Portfolio, Churn & Customer Value

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

2022-01-17

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, it has been decided to address the following problematic:

How to estimate the overall value of a customer portfolio?

Several key steps will be focused on in order to provide insights on the previous question:

- 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

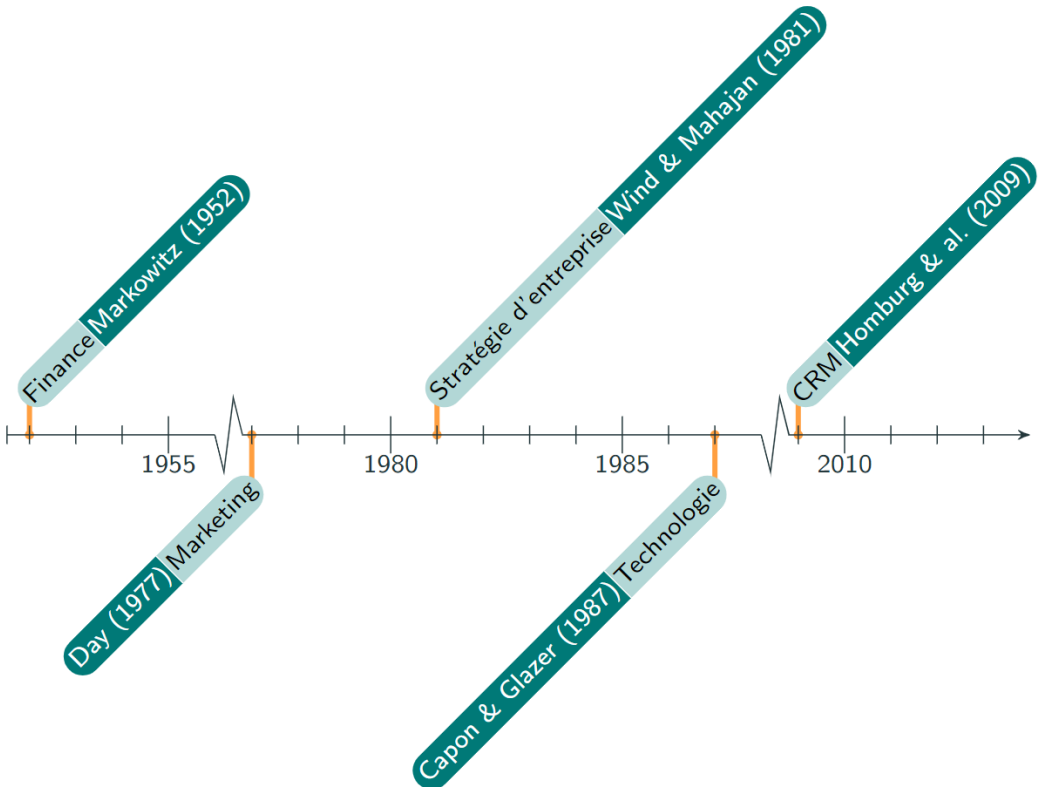

Figure 2.1: A timeline on the concept of portfolio

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.

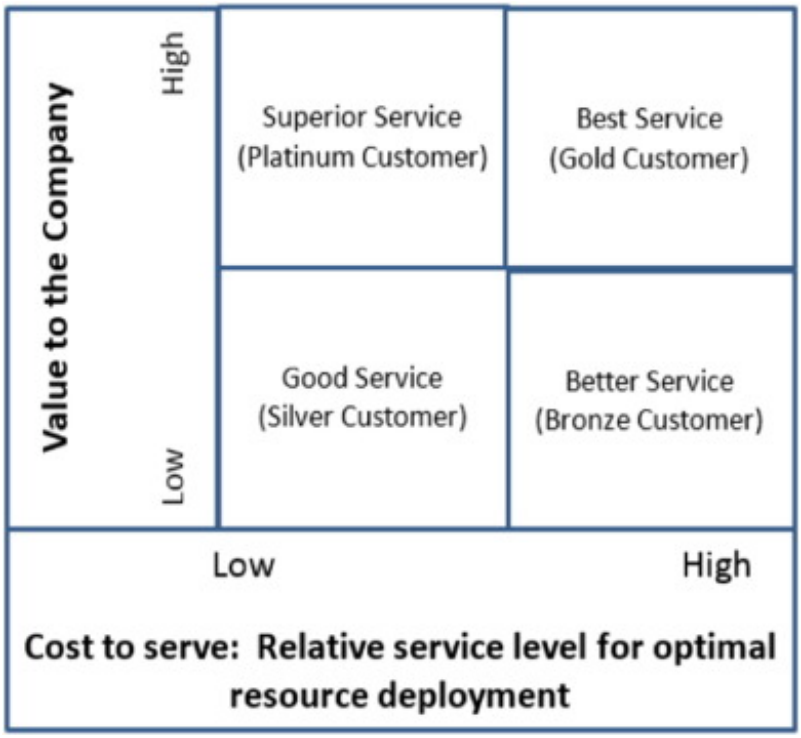

Figure 2.2: Csutomer Portfolio Management (CPM) Matrix

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.

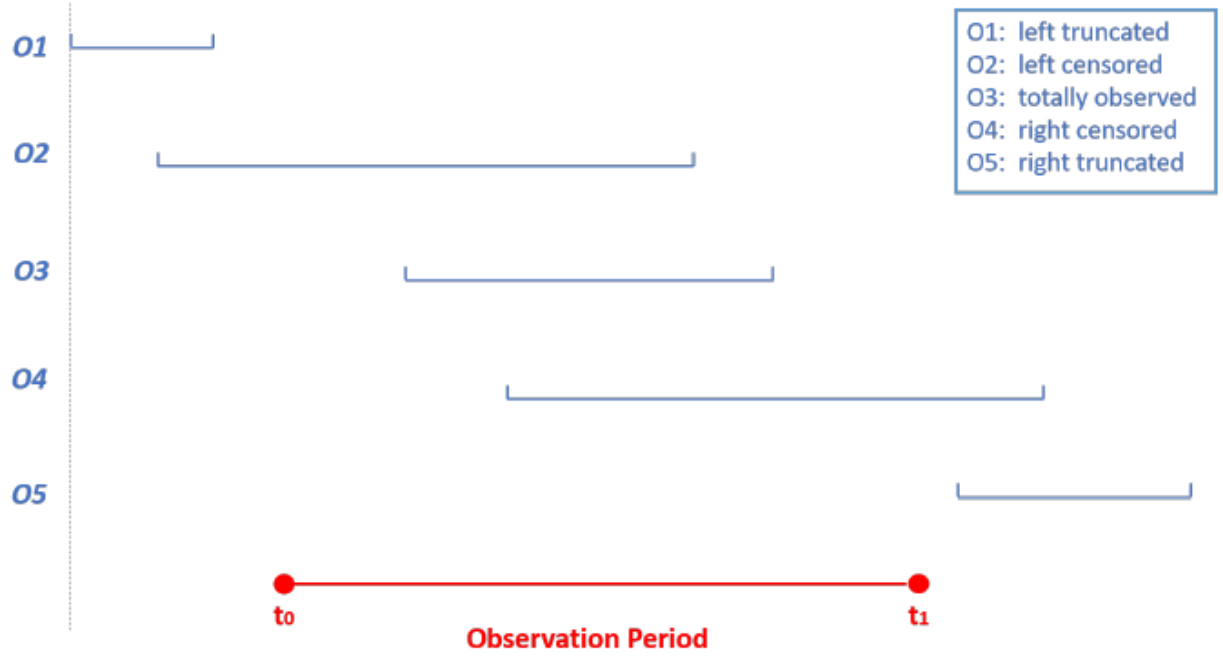

Figure 3.1: Censored and truncated data

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:

