios to determine how customer lifetime raw value and the portfolio value might evolve *ceteris paribus*. This section is not exhaustive in inasmuch as only a few parameters have been selected to perform the simulations.

Influence of the discount factor on CLRV

In the CLRV formula (equation (6.1)), the discount factor a plays a central role since it monitors the importance of future values in the customer value. In this context, it is decided to analyse the evolution of both CLRV and portfolio value after a change in a. The simulation results reproduced in table 6.13 lead us to suppose that the portfolio overall value decreases as the discount factor increases. V is indeed reduced by almost 16% when a goes from 1% to 8%.

Table 6.13: Portfolio value depending on discount rate

	V Lower	V	V upper
1%	20,894,006	21,721,297	22,674,662
2%	20,358,160	21,158,861	22,080,569
4%	19,357,495	20,108,832	20,971,852
8%	17,604,144	18,270,000	19,031,648

The reduction in V after an increase in the discount rate is actually the direct consequence of a decrease in every customer lifetime raw value. The monthly contribution to CLRV being negatively influenced by a as the number of months t increases, this phenomenon is all the more pronounced when a is large. On figure 6.14, one can observe that the higherthe discount factor, the more concave the shape of the cumulative value. In other words the higher the discount factor, the lower gain in CLRV when the number of months is high.

Influence of month-to-month contracts on CLRV

While it is obviously doable to simulate new customer lifetime raw values by playing with the model's parameters, one can also change the values of

