HEC MONTRÉAL

An Assessment of Practical Approximations for the Survival Analysis of Non-Contractual Customers' Lifetime

par

Volodymyr Sakun

Jean-François Plante
HEC Montréal
Codirecteur de recherche

Juliana Schulz
HEC Montréal
Codirectrice de recherche

Sciences de la gestion
(Spécialisation Data Science and Business Analytics)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Résumé

Pour les organisations, l'étude de la durée de vie des consommateurs est très importante. Elles utilisent toutes sortes de modèles pour prédire la fin de leur lien d'affaires, appelé churn. Lorsque les clients n'ont pas de contrat (par exemple en vente au détail), la date de leur départ ne sera pas observable, rendant le problème d'étude du churn encore plus difficile. L'auteur de cette thèse possède une longue expérience pratique en industrie où il a construit des modèles pour étudier le churn de consommateurs non-contractuels. En pratique, des approximations sont utilisées et les hypothèses énoncées sont rarement respectées à la perfection. Dans ce mémoire, nous utilisons les outils d'analyse de survie pour bâtir une stratégie pratique pour le churn non-contractuel, similaire aux solutions approximatives utilisées en pratique. Quelques modèles différents seront présentés. Deux jeux de données et une étude de Monte Carlo serviront à illustrer et comparer les différentes méthodes.

Mots clés : analyse de survie, COX, AFT, churn, CLV

Méthodes de recherche : quantitative, l'analyse des données, Étude de Monte-Carlo

Abstract

For organizations, studying the lifetime of customers and their value is very important. Models are made to predict the end of their business relationship, also called churn. In cases where customers do not have a contract (e.g. retail), the time when a client leaves will not be observed, making the study of churn even more challenging. The author of this thesis has a long experience in the industry where he built models for non-contractual churn. In practice, approximations are made, and assumptions are rarely completely verified. In this thesis, we present a strategy akin to those real-life approximate models, to study the lifetime of non-contractual customers while leveraging survival analysis tools. A few alternative models will be compared. Two datasets as well as a small Monte Carlo study are used to illustrate and assess the proposed methods.

Keywords : survival analysis, COX, AFT, churn, CLV

Research methods : quantitative, data analysis, Monte-Carlo study

Table of contents

List of tables

List of figures

List of abbreviations and acronyms

ANN : Artificial neural network[1]

AUC : Area under ROC curve[2]

CDF: Cumulative distribution function[3]

CI : Concordance index or Harrel's C-index, is a goodness of fit measure for models which produce risk scores[4]

CI_IPCW : CI for right-censored data based on inverse probability of censoring weights; it is an alternative to the CI estimator and does not depend on the distribution of censoring times in the test data

CLV: In marketing, customer lifetime value (CLV or often CLTV), lifetime customer value (LCV), or life-time value (LTV) is a prognostication of the net profit contributed to the whole future relationship with a customer[5].

CvM: Cramér–von Mises criterion[6]

F : Frequency, number of purchases un to day of study

GRU: Gated recurrent unit[7]

IAE : Integrated Absolute Error[8]

IBS : Integrated Brier Score[9]

[1] https://en.wikipedia.org/wiki/Artificial_neural_network (accessed on December 21, 2022)
[2] https://en.wikipedia.org/wiki/Receiver_operating_characteristic (accessed on December 21, 2022)
[3] https://en.wikipedia.org/wiki/Cumulative_distribution_function (accessed on December 21, 2022)
[4] https://statisticaloddsandends.wordpress.com/2019/10/26/what-is-harrells-c-index/ (accessed on December 21, 2022)
[5] https://en.wikipedia.org/wiki/Customer_lifetime_value (accessed on December 21, 2022)
[6] https://en.wikipedia.org/wiki/Cram%C3%A9r%E2%80%93von_Mises_criterion (accessed on December 21, 2022)
[7] https://en.wikipedia.org/wiki/Gated_recurrent_unit (accessed on December 21, 2022)
[8] See Chapter 2 Metrics section

IPI : Inter-purchase interval, duration between two consecutive purchases

ISE : Integrated Square Error[10]

KS : Kolmogorov–Smirnov test11

LSTM: Long short-term memory[12]

M : Monetary value, usually amount spent in $ per one purchase / transaction

MAE : Mean Absolute Error[13]

MLE : Maximum likelihood estimation[14]

PDF: Probability density function[15]

R : Recency, time since last purchase to day of study

RFM : Group of features R, F, M[16]

RNN: Recurrent neural network[17]

TTC : Time to ‘churn’, time interval from date of the last purchase to ‘churn’ event

TTD : Time to ‘death’, time interval from date of the last purchase to ‘death’ event

[9] https://en.wikipedia.org/wiki/Brier_score (accessed on December 21, 2022)
[10] See Chapter 2 Metrics section
[11] https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test (accessed on December 21, 2022)
[12] https://en.wikipedia.org/wiki/Long_short-term_memory (accessed on December 21, 2022)
[13] See Chapter 2 Metrics section
[14] https://en.wikipedia.org/wiki/Maximum_likelihood_estimation (accessed on December 21, 2022)
[15] https://en.wikipedia.org/wiki/Probability_density_function (accessed on December 21, 2022)
[16] https://en.wikipedia.org/wiki/RFM_(market_research) (accessed on December 21, 2022)
[17] https://en.wikipedia.org/wiki/Recurrent_neural_network (accessed on December 21, 2022)

Acknowledgements

I would like to express my deepest appreciation to my supervisors for their invaluable patience and feedback. I am also grateful to all professors from the university for their efforts and amazing lectures. Lastly, I would be remiss in not mentioning my family, especially my wife and my child for their patience and support while I was stressed writing this thesis. Their belief in me has kept my motivation high during this process. Also, I would like to thank my cat for the entertainment and emotional support.