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

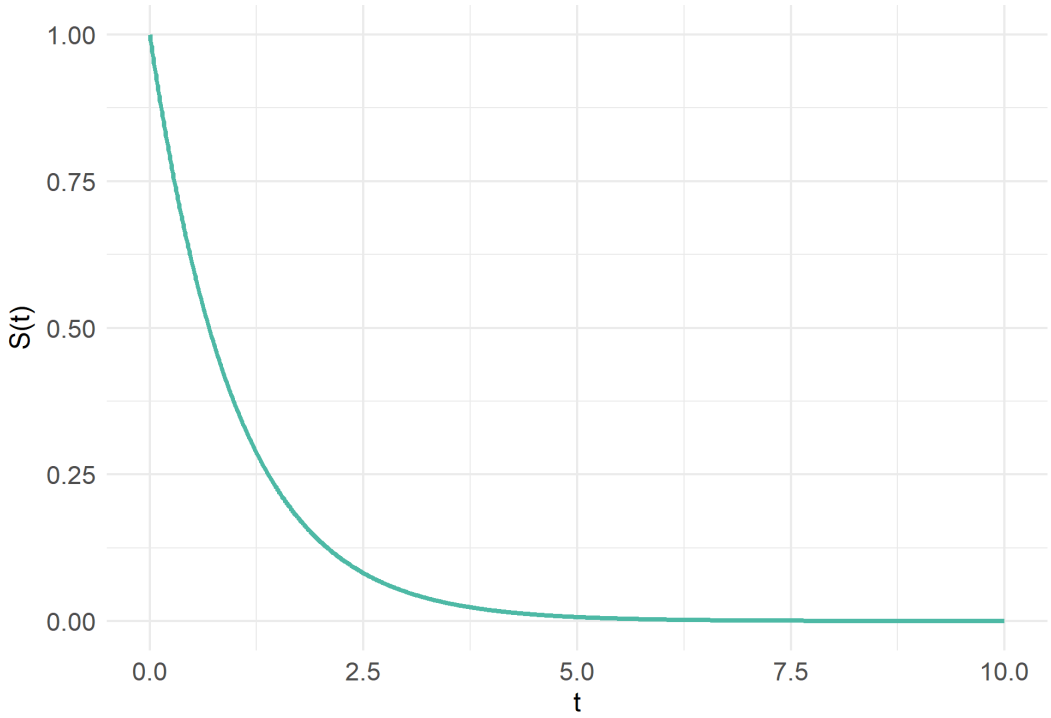

Figure 3.2: Survival function $S_T(t)$ with $T \sim \mathcal{E}(1)$

Using the Bayes formula, equation (3.3) can also be written as (see proof (5.2) 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:

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

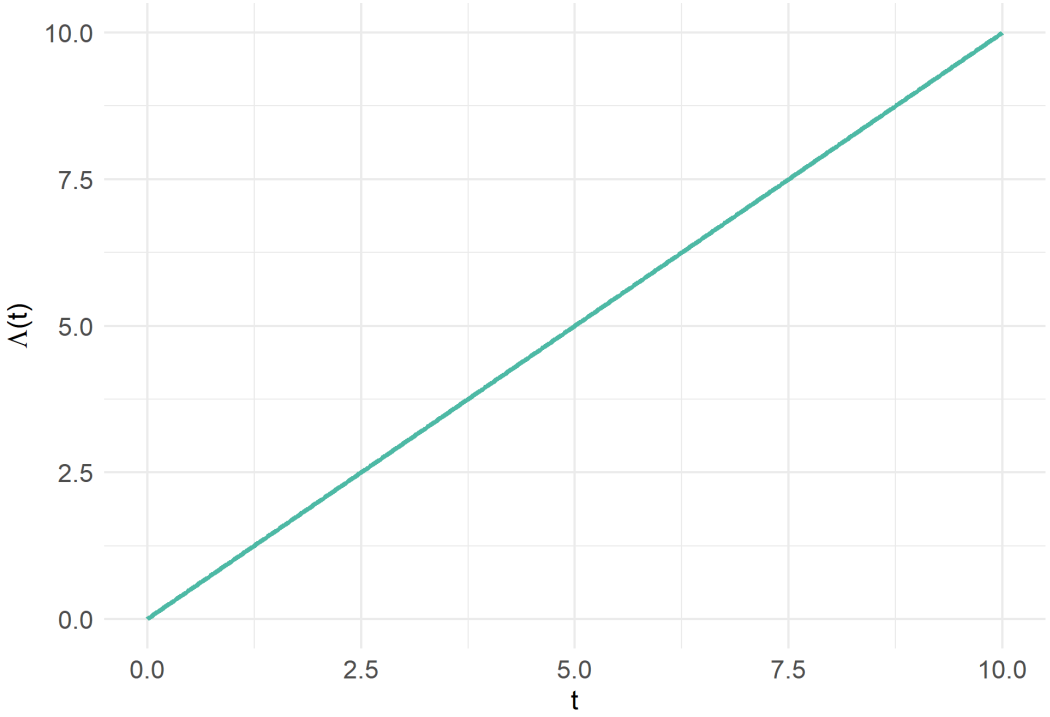

Figure 3.3: Cumulative Hazard function $\Lambda_T(t)$ with $T \sim \mathcal{E}(1)$

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 (5.3) 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).