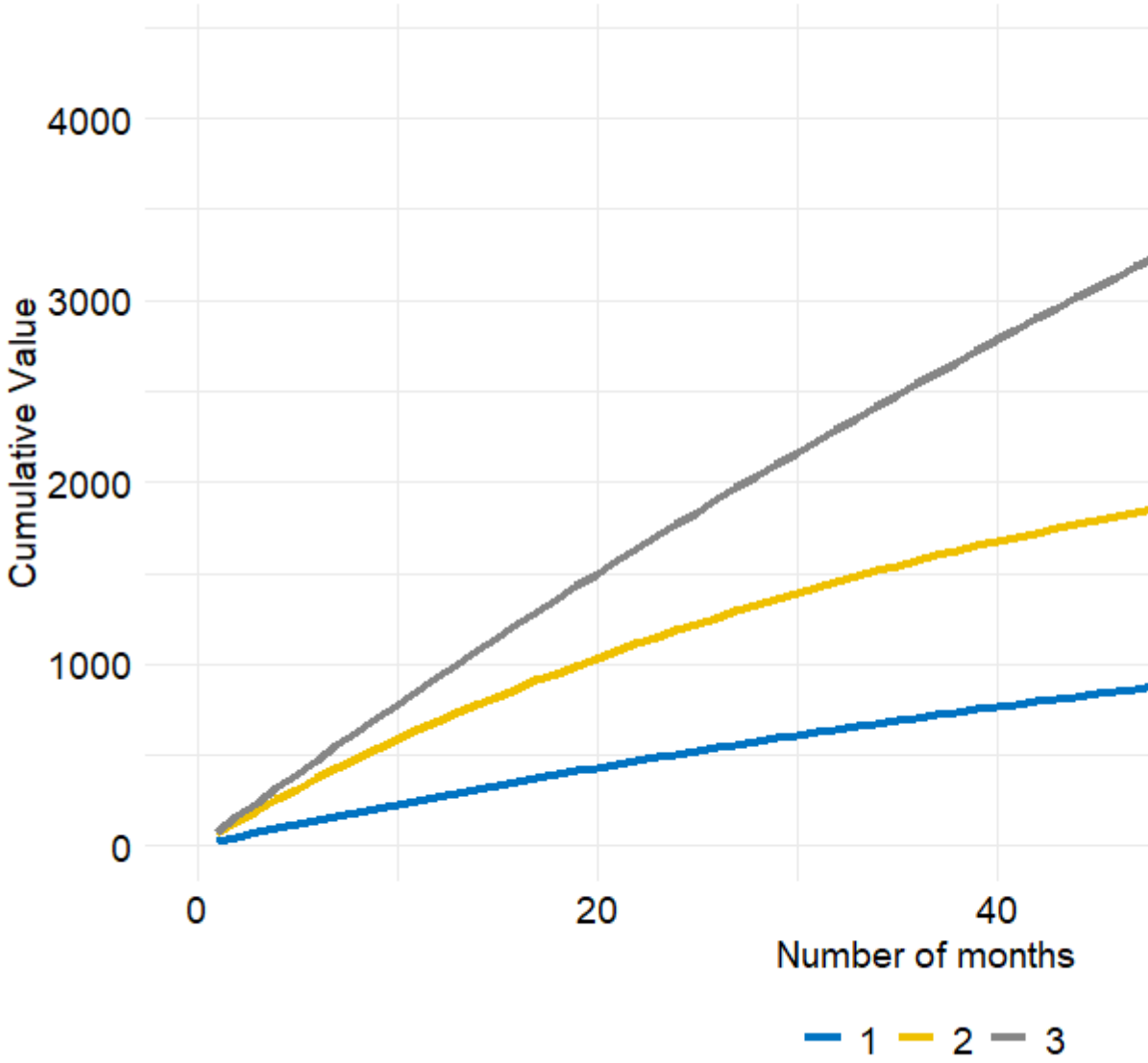

Figure 6.13: Customer cumulative value through time per cluster

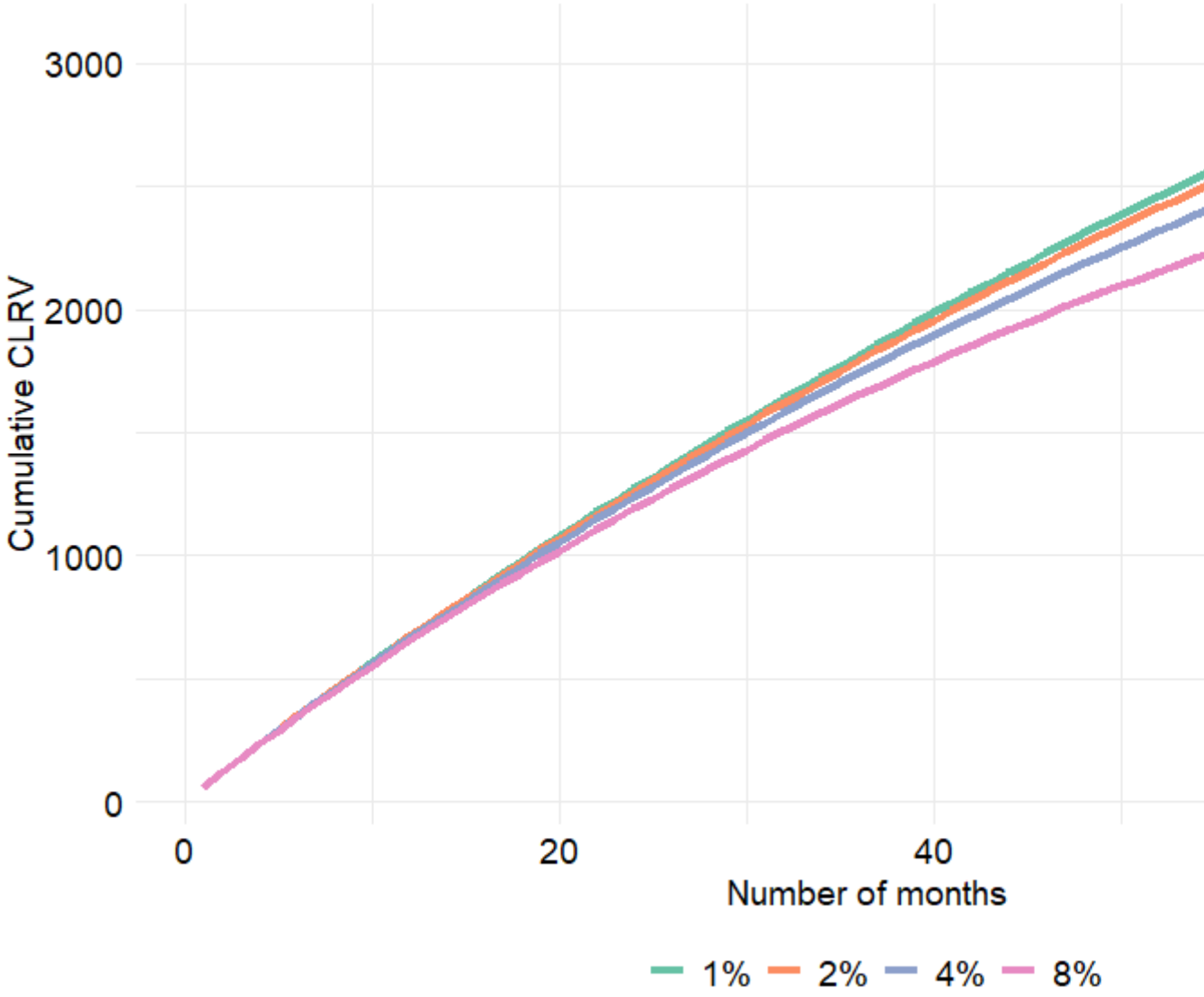

Figure 6.14: Customer Lifetime Raw Value evolution depending on discount rate

covariates used to estimate the survival probabilities $r_{i,t}$. In this part, the analysis is centred on the proportion of month-to-month contracts in the portfolio. The method consists in computing each client's CLRV and the portfolio value V for proportions of monthly contracts ranging from 10 to 90%. Table 6.4.4 examines the portfolio value for 4 interesting scenarios and compare them with the actual share of month-to-month contracts in the data. The results are quite counter-intuitive as V is not maximum for a 10% proportion whereas it has been demonstrated in figure 6.7 that clients enrolled in month-to-month contracts are 22 times more prone to churn than those with two-year contracts. Hence one might think that the lower share of those customers, the higher value of the portfolio.

\begin{table}[H]

\caption{Portfolio estimated value and 95% confidence interval depending on proportion of month-to-month contracts}

% month-to-month contracts	V lower	V	V upper
10%	22,526,500	23,009,290	23,535,331
30%	24,289,738	25,001,468	25,801,858
70%	19,763,605	20,706,782	21,796,005
90%	18,266,063	19,369,938	20,646,166
55.1% (reference)	17,604,144	18,270,000	19,031,648

\end{table}

The following figure illustrates the CLRV distribution for 4 scenarios depending on the proportion of month-to-month contracts. The results are closer to our expectations than those presented in the previous table since CLRV tends to be higher when there are 10% month-to-month contracts in the portfolio.

Ultimately when representing the evolution of customer raw equity, i.e. the portfolio value, according to the share of month-to-month contracts, a decreasing trend is identified which appears to be a rational result.

How to increase Customer Lifetime Raw Value?

At this point of the study, a consistent method to evaluate the firm's portfolio value through the computation of customer lifetime raw value has been implemented. In section 6.3 it has been shown that customers from the *Gold* cluster are characterized by a higher churn rate and lower survival probabilities. This lesser loyalty to the firm has a negative impact on the portfolio value. Furthermore, cluster 2 clients amount to 34% in the customer raw equity which proves their strong potential in terms of CLRV. Therefore, a method to increase *Gold* customers' lifetime is proposed in this last section.