Introduction

Context

The importance of determining a customer's lifetime should not be underestimated. While at first glance it might not look very important; having a tool to accurately model and predict clients lifetime leads to other important metrics widely used in business. For example, customer churn rates (the opposite of retention) are directly related to client lifetimes, which is one of the primary factors that determine the steady-state level of customers that a business will support. The time at which a client's lifetime ends can be defined as a churn event. This is a critical prediction for many types of business since; in general, acquiring new clients is much more expensive than retaining existing ones. Companies invest in clients through promotions, advertisement, discounts, special offers and many other marketing strategies to generate more revenue and minimize losses caused by customer churn. Such strategies also allow to target and retain high-value clients, thereby maximizing profit. Lifetime can be used as an input for modeling a customer's lifetime value (CLV)[18] which is in fact an important concept in business; it provides a tool to estimate the expected revenue generated by the client and encourages a company to shift its focus to the long-term health of their customer relationships. It also represents an upper limit on spending to acquire new customers. If there is a way to identify the most promising clients with the highest expected CLV, a company can focus on targeting them with certain marketing methods.

This thesis is focused on non-contractual businesses. There is a big difference between modeling contractual and non-contractual customer lifetimes. For contractual business, the client's defection is directly observable: the old contract ends and a new contract does not exist, or the current contract is interrupted at some point for some reason with the same result – no more profit from that customer. In this concept, defection refers to the time at which a client becomes inactive. That is, the client stops generating revenue and is considered to be lost forever. For non-contractual business,

[18] See abbreviations for more details

things are more complicated. For every client, any purchase can be the last one or just the next one in the long (or short) sequence of transactions that generates profit for the business. In this case, it is not a trivial task to identify the event of a client's defection. To illustrate this, consider a customer who buys only one product every week; following this pattern, it is expected that the client will make a purchase every seven days. If there are no transactions during the following 21 days, for example, it may be likely that the client has churned. On the other hand, for a customer who makes purchases from time to time and always buys different products, the expected time of the next visit is very different, and difficult to predict. A non-contractual customer can change their buying pattern without informing the company. Therefore, it is inappropriate to use a model designed for a contractual setting in the non-contractual case, where the first difficulty is to analytically define the moment when the client becomes inactive with respect to his previous activity (when they churn).

Objective of this research

The main topic of this research is the modeling of customer lifetime for non-contractual businesses. The author of the thesis has a long experience in the industry where he encountered the challenges previously described. In a real-life setting, multiple assumptions and approximations are made to yield a reasonable answer fast. The goal of this thesis is to reproduce a realistic strategy for modeling non-contractual churn similar to the methods that the author saw and used in real life. We will investigate how well this solution works on publicly available real transaction data and through simulations. Two datasets are considered, each corresponding to two different companies: CDNOW[19] and an online retail[20] company. The CDNOW dataset contains the entire purchase history up to the end of June 1998 of the cohort of 23,570 individuals who made their first-ever purchase at CDNOW in the first quarter of 1997. Each record in this file, 69,659 in total, comprises four fields: the customer's ID, the date of the transaction, the number of CDs purchased, and the dollar value of the transaction. For non-contractual data, it is typical to have such a simple and straightforward list of

[19] http://brucehardie.com/datasets/CDNOW_master.zip (accessed on December 21, 2022)
[20] https://archive.ics.uci.edu/ml/datasets/Online+Retail+II (accessed on December 21, 2022)

transactions. The Online Retail II data set contains a similar list of transactions occurring for a UK-based registered, non-store online retail company between the first of December 2009 and the ninth of December 2011.The company mainly sells unique all-occasion giftware. Many customers of the company are wholesalers. Database contains 33,112 transactions in total, made by 5,878 clients.

Since lifetime (in any units) is a quantitative measure, we will be using regression models to predict this outcome. Predictors must be determined from the transactions data. We will thus create a set of useful features that can be informative for describing customers' behavior and may also contain valuable information for estimating customer lifetime. Some of the features considered here originate from an RFM market analysis[21] and have been used in direct marketing for a number of decades because of their ease of implementation in practice. Interestingly, features that we defined based on intuition are also found in the literature. (Buckinx & Van den Poel, 2005) show that past behavioral variables, more specifically RFM[22] variables and their derivatives, are the best predictors of partial customer defection (when the customers switch some of their purchases to another store). (Fader, Hardie, & Shang, 2010) show that RFM predictors are sufficient statistics for a non-contractual class of CLV models and provide a theoretical justification for their use; it also implies that Recency, Frequency, and Monetary value provide a complete customer summary for CLV prediction.