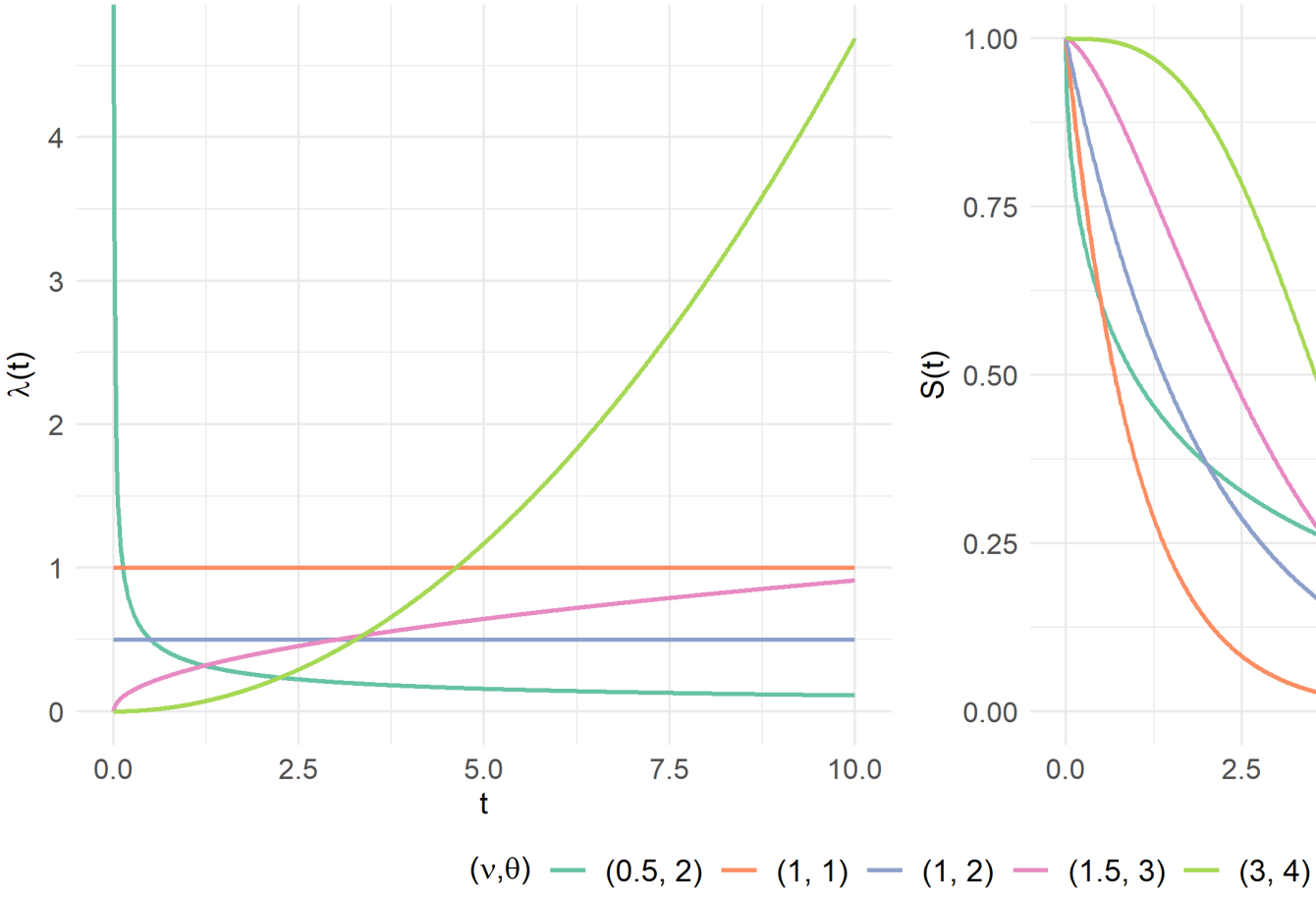

Figure 3.4: Hazard and Survival functions with $T \sim \mathcal{W}(\nu, \theta)$

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 (5.4) 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:

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

The latter formulation of the partial likelihood function is explained in more details in proofs (5.5) and (5.6) 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. $\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\mathrm{e}^{\mathbf{x}'\beta}$.

Chapter 4

Machine Learning

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

– Tom M. Mitchell –

This chapter introduces machine learning algorithms used to model customer *portfolio*. We firstly explain the techniques which aim to enrich the standard survival tools defined in the previous chapter. A second part depicts some regression models that are robust to predict customer lifetime value. Finally, the prediction performance associated to predictive methods are described.

4.1 Machine Learning for Survival Data

In chapter 3, 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.

4.1.1 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 4.1 (Scholler, 2021). The algorithm is a recursive partitioning which requires a criterion for choosing the *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

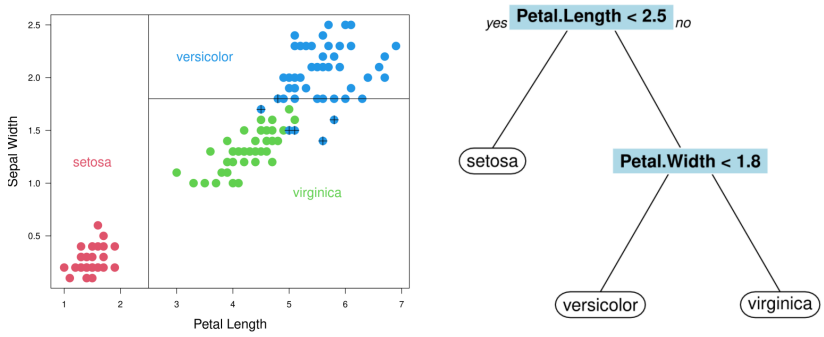

Figure 4.1: Clasification decision tree

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):

$$\chi^2_{\text{logrank}} = \frac{\left[\sum_{j=1}^{k} (d_{0j} - r_{0j} \times d_j/r_j)\right]^2}{\sum_{j=1}^{k} \frac{r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}} \tag{4.1}$$

4.1.2 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:

$$\widehat{H_l}(t) = \sum_{t_{j,l}<t} \frac{d_{j,l}}{r_{j,l}} \tag{4.2}$$

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.

4.1.3 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.

4.2 Regression methods

Customer Lifetime Value being a quantitative variable, machine learning regression models are adapted to predict this quantity. Regression analysis is a fundamental concept in the field of machine learning. It falls under supervised learning wherein the algorithm is trained with both input features and output labels. It helps in establishing a relationship among the variables by estimating how one variable affects the other. In this context, we present some famous machine learning algorithms that can be implemented to estimate the relationship between several features and a continuous target variable.

4.2.1 Linear models

Model formulation

Linear models are called that way as the target value is expected to be a linear combination of the features. Let $\hat{\mathrm{y}}$ be the vector of predicted values, β the set of parameters to optimise and $\mathbf{x}$ the feature vector with $\mathbf{x} = [\mathbf{1}, \mathrm{x}_1, \ldots, \mathrm{x}_P]$. The linear relationship between $\hat{\mathrm{y}}$ and $\mathbf{x}$ can be written as follows:

$$\hat{\mathrm{y}} = \beta_0 + \beta_1 \mathrm{x}_1 + \cdots + \beta_P \mathrm{x}_P \tag{4.3}$$

with P the number of explanatory variables.

Equation (4.3) can be rewritten in a vectorized way such that $\hat{\mathrm{y}} = \mathbf{x}\beta$.

Ordinary Least Squares

The most standard regression algorithm is linear regression in which the loss function is the residual sum of squares between the observed targets in the dataset y and the targets predicted by the linear formulation $\hat{\mathrm{y}}$. This method is called ordinary least squares (OLS). Mathematically, the set of parameters β is chosen with a view of minimizing:

$$l_{OLS} = ||\mathrm{y} - \mathbf{x}\beta||_2 = (\mathrm{y} - \mathbf{x}\beta)^{'}(\mathrm{y} - \mathbf{x}\beta) \tag{4.4}$$

Regularization

In order to address some of the problems encountered with the OLS method, regularization can be used. Regularization techniques can be employed when the explanatory variables are highly correlated.

Ridge regression is a linear regression with a quadratic constraint on the coefficients. Here, the coefficients minimize a penalized residual sum of squares: the higher the penalty term, the more large coefficients are discouraged and the less risk of overfitting. Formally, the ridge loss function is based on l_2-norm and is expressed as:

$$l_{Ridge} = ||\mathrm{y} - \mathbf{x}\beta||_2 + \alpha||\beta||_2 \tag{4.5}$$

where α is the shrinkage parameter which controls the penalization on the value of the model's coefficients.

Lasso regression is another example of penalized estimation technique with this time a linear constraint on the β vector. It is helpful in reducing the number

of features upon which the target variable is dependent. The loss function is based on l_1-norm and can be written as follows:

$$l_{Lasso} = ||\mathrm{y} - \mathbf{x}\beta||_2 + \alpha||\beta||_1 \tag{4.6}$$

Finally, ElasticNet is a linear regression model trained with both l_1 and l_2-norm regularization of the coefficients and is useful when there are multiple features that are correlated with one another. In this context, the loss function is derived as follows:

$$l_{EN} = ||\mathrm{y} - \mathbf{x}\beta||_2 + \alpha\rho||\beta||_1 + \alpha\frac{1-\rho}{2}||\beta||_2 \tag{4.7}$$

Gradient Descent

The loss functions presented above need to be optimized to obtain the optimal set of parameters that best represents the linear relationship between y and $\mathbf{x}$. In other words, a minimization problem needs to be solved. Gradient descent is an algorithm whose goal is to find the maximum (or minimum) of a given function f. The gradient of f is defined as the vector of partial derivatives and gives the input direction in which f most quickly increases. The gradient descent approach consists in picking a random starting point, computing the gradient, taking a small step in the opposite direction of the gradient and repeating with the new starting point until some criterion is met.

4.2.2 More advanced regression models

Linear models are considered the most common and easy-to-understand models when it comes to predict a quantitative variable such as customer *value*. Nonetheless, more advanced techniques can sometimes be implemented in order to obtain more accurate predictions.