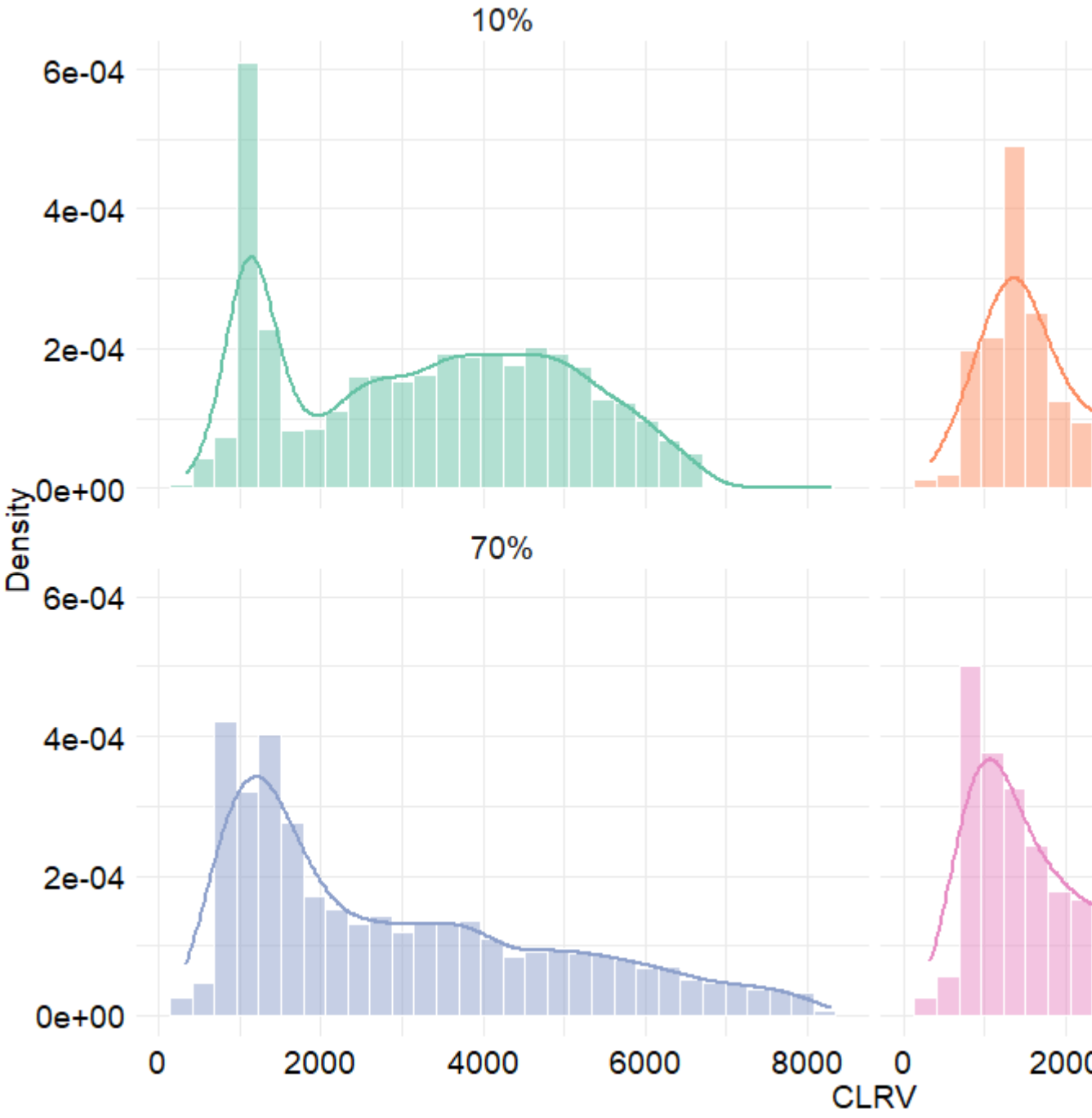

Figure 6.15: CLRV distribution depending on the proportion of month-to-month contracts

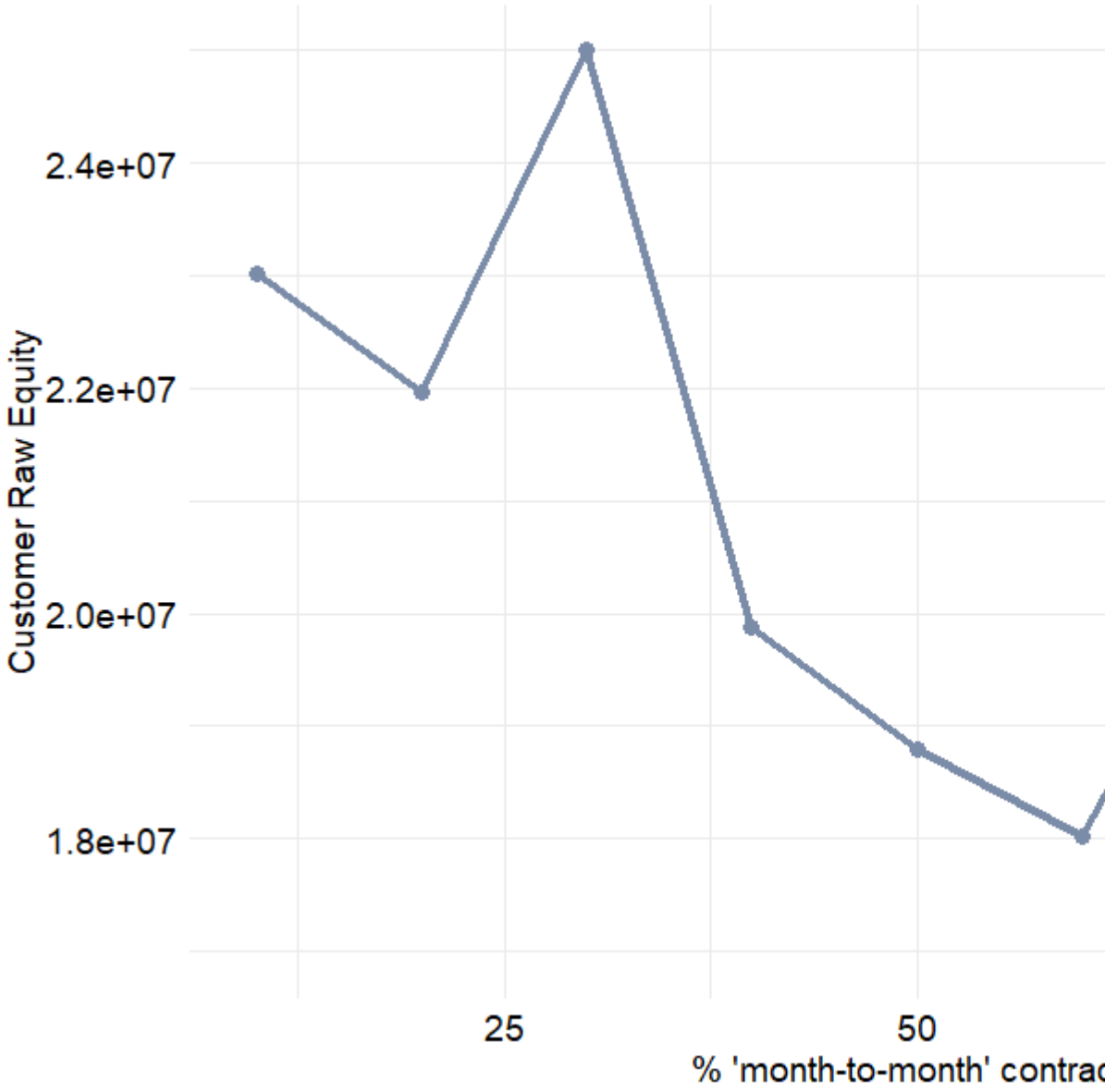

Figure 6.16: Portfolio estimated value (Customer Raw Equity) depending on the proportion of month-to-month contracts

To that end, it is decided to offer additional services for free to every cluster 2 client. These services are represented by the following variables: `Online_Security`, `Online_Backup` and `Tech_Support`. The underlying purpose is to increase the survival probabilities estimated by the Cox model as it has been proved that they have a negative impact on the churn hazard (see figure 6.7).

The results presented in the table below support our hypothesis since the *Gold* cluster total value increases by almost 50% after having provided the clients with free additional options.

Table 6.14: Increase in portfolio value after having provided cluster 2 clients with additional services

	Cluster 2 Value	% Variation
Reference	6,211,027	0.00
With additional options	9,218,937	48.43

Figure 6.17 illustrates the mechanism that leads to an increase in the portfolio value after the "gift" to cluster 2 clients. One can notice a significant shift to the right of the CLRV distribution. This shift is the consequence of an increase in $r_{i,t}$ - client i's survival probability at time t - which is a positive driver of customer value. We are thus able to propose a marketing strategy which aims at boosting the firm's lifetime revenues. Note, however, that our methodology does not take costs into account which implies that the highlighted points should be mitigated.