Since the end of a non-contractual customer's relationship with a company is not explicitly observed, we will have to make certain assumptions based mostly on our practical experience. The most crucial one is at which point of time we can consider that a customer has experienced a total defection: he stopped his business with the company and will never make any new purchase. The consequences of this assumption will be investigated with a Monte Carlo simulation to assess its compatibility with reality and to measure if this assumption is a significant source of error.

While common sense suggests that time-varying covariates should be meaningful to the prediction of lifetime, they cannot easily be leveraged because we would need to

[21] https://en.wikipedia.org/wiki/RFM_(market_research) (accessed on December 21, 2022)
[22] See abbreviations

know the values of those covariates in the future to make predictions. For survival regression we will still consider time-varying covariates but will use the last observed value as a proxy to make predictions. For models that ignore the time-varying variable, we compensate at least slightly for a loss of the behavioral dynamics by using additional predictors that capture features trends (short and long). We believe that leveraging the history of features through their trends will provide valuable (but unfortunately incomplete) dynamic information to the regression model. Lastly, we will construct a regression (non-survival) model based on artificial neural network that uses time-varying covariates as predictors. There are in practice many instances of trying shiny new models for solving problems they were not specifically designed for, hereby ignoring the censoring, for instance. We call this the 'dazzle effect'. We expect that this model will significantly underestimate the remaining survival life, since it will not take into consideration the concept of censorship, but its learning capacity includes learning from sequential data, so it will be interesting to compare the performances of traditional survival regression and ANN[23].

In practice, different algorithms are typically considered. In the context of churn, this includes leveraging survival analysis tools, but also proposing methods that ignore censoring. We will compare results obtained from different algorithms to see their relative performance on the different data including real and simulated datasets. Using the simulated data, we predict the remaining customer lifetime according to the methods explored, and compare the predicted values with the 'true' remaining lifetime obtained by the simulation. While a simulation cannot reflect all aspects of real customers' behavior, especially at the end of their lifetime (before they churn), this method nonetheless provides a rough approximation of a true data-generating process for which we know the truth. It allows us to use additional quantitative metrics such as the mean absolute error between true and predicted remaining life, which is impossible to compute with real non-contractual customer transaction data. We use R to simulate transactions and Python for all other tasks: data preparation, feature engineering, survival analysis and multivariable sequence learning artificial neural network.

[23] See abbreviations

Structure of thesis

This work starts with a literature review of models for the non-contractual setting. Two types of models, probabilistic and non-probabilistic, are proposed in the literature to predict different outcomes such as churn, number of future purchases, or CLV. Particular attention is devoted to exploring various definitions of churn events from the literature.

We describe and explain our own definition of churn in Chapter 1. A cornerstone is the description of our vision of customer's lifetime when the death event is not explicitly observed. Further steps of the proposed methodology are also presented, including the modeling framework (i.e., how we construct a discrete-time survival model) and feature engineering. A special section is devoted to a conceptual explanation of the transition from transaction data into variables and features that can sustain survival analysis. Again, these solutions are based on real-life solutions driven by intuition and common sense, but they may not respect all usual assumptions.

Chapter 2 reviews models and metrics that are used in further chapters, including the description of the survival regression models that are considered. The two real datasets previously described are analyzed in Chapter 3 using different algorithms. We consider commonly used evaluation metrics for survival, notably two versions of the concordance index (CI).

In Chapter 4, we know the 'truth' in the synthetic data produce in the Monte Carlo simulation. As a metric, the mean absolute error (MAE)[24] allows to compare with the 'true' remaining life for censored customers. We also make simplifying assumptions on the CLV to determine a 'true' spending for the censored customers based on the individuals' remaining life and average spending. Customers spending made during their remaining live can be seen as CLV and describes how much revenue the company can expect to have in the future from existing clients (without taking into consideration new arrivals). The interpretation of the role of the features in the models is in line with

[24] See abbreviations

the intuition, and the results shed a new light on the role of assumptions to get good predictions

Finally, our main findings will be summarized in the conclusion.

Literature review

Introduction

The prediction of churn is gaining attention across different fields; both from a business and an academic perspective, and a variety of sophisticated models that predict churn have already been proposed. In a contractual business setting, the definition and observation of churn is relatively simple, because it is directly related to the changes to an existing contract. This can be through the termination of the service unless the customer takes a specific action such as a subscription or membership renewal, or by an active service until the customer actively cancels the contract (Ascarza, Netzer, & Hardie, 2018b). One of the most important challenges with churn in non-contractual business settings is the lack of a good definition for the churn itself. This is especially important since we typically need a working definition of a customers' defection to build a churn prediction model. That churn definition is highly subjective and definitely influences any model and its results.