Generalized Additive Model (GAM)

GAM is a statistical model in which the response variable y depends linearly on unknown smooth functions of some feature vector $\mathbf{x}$, and interest focuses on inference about these smooth functions called f_p. The f_p functions may be either parametric (polynomial), semi-parametric or nonparametric (smoothing splines) leading to more flexibility in the model's assumptions on the actual relationship between y and $\mathbf{x}$. A link function g can also be introduced to specify this relationship and the model's general formulation is as follows:

$$g(\mathrm{E[y]}) = \sum_{p=1}^{P} f_p(\mathrm{x}_p) \tag{4.8}$$

Other models

State-of-the-art algorithms can also be employed to describe the relationship between the continuous variable and the feature vector. Among them, one can find random forest of regression trees, gradient boosting as well as multi-layer perceptron (MLP) regressor.

4.3 Performance Metrics

Once several models have been trained, comparing their performance is an essential step to choose the right model. Statistical metrics helps in the model selection stage by providing an indication of goodness of fit. In other words, they are a measure of how well unseen samples are likely to be predicted by a given model. As two families of algorithms have been introduced, two types of statistical metrics are required.

4.3.1 Metrics for Regression models

Notations

Let us consider a set of n samples where y_i is the true value and $\hat{y}_i$ the predicted value related to the i^{th} sample.

Coefficient of determination

Historically, the R^2 score or coefficient of determination is the most common tool to compare the performance of two regression models. It computes the proportion of variance in the target variable that has been explained by the independent variables in the model. Mathematically, we have:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(y_i - \overline{y})^2} \tag{4.9}$$

From equation (4.9) it can be derived that $R^2 \in [0, 1]$. A perfect model would obtain an R^2 score of 1. Conversely, a constant model that always predicts the expected value of the target, disregarding the input features, would get a score of 0. The R^2 value remains to be carefully analysed as a large value does not

necessarily mean a high quality model. Indeed, the R^2 score being the square of the correlation coefficient, if the latter is close to 1 (or -1), the coefficient of determination will be close to 1. But if the correlation is spurious, the R^2 value will be meaningless as well as the trained model (see Bousquet and Concettini (2019) for more details).

Mean Squared Error (MSE)

MSE can be regarded as a risk metric corresponding to the expected value of the squared error or loss. Formally, MSE is expressed as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.10}$$

4.3.2 Metrics for Survival models

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 $(y_1; \hat{y}_1)$ and $(y_2; \hat{y}_2)$. y_i and $\hat{y}_i$ 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.

$$c = \mathbb{P}(\hat{y}_1 > \hat{y}_2 | y_1 > y_2) \tag{4.11}$$

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 (4.11) can then be rewritten as follows:

$$c = \frac{\#\text{concordant pairs}}{\#\text{concordant pairs} + \#\text{discordant pairs}}$$
$$c = \frac{\sum_{i \neq j} \mathbf{1}_{\eta_i < \eta_j} \mathbf{1}_{T_i > T_j} d_j}{\sum_{i \neq j} \mathbf{1}_{T_i > T_j} d_j} \tag{4.12}$$

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.

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:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{1}_{\{t_i > t\}} - \widehat{S}(t|\mathrm{x}_i) \right)^2 \tag{4.13}$$

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	CLTV
3668-QPYBK	53.85	DSL	2	1	3239
9237-HQITU	70.70	Fiber optic	2	1	2701
9305-CDSKC	99.65	Fiber optic	8	1	5372
7892-POOKP	104.80	Fiber optic	28	1	5003
0280-XJGEX	103.70	Fiber optic	49	1	5340
4190-MFLUW	55.20	DSL	10	1	5925

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, the `CLTV` variable represents each customer's *value* through measurement of customer lifetime value and is used as target in regression models.

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

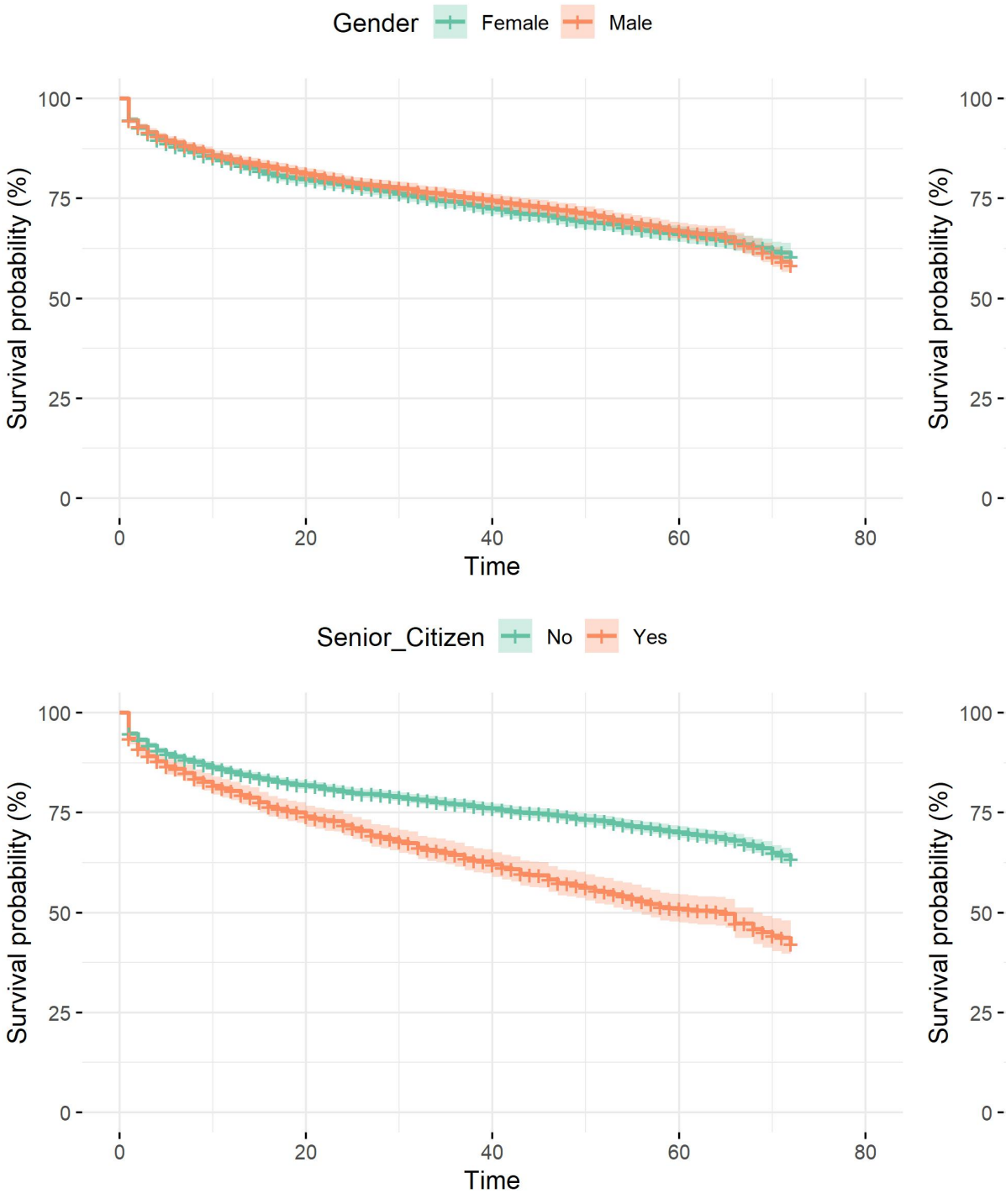

Figure 5.1: Kaplan-Meier survival function depending on demographic information

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 5.3 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-