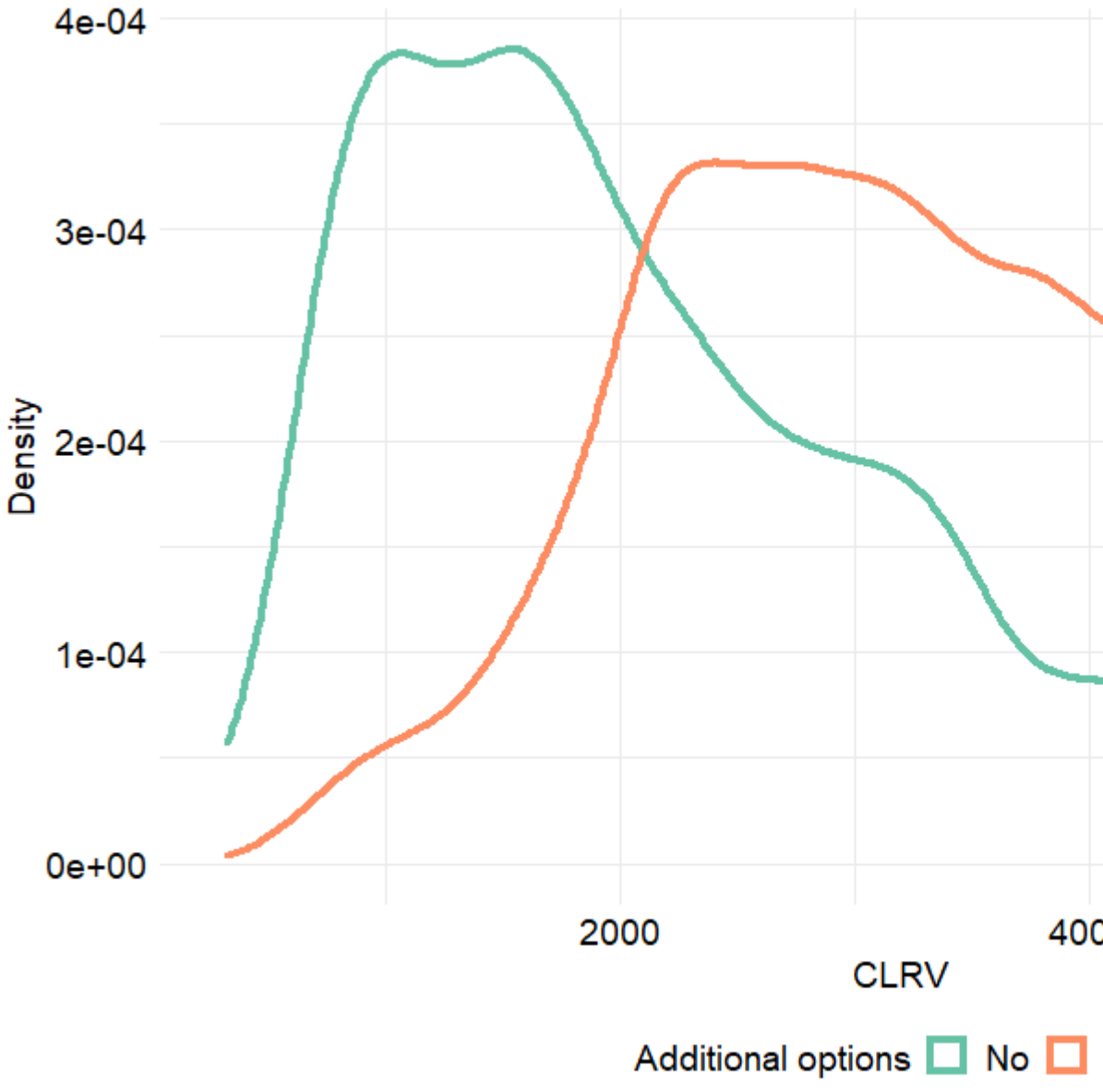

Figure 6.17: Increase in cluster 2 clients' CLRV after after having provided them with additional services

Conclusion

Our study on customer *portfolio*, *attrition* and customer *value* has a threefold purpose. To begin with, the literature review helps the readers to understand the important notions behind the topic's key points. Although it is far from being exhaustive, it summarizes the important Econometrics and Data Science improvements that have been made in terms of portfolio management, churn analysis and customer value estimation. Secondly, a large part of the study centers on the theoretical foundations of duration models which are essential to predict the risk of churn. Finally, chapter 6 explains the ins and outs of a decision support tool whose goal is to manage a portfolio of customers. The method consists in identifying groups of customers who have similar consumption habits, then estimating customer lifetime by the means of duration models to eventually calculate the overall lifetime value of the portfolio based on the computation of customer lifetime raw value.

Nevertheless, some improvements need to be made in order to make our project more consistent to be applied in a business context. The Cox model introduced in section 6.3 presents some limitations since it does not manage to properly estimate the risk of churn. A more reliable estimation method might be adopted such as more complex parametric models or machine learning algorithms for survival data. In addition, it could be relevant to determine the churn reason using competing risk models. As for the data, we think our study would be more impactful with time-varying covariates and more observations. Eventually, as said at the end of the previous chapter, costs are not taken into account in the estimation method. An area for improvement might be to add simulated costs to the data set or retrieving online data on the costs of telecommunication service providers.

Appendix

In this section some proofs of the mathematical concepts used in the study are derived, specifically related to duration analysis. It also consists of additional data visualisation related to the chapter 6.

Hazard function

$$
\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t] / P[T \geq t]}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{(F(t + \Delta t) - F(t))/\Delta t}{S(t)} \\
&= \frac{\mathrm{d}F(t)/\mathrm{d}t}{S(t)} \qquad (6.3) \\
&= \frac{f(t)}{S(t)} \\
&= -\frac{\mathrm{d}S(t)/\mathrm{d}t}{S(t)} \\
\lambda(t) &= \frac{-\mathrm{d}\ln\left(S(t)\right)}{\mathrm{d}t}
\end{aligned}
$$

Link between cumulative hazard and survivor functions

$$\begin{aligned}
&\Lambda(t) = \int_0^t \lambda(s)ds \\
&\iff \Lambda(t) = \int_0^t \frac{f(s)}{S(s)}ds \\
&\iff \Lambda(t) = -\ln\big(S(t)\big) \\
&\iff S(t) = \exp\big(-\Lambda(t)\big)
\end{aligned} \tag{6.4}$$

Contribution to the partial likelihood function in PH models

$$\begin{aligned}
\mathbb{P}\big[T_j = t_j | R(t_j)\big] &= \frac{\mathbb{P}\big[T_j = t_j | T_j \geq t_j\big]}{\sum_{l \in R(t_j)} \mathbb{P}\big[T_l = t_l | T_l \geq t_j\big]} \\
&= \frac{\lambda_j(t_j|\mathrm{x_j}, \beta)}{\sum_{l \in R(t_j)} \lambda_l(t_l|\mathrm{x_l}, \beta)} \\
&= \frac{\lambda_0(t_j, \alpha)\phi(\mathrm{x_j}, \beta)}{\sum_{l \in R(t_j)} \lambda_0(t_j, \alpha)\phi(\mathrm{x_l}, \beta)} \\
\mathbb{P}\big[T_j = t_j | R(t_j)\big] &= \frac{\phi(\mathrm{x_j}, \beta)}{\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)}
\end{aligned} \tag{6.5}$$