Churn definitions for non-contractual settings

Non-contractual business settings pose many more challenges. As customers can leave without saying a word to a company or terminating a contract, the loss of customer is not directly observed. Unfortunately, the decision of whether a customer has churned or not is rather subjective and usually relies on heuristic rules, set by the industry officials (Kaya, et al., 2018). One approach to a churn definition for non-contractual business includes two basic parameters: customers' activities and some threshold fixed by certain business (Clemente-Císca, San Matías, & Giner-Bosch, 2014).

(Buckinx & Van den Poel, 2005) do not take into account any monetary condition to identify behaviourally loyal customers. Focusing on frequent customers they define their loyal segment based on two behavioural attributes: the frequency of purchases and the time between their purchases, so customers should satisfy the following two conditions: frequency of purchases is above the average and the ratio of the standard deviation of the inter-purchase time to the mean inter-purchase time is below average.

The first criterion serves as an indication of a customer's loyalty and potential profitability (Wu & Chen, 2000). The second ensures that the times between customer visits are regular. So, if one of the aforementioned conditions is not fulfilled, (Buckinx & Van den Poel, 2005) classify a customer as partially defecting. (Clemente-Císca, San Matías, & Giner-Bosch, 2014) use an approach similar to (Buckinx & Van den Poel, 2005) where a customer is considered as a churner if they changed a predefined status from loyal to non-loyal, whereas loyal customers are those who shop frequently and have a regular buying pattern.

(Karnstedt, et al., 2010) propose several definitions of churn, notably global, individual, and gradient. Global churn considers a customer as a churner if their average activity level within a certain time window is lower than a fraction of the average activity level of the population in the same time window. Their definition of individual churn follows a similar principle, but the individual average activity is compared to the average level of activity of the same customer in a prior time window. The idea of gradient churn is then similar to individual churn, but the change through time is measured as a ratio akin to a derivative measuring the magnitude of the rate of change in the individual's level of activity.

Working with different definitions of churn is frequent as it may answer different business questions. In the mobile gaming industry, (Perisic, Jung, & Pahor, 2022) use definitions of churn for no activity (absence churn) as well as a decline in engaging activities (starting absence churn). They also define their own version of the gradient churn of (Karnstedt, et al., 2010) to detect changes in the behaviour of customers.

For the online gambling industry (Coussement & De Bock, 2013) assume gamblers are churners when they have not placed a bet over a period of four month. The model proposed by (Jahromi, Stakhovych, & Ewing, Managing B2B customer churn, retention and profitability, 2014) use half a year as a unit of measurement and analyses two consecutive periods: define churn as being inactive in the second half of the year (prediction period) while being active in the first half of the year (calibration period). A clustering approach can be found in (Jahromi, Sepehri, Teimourpour, & Choobdar,

2010) who model churn in non contractual setting in the case of telecommunication service providers. Customers' population is divided into groups (clusters) according to their intensity of cell phone usage. Clients are considered as churners if the period of their inactivity is greater than the mean value that corresponds to the cluster to which they belong.

(Bayrak, Guven, Bahadır, & Yalcınkaya, Comparative Methods for Personalized Customer Churn Prediction with Sequential Data, 2022) use sequences of times of transactions to determine customers' churn. They focus on average-order-day frequency (the average of the time differences between orders in day) calculated individually for each regular customer, as well as a similar average for the entire population. They include only customers with a minimum of five transactions. Interestingly, they follow customers after they churned, and study the individual buying patterns – a sequence of churn / no churn status for each customer. Their analyses focus on a short-term definition of churn and the assumption that there will be multiple churns for a typical customer. They use those patterns as an input feature in sequential neural networks algorithms.

(Glady, Baesens, & Croux, 2009) use a modified Pareto/NBD model to predict the future number of transactions and CLV as a sub-model to define churn: a churner is defined as someone with a customer lifetime value decreasing over time.

Survival for time-to-churn prediction

To paraphrase (Liu, 2012), the practice of survival analysis is the use of reason to describe, measure, and analyze features of events for making predictions about not only survival but also 'time-to–event processes' – the length of time until the change of status or the occurrence of an event – such as from living to dead, from single to married, or from healthy to sick. In medical research, scientists apply survival analysis to compare the risk of death or recovery from disease between or among population groups receiving different medications or treatments

In survival analysis specific methods are required because we rarely observe the event of interest for all participants but the 'censored' information about the not-yet-observed events must be taken into consideration. As (Liu, 2012) mentions, methodologically, censoring is defined as the loss of observation on the lifetime variable of interest in the process of an investigation. In survival data, censoring frequently occurs for many reasons. In a clinical trial on the effectiveness of a new medical treatment for disease, for example, patients may be lost to follow - up due to migration or health problems. In a longitudinal observational survey, some baseline respondents may lose interest in participating in subsequent investigations because some of the questions in a previous questionnaire are considered too sensitive.