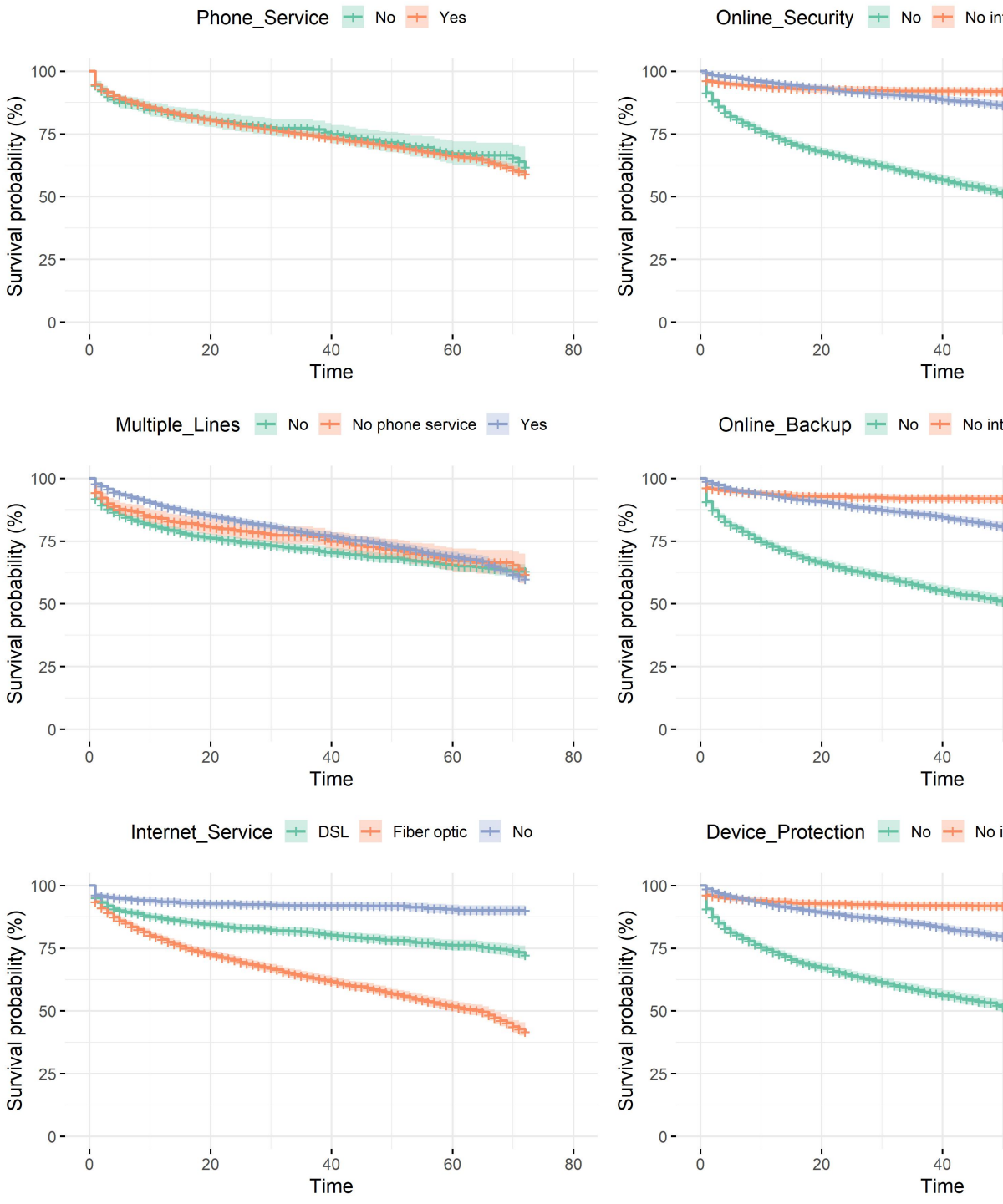

Figure 5.2: Kaplan-Meier survival function depending on services subscribed

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 CLTV: Customer Lifetime Value

In the dataset, the *value* of the fictional TSP's customers is measured by the discrete quantitative variable called `CLTV`. In this context, it is decided to draw histogram and density plots related to `CLTV` depending on the three types of explanatory variables. Anova tests are also implemented to verify whether `CLTV` has different values in the treatment variables' categories. Anova is a generalization of the Student's t test allowing to compare more than two groups. The test's statistic computes the ratio between variance between sample and variance within samples and is Fisher distributed. A low ratio indicates that there is no significant difference between the means of the samples being compared.

Demographic data

Based on the Anova tests' results, `Partner` and `Dependents` seem to be statistically discriminant in terms of customer lifetime value which is not the case for `Gender` and `Senior_Citizen`.

Table 5.5: Anova test between CLTV and demographic variables

	F statistic	Df1	Df2	p-value
Partner	139.76	1	7030	6e-32
Dependents	24.89	1	7030	6.2e-07
Gender	0.39	1	7030	5.3e-01
Senior_Citizen	0.09	1	7030	7.6e-01

The figure below illustrates how different CLV is between customers with a partner and those without, as well as between those with children or other dependents and those without. To put it another way, having a partner in life of dependents tends to increase customer lifetime value as shown by the last two plots.

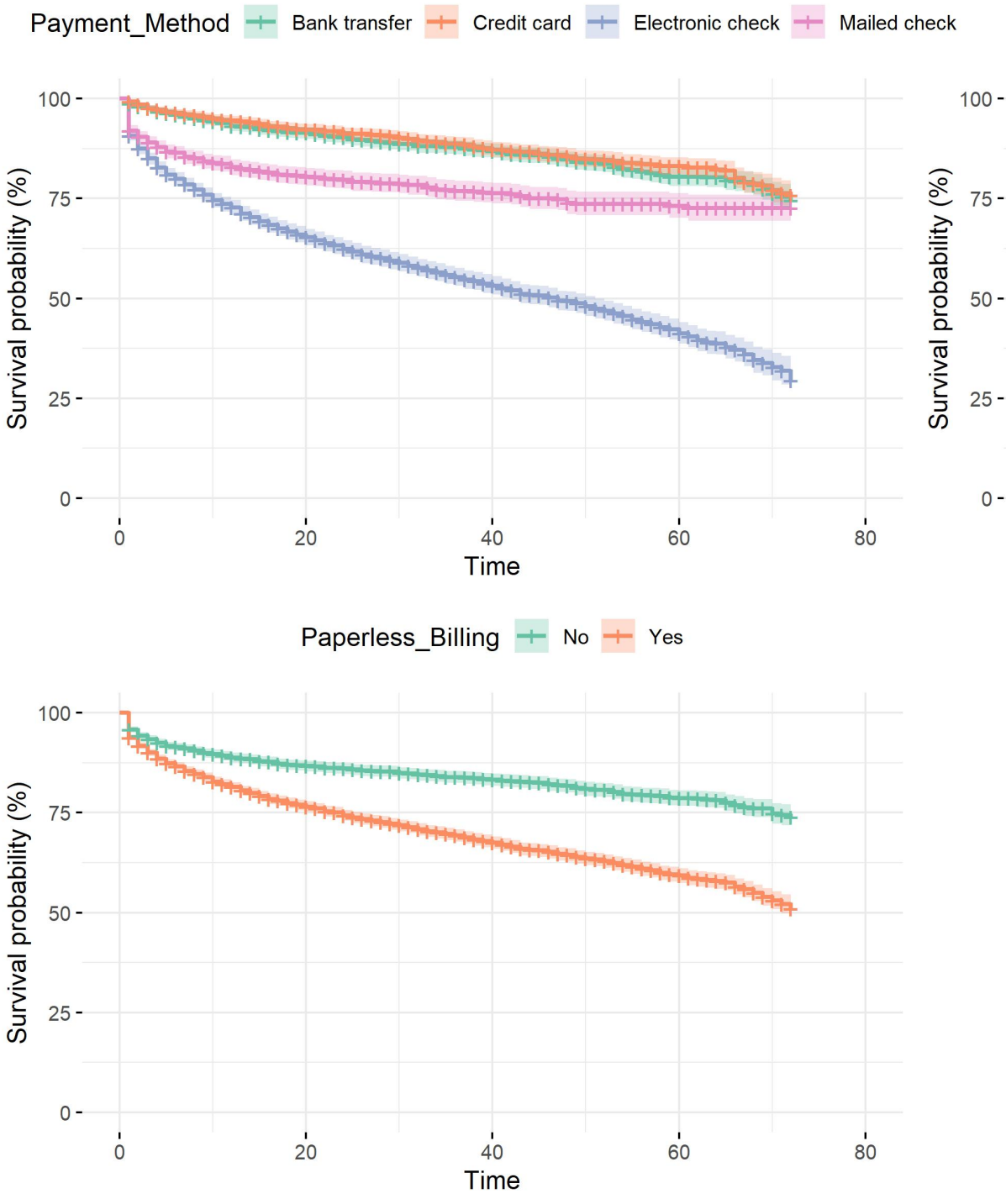

Figure 5.3: Kaplan-Meier survival function depending on customer account information