Partial likelihood function in PH models

Based on equation (3.18), one can derive the probability that all spells completed at t_j ends in the j^{th} failure time, such that:

$$
\begin{aligned}
\mathcal{L}_{p,\ t_j} &= \mathbb{P}[T_1 = t_j, \ldots, T_{d_j} = t_j \mid R(t_j)] \\
&= \Pi_{m \in D(t_j)}\ \mathbb{P}[T_m = t_j | R(t_j)] \\
&= \Pi_{m \in D(t_j)}\ \frac{\phi(\mathrm{x_m}, \beta)}{\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)} \\
&= \Pi_{m \in D(t_j)}\ \phi(\mathrm{x_m}, \beta) \times \Pi_{m \in D(t_j)}\ \frac{1}{\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)} \\
\mathcal{L}_{p,\ t_j} &= \frac{\Pi_{m \in D(t_j)}\ \phi(\mathrm{x_m}, \beta)}{\left[\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)\right]^{d_j}}
\end{aligned}
\tag{6.6}
$$

The joint probability over the k ordered discrete failure times then becomes:

$$
\begin{aligned}
\mathcal{L}_p &= \Pi_{j=1}^{k}\ \mathcal{L}_{p,\ t_j} \\
\mathcal{L}_p &= \Pi_{j=1}^{k}\ \frac{\Pi_{m \in D(t_j)}\ \phi(\mathrm{x_m}, \beta)}{\left[\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)\right]^{d_j}}
\end{aligned}
\tag{6.7}
$$