A key publication in survival analysis, (Cox, 1972) proposes semi-parametric regression for survival data. His trick is to have a nonparametric base hazard function multiplied by a factor determined by regression coefficients. A parametric alternative to Cox regression is the Accelerated failure time models (Wei, L.J.: The Accelerated Failure Time Mode: A Useful Alternative To the Cox Regression Model in Survival Analysis, Statistics in Medicine,11, 1992, 1871-1879.). Generally speaking it determines the way in which the explanatory variables influence the survival time or in another words, how much the covariates accelerate or decelerate the lifetime of an individual comparing to the baseline. In recent publications non-traditional applications of survival analysis are found in fields quite far from medicine. For example, Cox model has been used to predict credit card default (Djeundje & Crook, 2019). Also, Cox proportional hazards model is used by (Li, Li, & Li, 2019) to study credit card default problems. An extended Cox proportional hazards model is applied by (Hu, Chen, & Chen, 2021) to discover the impact of different factors that influence customer churn in the car sharing industry. In the insurance field, Cox regression model is proposed to analyze clients' repurchasing behavior with lifestyle segmentation (Ansell, Harrison, & Archibald, 2007). Research by (Chen, Zhang, Zhao, & Xu, 2022) has been conducted on car insurance renewal problems. Their paper proposes the Cox model with variable penalties to predict a customers' churn and distinguish regular clients for vehicle insurance.

CLV

Most models that study non-contractual customers' life attempt to predict either their number of future purchases followed by estimation of future spending or CLV directly. (Wong, 2011) models customer time to churn by Cox regression, but for contractual settings where the 'death' event is explicitly observed. However, his work focused mostly on the attempt to identify customer segments that are sensitive to churning behaviors. Some papers related to non-contractual settings propose probability models. An interesting overview of a class of parsimonious models is has been done by (Fader, Hardie, & Shang, 2010). They cover both: non-contractual (customers 'death' is not observed) and contractual setting. Those models assume that customers buying behavior follow some probability distribution. Models use this assumption to estimate the probability of customers 'death' or number of future purchases (buy till you die). One of the first probabilistic models for non-contractual setting that tries to predict customer's future purchases is Pareto/NDB proposed by (Schmittlein, Morrison, & Colombo, Counting your customers: Who are they and what will they do next?, 1987). They assume that purchases of active customers are characterized by a Poisson distribution and the time between purchases can be represented by an exponential distribution with a certain rate parameter that corresponds to customers mean purchase rate. Each customer has their own rate which is assumed to come from a gamma distribution, resulting in the negative binomial distribution (NBD) model for repeating purchases at the population level (Morrison & Schmittlein, 1988). One development of Pareto/NBD model is proposed by (Fader, Hardie, & Shang, 2010). They call it beta-geometric/beta-Bernoulli (BG/BB) model that captures customers' purchases while clients are active and the time until each customer "dies". Another modified Pareto/NDB approach was proposed by (Glady, Baesens, & Croux, 2009) to predict the future number of transactions and customer lifetime value simultaneously. The churning customer is defined as someone who's CLV is decreasing. Generalizations of the Pareto/NBD can be found, for example in (Jerath, Fader, & Hardie, 2011). Non-homogeneous hidden Markov proposed by (Netzer, Lattin, & Srinivasan, 2008) captures the dynamics of customer relationships incorporating the effect of the sequence of customer-company encounters and buying behaviour.

In addition, survival analysis allows having both lifetime and probability of survival, so it could in fact be used to solve both problems since the ‘hazard function’ can be viewed as the probability that a customer will churn away (Wong, 2011).

Multiple strategies have been developed to study the life of customers. Some look at specific definitions of churn, others try to predict CLV. Since churn is an ‘event’ of interest, using survival analysis would seem natural, yet standard tools are not typically leveraged. The approach that we assess in this thesis focuses on customers’ lifetime and makes use of different survival analysis methods.

Chapter 1
From transactions to survival

Churn refers to an end-of-life event in the sense that a client eventually ceases to purchase from the company. The data produced by non-contractual clients are a sequence of transactions, which may be transformed into survival data. In this chapter, we present relevant definitions and strategies to transform those transactions into relevant times for survival analysis.

Why survival?

The main goal of this work is to estimate customers' lifetime. Survival analysis is a branch of statistics for analyzing the duration of time until one event occurs, such as death in biological organisms and failure in mechanical systems. Among the others, the two main goals of survival analysis are to determine the effect of covariates on survival time and to predict the moment of failure or individuals' death. The inability to observe the death event in a non-contractual setting involves that we cannot know the real time of churn. Our setting contains time-varying covariates, which creates additional complications for modeling. To predict duration, we would need to know the covariates values beyond the observed times. However, if we knew that, we would also know if the subject was still alive or not! COX time varying and discrete-time proportional odds models allow to compute hazard rate of subjects at known observations, the baseline cumulative hazard rate and baseline survival function. However, it is not trivial to estimate expected duration (lifetime) for any individual. Accelerated failure time survival models have the ability to extrapolate lifetime estimation, but the problem lies in feature engineering: they are time-varying. One simple possible solution is to use features that correspond only to last known states regardless of whether customer is active or churned. However, this simplified model might not be able to learn from customers purchasing history and be unable to extract useful information from changing clients purchasing activities during the study period. For example, one customer had very regular buying pattern during some period (purchased something every week for example). Then he started to buy in irregular manner (one month delay, than few regular

transactions again). Finally, he leaves. Intuitively, he started to churn when he changed his buying behavior. Therefore, a feature that describes regularity (clumpiness) had values close to zero at the beginning, but by the time of churning event, value got closer to 1 (0.5 for example). If we look only at the final state, we will see that this client has an irregular buying pattern, but we cannot know that he purchased regularly before. Another example is for the customer whose mean daily spending was $10 at the beginning for some time, than he started to buy less and less ($2 daily), and finally left. The inability to provide changes of any predictor in time might be compensated by using their linear trends.