Figure 5.4: Histogram and density plots of customer lifetime value depending on demographic information

Data on services subscribed

When one wants to identify factors influencing customer lifetime value, it is relevant to consider variables related to services subscribed by customers. The Anova tests' results indicate that `CLTV` has significant different values in each group of all services related variables, except for `Internet_Service`. The most important difference can be noted for `Online_Backup` and `Online_Security` variables, whereas there is more homogeneity in the `Phone_Service` groups.

Table 5.6: Anova test between CLTV and services information variables

	F statistic	Df1	Df2	p-value
Online_Backup	138.57	1	7030	1.1e-31
Online_Security	138.21	1	7030	1.3e-31
Device_Protection	105.23	1	7030	1.6e-24
Tech_Support	101.85	1	7030	8.7e-24
Streaming_Movies	90.96	1	7030	2e-21
Streaming_TV	79.58	1	7030	5.8e-19
Phone_Service	3.65	1	7030	5.6e-02
Multiple_Lines	3.65	1	7030	5.6e-02
Internet_Service	0.56	2	7029	5.7e-01

From figure 5.5 one can note that subscribing to additional services like having multiple lines, online security and backup, device protection or using the streaming movie service significantly enhance customer lifetime value. These variables may be interesting predictors of `CLTV` in regression models.

Customer account data

Table 5.7 depicts that `CLTV` is statistically different according to the type of contract and the payment method. However, paperless billing does not seem to influence customer lifetime value.

Table 5.7: Anova test between CLTV and customer account data

	F statistic	Df1	Df2	p-value
Contract	274.28	2	7029	2e-115
Payment_Method	52.66	3	7028	1.2e-33
Paperless_Billing	0.77	1	7030	3.8e-01

The three plots below provide details to previous results as it can be noticed

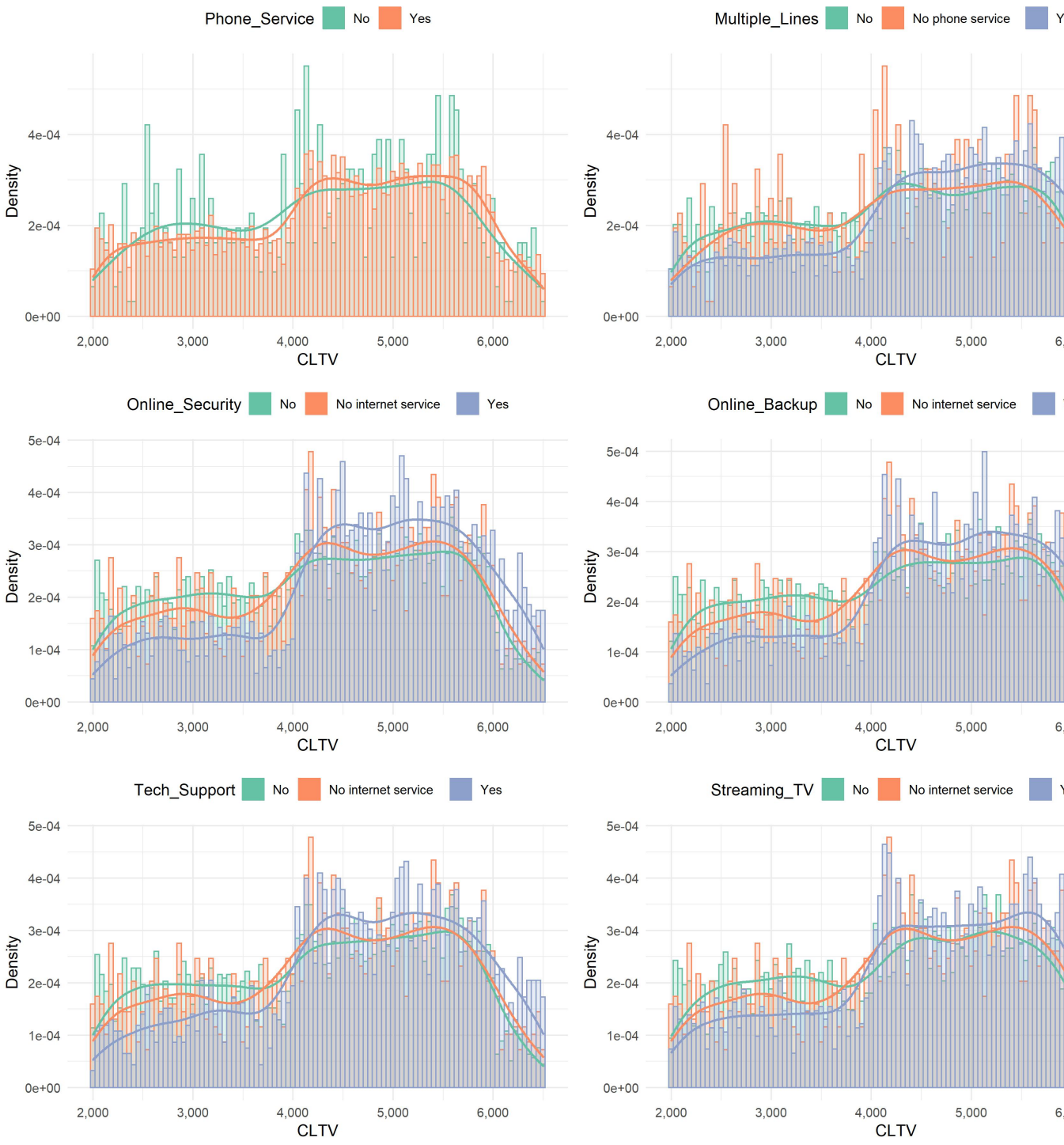

Figure 5.5: Histogram and density plots of customer lifetime value depending on services subscribed

that customers paying by credit card of bank transfer have higher CLV than those paying by e-check of mailed check. Besides, clients enrolled in one-year or two-year contracts have greater *value* to the firm than those who pay on a monthly basis.

Correlation between `CLTV` and explanatory quantitative variables

Plotting the correlation matrix allows to have an overview on the links between the dataset's quantitative variables. The method used is the Pearson correlation coefficient which is defined as follows:

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{5.1}$$

where X and Y are two quantitative random variables, cov the covariance function and σ the standard deviation.

On figure 5.7, non-significant correlations are crossed-out. It can be noticed that `CLTV` has the strongest correlation with `Tenure_Months` (40%), followed by `Total_Charges` (34%) then `Monthly_Charges` (10%).

5.4 Churn, duration and customer *value*

The final step in the data exploration consists in analyzing the relationship between the three target variables: `CLTV`, `Churn_Label` and `Tenure_Months`.

Looking at figure 5.8, customer lifetime value seems to have higher values for retained customers than for churners as the density is more right-oriented. This result makes sense as retained customers may have longer lifetime leading to higgher CLV.

The following histograms are interesting to the extent that the distribution of `Tenure_Months` depends on the churn status. From figure 5.9, 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 last's tenure months distribution is decreasing and looks like a Poisson distribution with an inflation of low values.

Eventually, 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.