Multiple correspondence analysis

Hierarchical clustering on principal components

Cluster visualisation

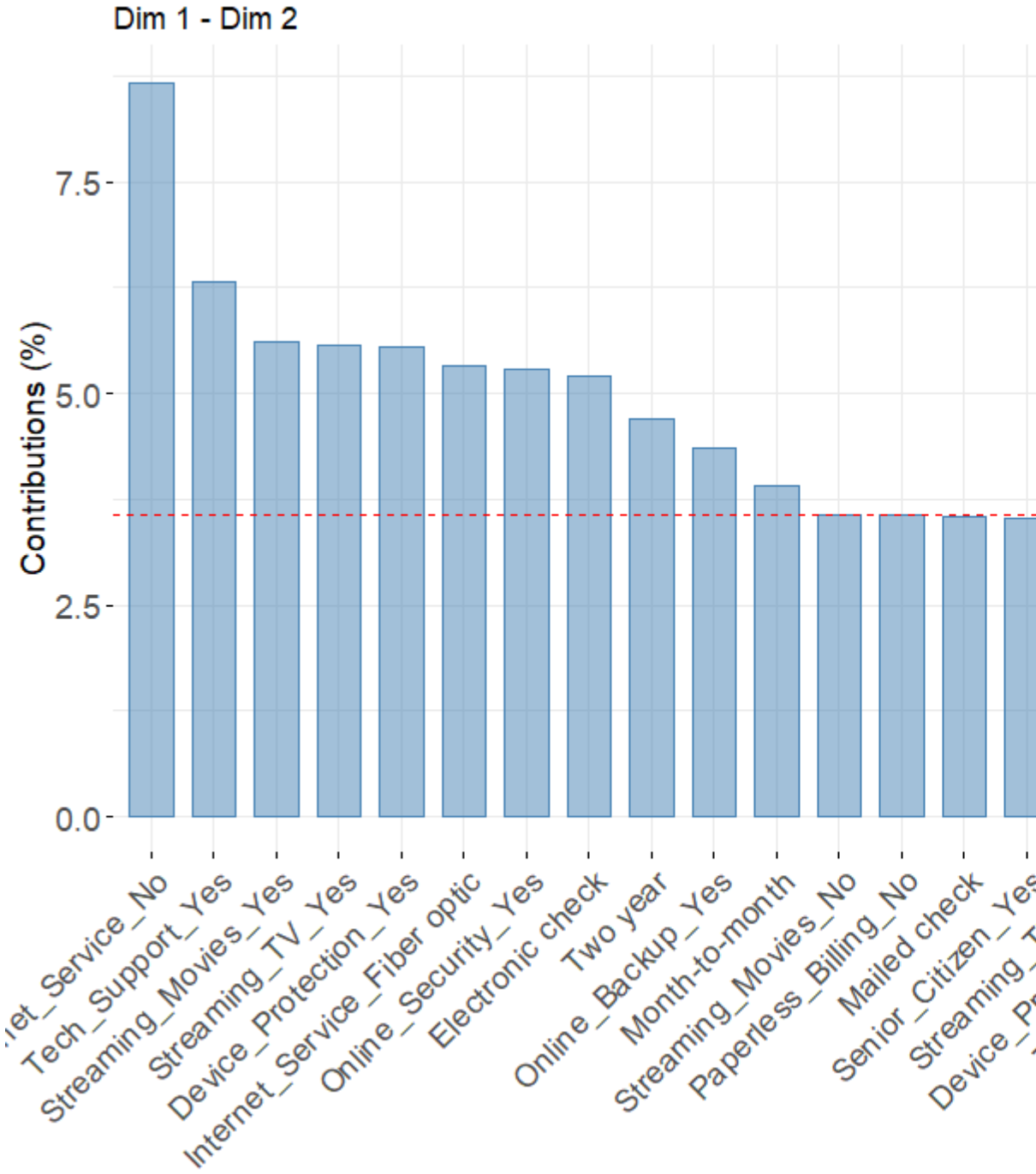

Figure 6.18: MCA - Categories contribution to axes 1 and 2

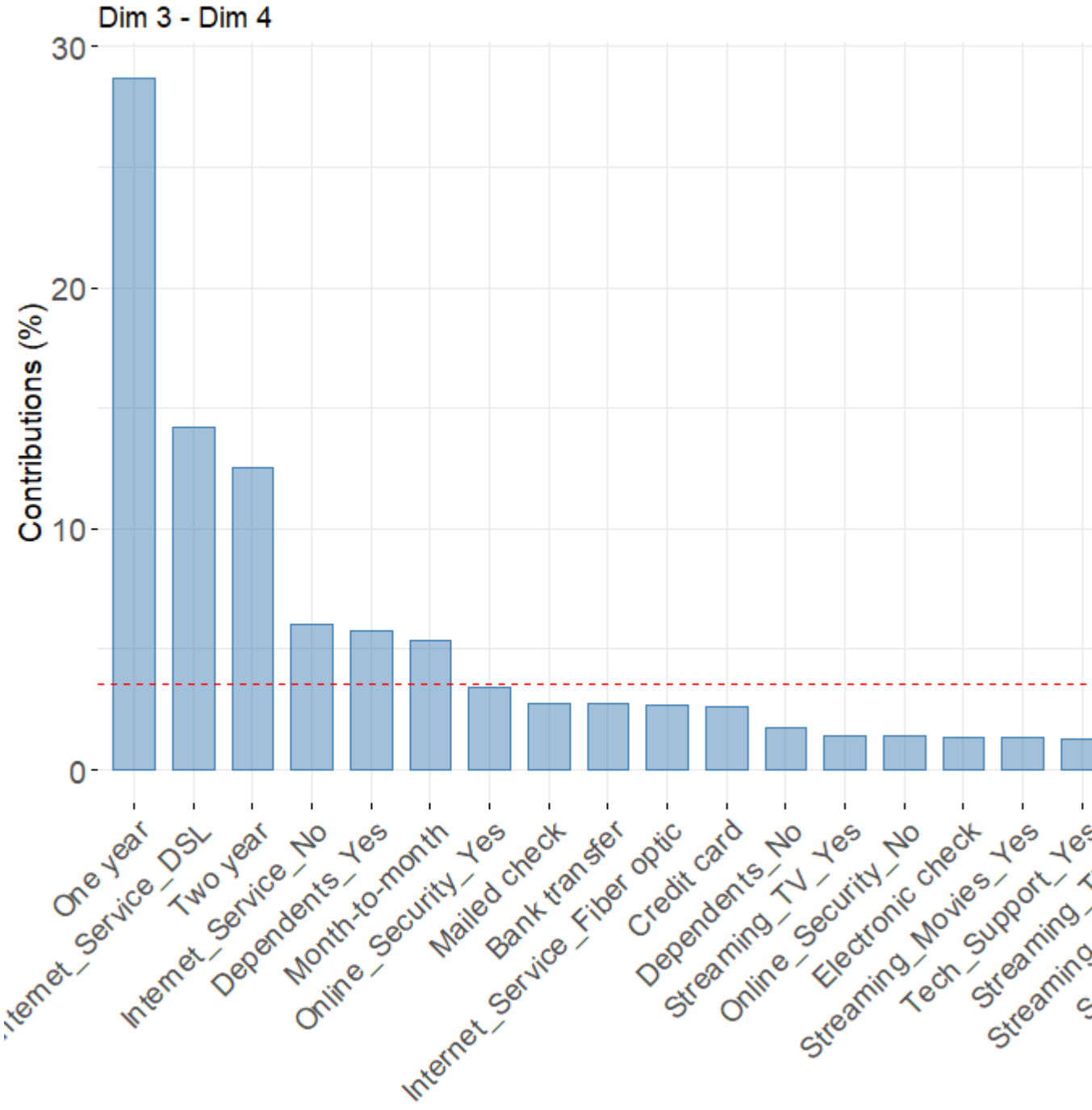

Figure 6.19: MCA - Categories contribution to axes 3 and 4

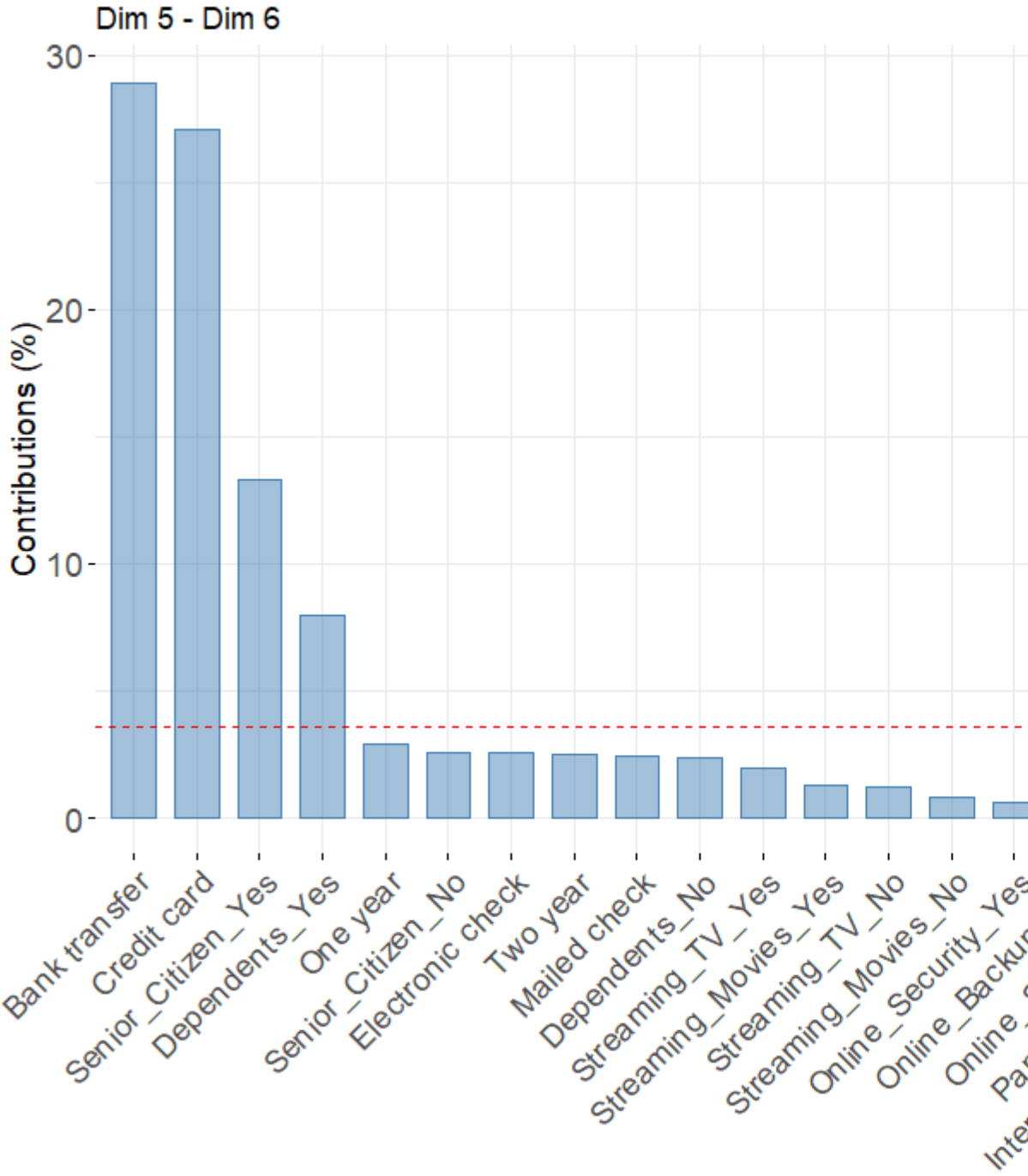

Figure 6.20: MCA - Categories contribution to axes 5 and 6

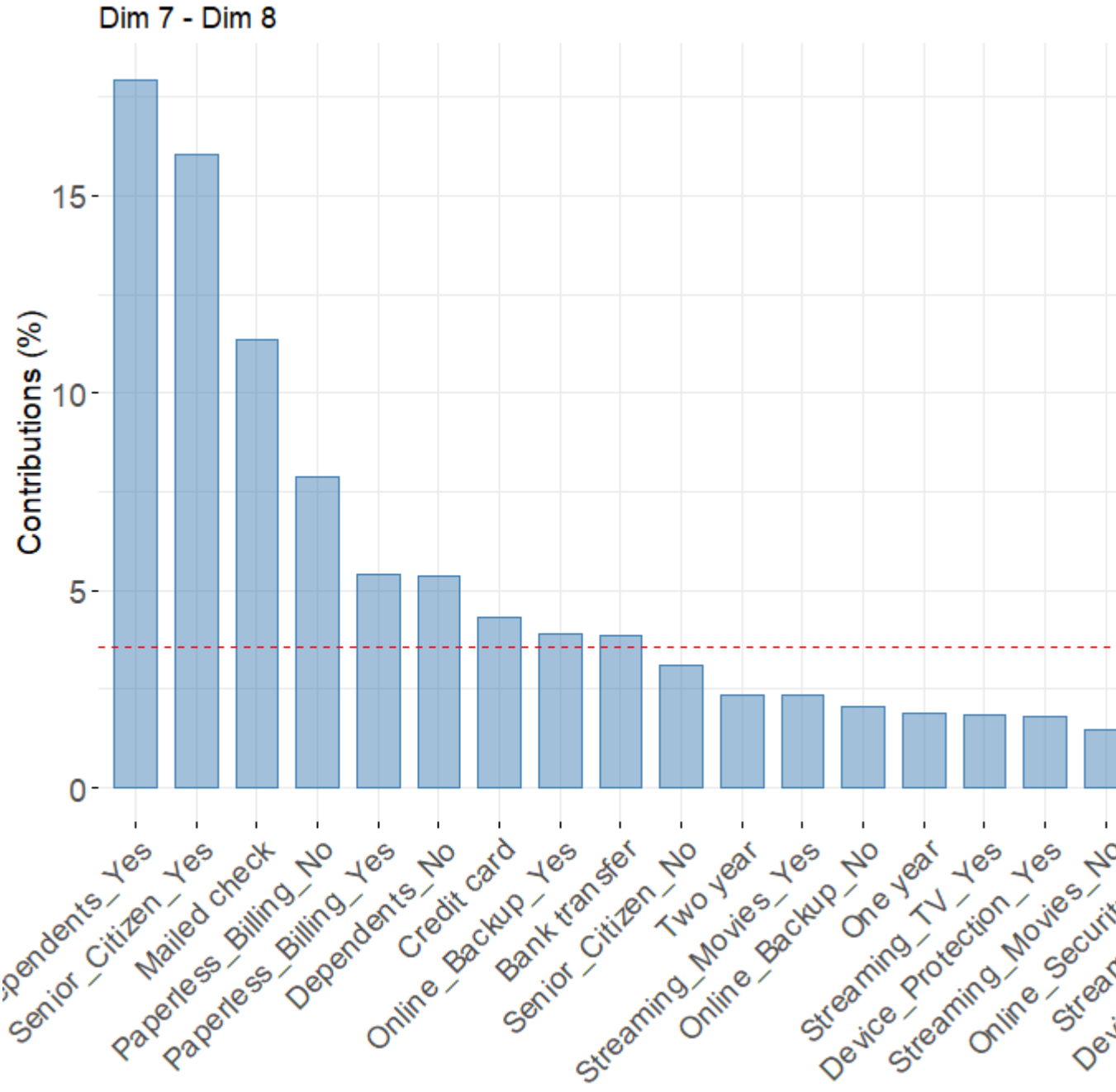

Figure 6.21: MCA - Categories contribution to axes 7 and 8

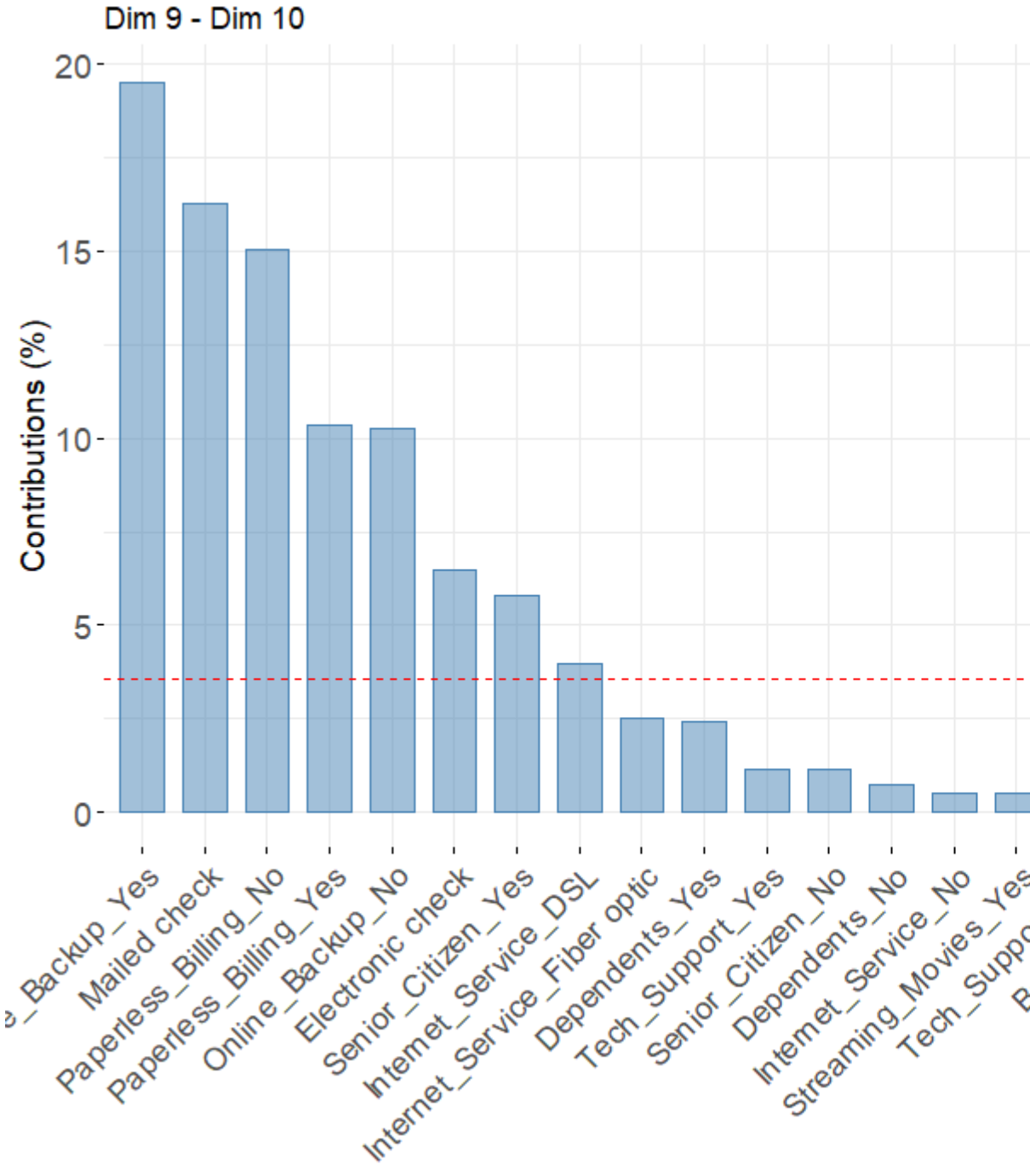

Figure 6.22: MCA - Categories contribution to axes 9 and 10

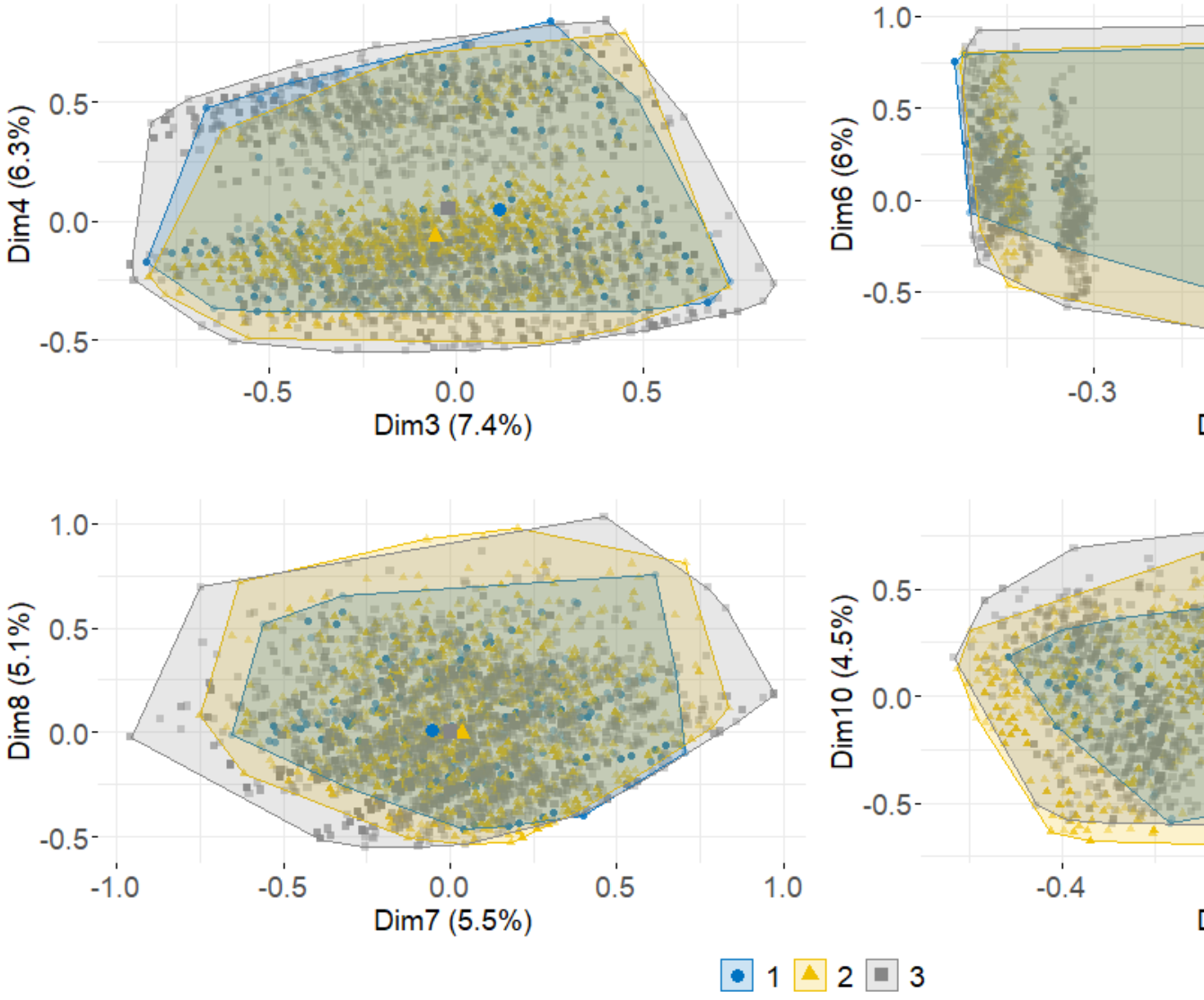

Figure 6.23: Cluster visualisation onto MCA axes

Bibliography

Bellani, C. (2019). *Predictive Churn Models in Vehicle Insurance*. PhD thesis, Universidade Nova de Lisboa.

Binder and Schumacher (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9(14).

Blattberg and Deighton (1996). Manage marketing by the customer equity test. *Harvard Business Review*, pages 136–144.

Bley, Ng, and Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 36.

Borle, S. and Singh, S. S. (2008). Customer lifetime value measurement. *Management Science*, 54(1):110–112.