Durations and censorship

A typical survival outcome is summarized by two pieces of information: a time and an indicator saying if that time corresponds to an event or to censoring. In this work, we consider right-censoring, i.e. the situation where we either know the time of an event, or a lower bound for that time, as this is the typical type of data for a churn study.

In a non-contractual setting the death and churn events are not observed. To define the survival outcome, we must first decide on a working definition for those important events. The mathematical definition of those events is presented in the next chapter, but given our constraints, they are defined from the sequences of purchase times. **Figure 1** shows three different customers with their purchase history within the time frame of a churn study period. Client 1 has his ‘death’ event within the study interval, therefore his lifetime will be considered observed, he is not censored. The time to event is defined as the time from ‘birth’ to ‘churn’ (see details in defining churn section). The ‘death’ events of Clients 2 and 3 lie outside the study interval. Even if the ‘churn’ event of Client 2 happens before the end of the study, we do not yet know about his ‘death’; so both Client 2 and 3 are censored. Their times of censoring are equal to the duration from ‘birth’ event to the last time of activity, which corresponds to the customers’ last transaction.

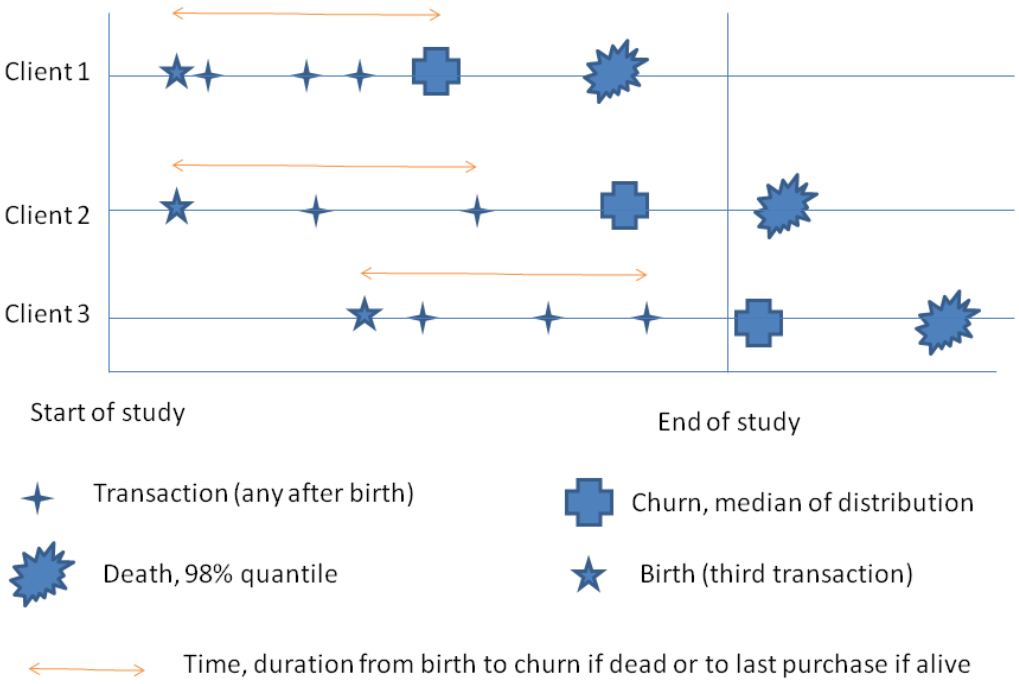

Figure 1. Visual explanation of the process of defining 'event' shown on three different customers.

Definition of churn

Before defining churn for non-contractual business, two questions arise: when does a client become inactive (dead) and can we be sure that the death is true? In other words, when can we truly ascertain whether a customer has stopped his business with the company and left forever? Such unobserved states can unfortunately not be determined with certainty. The literature review in the previous section describes a variety of strategies when it comes to defining customer churn in a non-contractual setting. All have their advantages but need to be adapted to the type of business it is applied to. In practice, the definitions of churn will be inspired by the literature but represent the particularities of a given company. This work proposes a definition of churn based on customers' individual buying patterns that could likely be developed by an organization, and that is in fact very similar to definitions that the author has used in the industry.