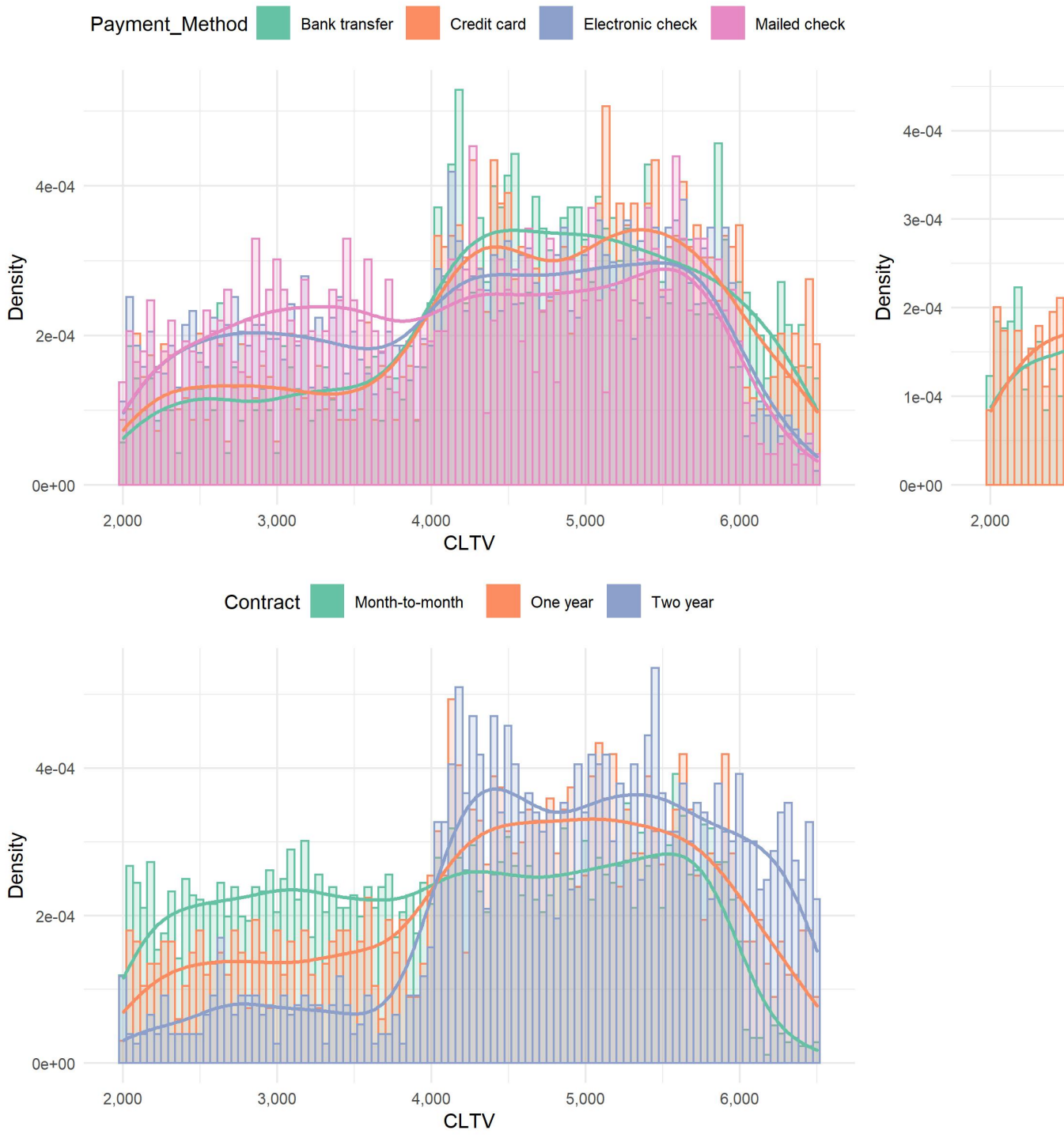

Figure 5.6: Histogram and density plots of customer lifetime value depending on customer account data