Bousquet, A. (2021). Gestion optimale de portefeuilles de brevets.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*.

Capon and Glazer (1987). Marketing and technology: a strategic alignment. *Journal of Marketing*, 51:1–14.

Day, G. (1977). Diagnosing the product portfolio. *Journal of Marketing*, 41:29–38.

Fader and al. (2005). "counting your customers" the easy way: An alternative to the pareto/nbd. *Management Science*, 24(2):275–284.

Gupta, Hanssens, and Hardie, D. (2006). Modeling customer lifetime value. *Journal of Service Research*, 9(2):139–155.

Gupta and Lehmann, D. R. (2003). Customers as assets. *Journal of Interactive Marketing*, 17(1):9–24.

Harrell, F. (1984). *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*.

Homburg, C., Steiner, V. V., and Totzek, D. (2009). Managing dynamics in a customer portfolio. *Journal of Marketing*, 73(5):70–89.

Ishwaran, H. and al. (2011). Random survival forests for high-dimensional data.

Ishwaran, H., Kogalur, U., Blackstone, E., and Lauer, M. (2008). Random survival forests. *Ann. Appl. Statist.*, 2(3):841–860.

Jackson, C. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*, 70(8):1–33.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88:457–467.

Liu, W. (2019). *Inclusive Underwriting: the case of Cardiovascular Risk Calculator*. PhD thesis, ENSAE ParisTech.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1):71–91.

Pérez Marín, A. M. (2006). *Survival methods for the analysis of customer lifetime duration in insurance*. PhD thesis.

Reinartz and Kumar. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, 67(1):77–99.

Scholler, J. (2020). *M1 Analyse de données exploratoire - Classification non supervisée*. Université de Tours.

Scholler, J. (2021a). *M1 Analyse de données exploratoire - ACM*. Université de Tours.

Scholler, J. (2021b). *M1 Data Mining - Decision trees*. Université de Tours.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, 44(1):35–47.

Slof, Frasincar, and Matsiiako (2021). A competing risks model based on latent dirichlet allocation for predicting churn reasons. *Decision Support Systems*, 146.

Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Thakur, R. and Workman, L. (2016). Customer portfolio management (cpm) for improved customer relationship management (crm): Are your customers platinum, gold, silver, or bronze? *Journal of Business Research*, 69:4095–4102.

Wind, Y. and Mahajan, V. (1981). Designing product and business portfolios. *Harvard Business Review*, 59(1):155–165.