A very common definition of churn is to fix a set time (say four months) after which the client is considered ‘dead’. This definition is found, e.g. in (Coussement & De Bock, 2013). In some businesses, clients will have heterogeneous buying frequencies. For customers that buy every day, a death event after four months lies way too far from the unobserved time when they quit. From the other side, clients that make purchases once per year experience ‘death’ after each transaction and are wrongly labeled as ‘dead’ after each purchase. Defining a longer (or shorter) period of inactivity for the entire population of customers will only lead to an aggravated problem at one end of that spectrum. A definition of churn that uses a common time threshold for all customers would be appropriate when customer patterns are homogeneous, but this is not the case in the data that we consider. From experience, we also observed that clients’ buying patterns can vary a lot for different companies depending on their occupation, product assortment, size and many other characteristics. In one company that I worked for, a small pool of customers were buying every day products and services, and hence inactivity for only few days meant they turned over. Even within the same company, another category of customers were buying from time to time with inter-purchase intervals that varied from days to years. Seasonality patterns may also be a reality: every summer some clients buy a lot but they are inactive for the rest of the year.

We prefer adaptive thresholds to define ‘death’, an idea that is found in (Buckinx & Van den Poel, 2005) although they predict customer partial defection rather than their lifetime. For our adaptive proposal, each customer has their own individual time to ‘death’ (TTD), the time interval from the last transaction to the day when a customer is considered to be lost. The personalization of TTD leverages the individual distribution of purchase history. The inter-purchase interval (IPI) times are of particular interest. The IPIs represent the time between two consecutive transactions. To have a sufficient purchase history, we limit the pool of customers to those who completed their third purchase, hence making that date of purchase their ‘birth’ event for the purpose of the churn study. To represent each customer’s IPI distribution, we fit parametric models to their data using Maximum Likelihood Estimation (MLE). If we knew that all clients are still alive at the end of the period, we should have included the censored value of the last purchase time to the end of observations. Since we do not know their status, we ignore

that last censored time, but acknowledge that this approximation could bias in the estimation. Each customer buying pattern therefore gets summarized by their estimate of the parameters for their IPI distribution. With a Gamma model, for instance, each customer will have a shape and a scale parameter of his own.

To illustrate, let us consider the following list of transaction dates for one typical customer in the CDNOW dataset: 1997-01-01, 1997-01-15, 1997-02-05, 1997-04-08, 1997-05-27, 1997-06-17, 1997-07-22, 1997-09-16, 1997-12-09, 1997-12-30, 1998-02-18. These dates yield the corresponding sequence of IPI in days: 14, 21, 63, 49, 21, 35, 56, 84, 21, 50. They are referred as the sample for this customer. The distribution of IPI may vary from a business to another, and we thus consider multiple popular distributions as potential parametric models, namely: Normal, Lognormal, Fisk, Weibull, Exponential, and Gamma. For this specific sample, we find the following parameter estimates for these distributions:

- Normal: mean = 41.4, std = 21.6019,
- Lognormal: stdLog = 0.565, mean = 35.6273, meanLog = 3.5731,
- Fisk: shape = 2.8731, median = 36.2365,
- Exponential: scale = 41.4, rate = 0.0242,
- Gamma: shape = 3.4874, scale = 11.8712, rate = 0.0842,
- Weibull: shape = 2.0546, scale = 46.9612.

These parameters yield probability density functions (PDF) that are shown on **Figure 2** along with their median and 98% quantiles.

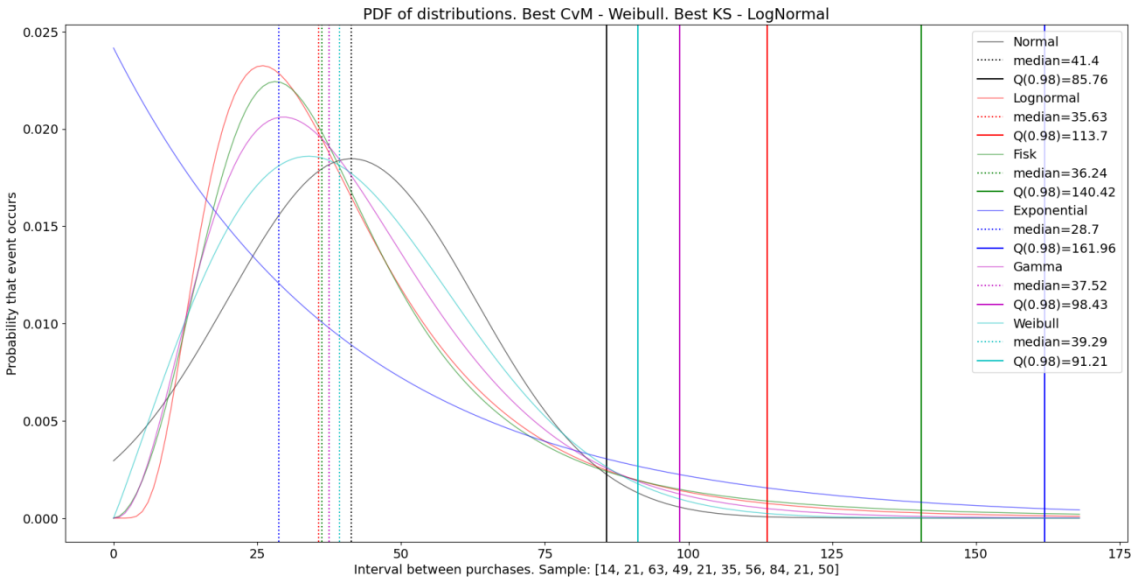

Figure 2. PDF of Normal, Lognormal, Fisk, Exponential, Gamma and Weibull distributions with parameters obtained by MLE on sequence of inter-purchase intervals from the CDNOW dataset, namely {14, 21, 63, 49, 21, 35, 56, 84, 21, 50 }. Dotted vertical lines with corresponding colors show median of each distribution. Solid vertical lines mark 98% quantile of corresponding distribution.