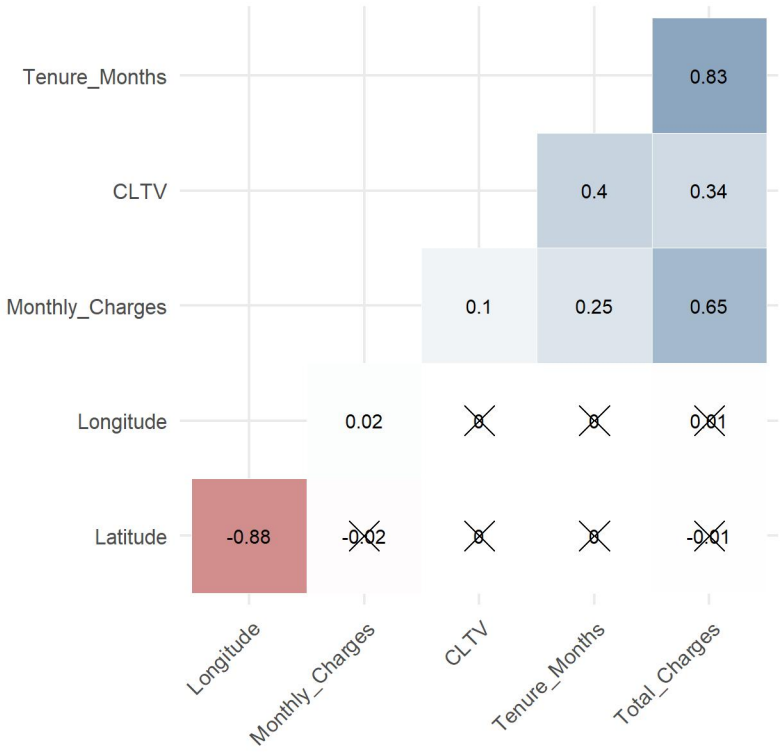

Figure 5.7: Correlation plot

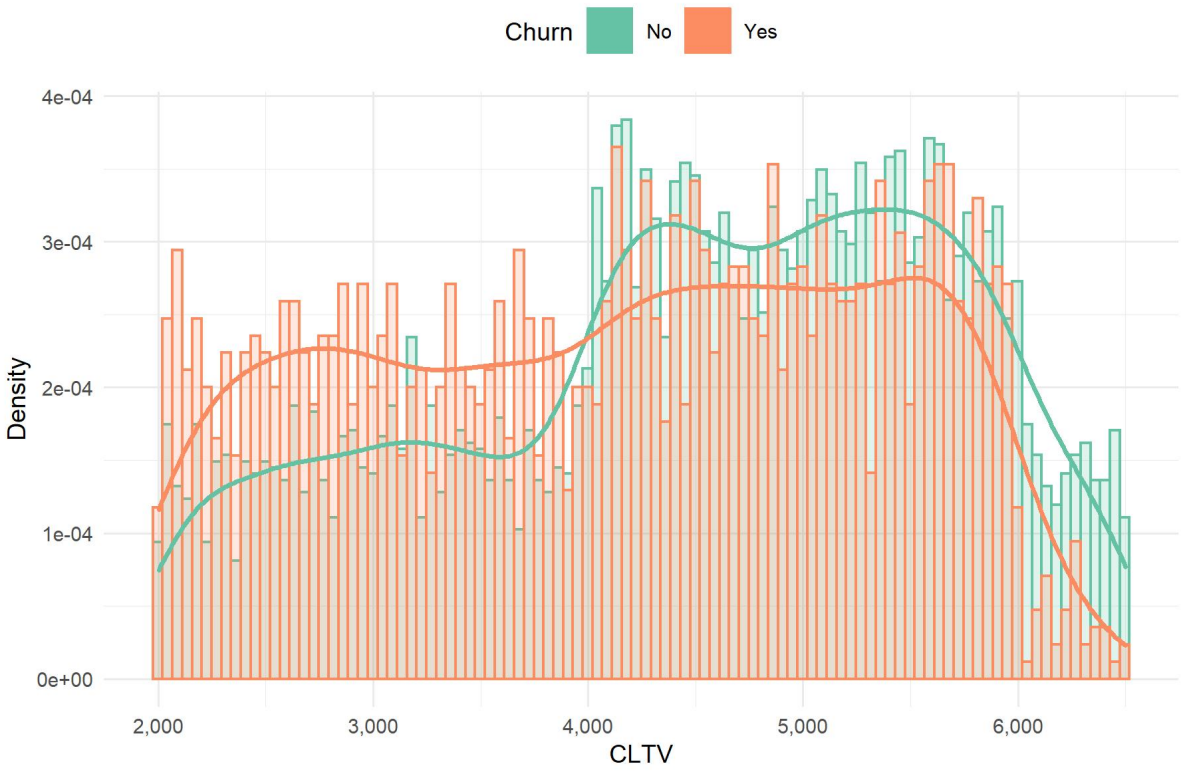

Figure 5.8: Customer lifetime value depending on churn status

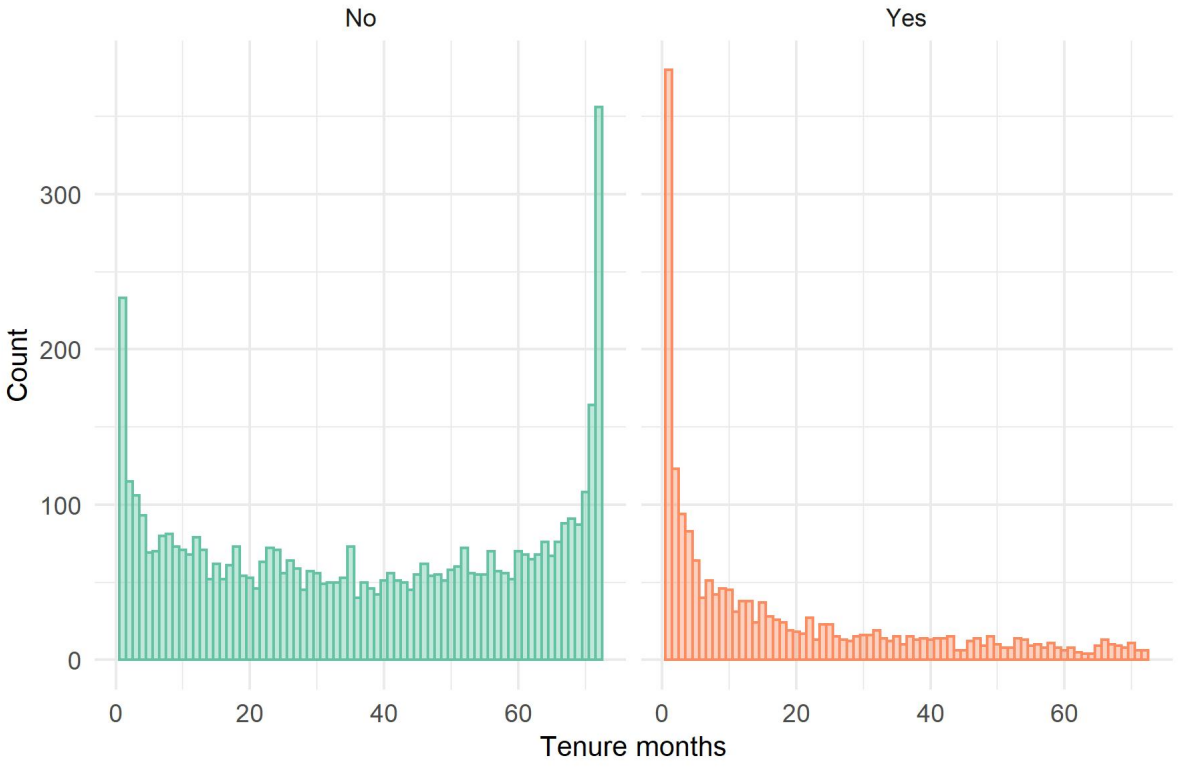

Figure 5.9: Tenure months depending on churn status

Table 5.8: Anova test between CLTV and churn status

	F statistic	Df1	Df2	p-value
Churn_Label	117.57	1	7030	3.5e-27

Appendix

Some proofs of the mathematical concepts used in the study are derived, specifically related to duration analysis.

Hazard function

$$
\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t]/P[T \geq t]}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{(F(t + \Delta t) - F(t))/\Delta t}{S(t)} \\
&= \frac{\mathrm{d}F(t)/\mathrm{d}t}{S(t)} \qquad (5.2) \\
&= \frac{f(t)}{S(t)} \\
&= -\frac{\mathrm{d}S(t)/\mathrm{d}t}{S(t)} \\
\lambda(t) &= \frac{-\mathrm{d}\ln\left(S(t)\right)}{\mathrm{d}t}
\end{aligned}
$$

Link between cumulative hazard and survivor functions

$$\begin{aligned}
&\Lambda(t) = \int_0^t \lambda(s)ds \\
\iff &\Lambda(t) = \int_0^t \frac{f(s)}{S(s)}ds \\
\iff &\Lambda(t) = -\ln\big(S(t)\big) \\
\iff &S(t) = \exp\big(-\Lambda(t)\big)
\end{aligned} \tag{5.3}$$

Contribution to the partial likelihood function in PH models

$$\begin{aligned}
\mathbb{P}\big[T_j = t_j | R(t_j)\big] &= \frac{\mathbb{P}\big[T_j = t_j | T_j \geq t_j\big]}{\sum_{l \in R(t_j)} \mathbb{P}\big[T_l = t_l | T_l \geq t_j\big]} \\
&= \frac{\lambda_j(t_j|\mathrm{x_j}, \beta)}{\sum_{l \in R(t_j)} \lambda_l(t_l|\mathrm{x_l}, \beta)} \\
&= \frac{\lambda_0(t_j, \alpha)\phi(\mathrm{x_j}, \beta)}{\sum_{l \in R(t_j)} \lambda_0(t_j, \alpha)\phi(\mathrm{x_l}, \beta)} \\
\mathbb{P}\big[T_j = t_j | R(t_j)\big] &= \frac{\phi(\mathrm{x_j}, \beta)}{\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)}
\end{aligned} \tag{5.4}$$

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:

$$\begin{aligned}
\mathcal{L}_{p,\ t_j} &= \mathbb{P}\big[T_1 = t_j, \ldots, T_{d_j} = t_j \mid R(t_j)\big] \\
&= \Pi_{m \in D(t_j)}\ \mathbb{P}\big[T_m = t_j | R(t_j)\big] \\
&= \Pi_{m \in D(t_j)}\ \frac{\phi(\mathrm{x_m}, \beta)}{\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)} \\
&= \Pi_{m \in D(t_j)}\ \phi(\mathrm{x_m}, \beta) \times \Pi_{m \in D(t_j)}\ \frac{1}{\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)}
\end{aligned} \tag{5.5}$$

$$\mathcal{L}_{p,\ t_j} = \frac{\Pi_{m \in D(t_j)}\ \phi(\mathrm{x_m}, \beta)}{\Big[\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)\Big]^{d_j}}$$

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

$$\begin{aligned}
\mathcal{L}_p &= \Pi_{j=1}^{k}\ \mathcal{L}_{p,\ t_j} \\
\mathcal{L}_p &= \Pi_{j=1}^{k}\ \frac{\Pi_{m \in D(t_j)}\ \phi(\mathrm{x_m}, \beta)}{\Big[\sum_{l \in R(t_j)} \phi(\mathrm{x_l}, \beta)\Big]^{d_j}}
\end{aligned} \tag{5.6}$$

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.

Bousquet, A. and Concettini, S. (2019). *Econometrie Introduction*. Université de Tours.

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.

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. (2021). *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.

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.