For a given churn analysis, we select one of the models using goodness-of-fit tests, namely Kolmogorov-Smirnov and Cramér-von-Mises. While letting each individual have their own best fitting model would seem to provide additional flexibility, such a strategy would also be more unstable for customers with a very short buying history. We therefore retain the parametric model that provides the best average fit for all customers based on a majority vote strategy.

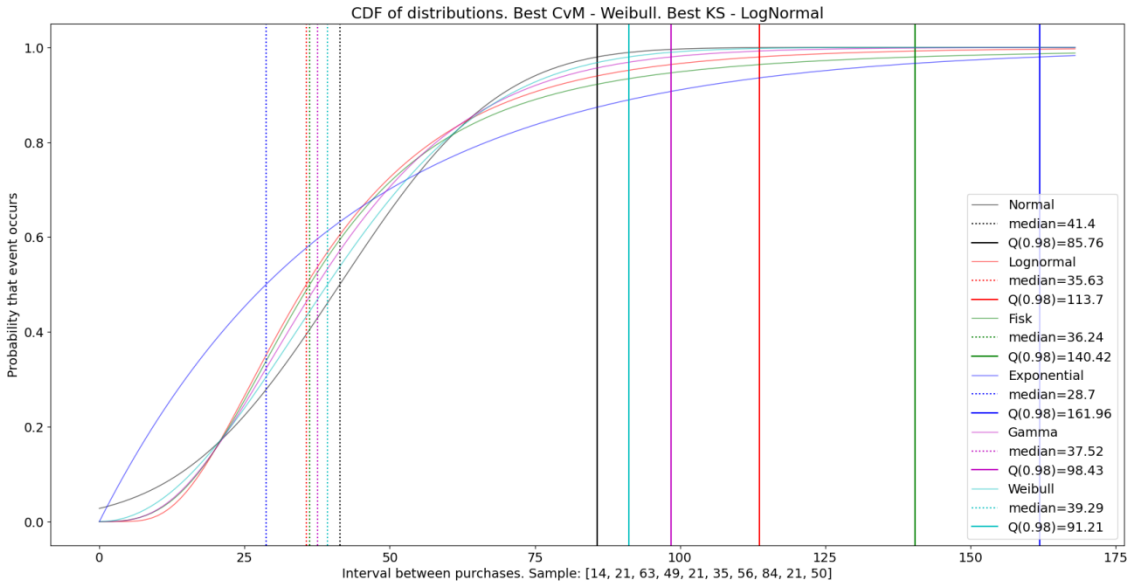

Figure 3. CDF of Normal, Lognormal, Fisk, Exponential, Gamma and Weibull distributions with parameters obtained by MLE on sequence of inter-purchase intervals {14, 21, 63, 49, 21, 35, 56, 84, 21, 50 }. Dotted vertical lines with corresponding colors show median of each distribution. Solid vertical lines mark 98% quantile of corresponding distribution.

Our definitions of churn and death will involve quantiles that are more naturally visible from the cumulative distribution function (CDF). **Figure 3** shows the estimated CDF for the same CDNOW client considered earlier. All CDFs (except exponential) look similar. While the solid vertical lines show the 98% quantile of the distributions, the dotted vertical lines represent the median of each CDF. Remembering that the CDF of a random variable X is the function given by: $F_X(x) = P(X \leq x)$ where the right-hand side represents the probability that X takes values less than or equal to x. This means that for Weibull distribution, for example, estimates that the probability of a customer making a purchase within 91.21 days of its last purchase is 98% as may be seen from the plot. Or having parameters of Weibull distribution determined by MLE: shape: 2.0546, scale: 46.9612 and remembering that for $x > 0$ $\mathrm{CDF} = 1 - \mathrm{e}^{-(\frac{\mathrm{x}}{\mathrm{scale}})^{\mathrm{shape}}}$ at $x = 91.21$ we have $CDF = 0.98$. In other words, if this customer is still alive there is only a 5% chance that he does not buy something during 91.21 days. We would like to translate this probability into the probability of being alive given the time since the last transaction, but that would involve using the Bayes formula which requires the a priori knowledge on the probability of being alive until that time. This situation is akin to tests of hypotheses where we need to make a decision on the rejection of H_0 but have no easy

way to determine the probability of it being true. We propose a similar solution: when there is a low enough probability that an alive client waits so long to do a purchase, we reject the 'hypothesis' that he is still alive. With this definition of the 'death' event, after a given purchase there is approximately one chance out of fifty that a customer will experience a false 'death' event because of the length of time between two consecutive transactions exceeding the TTD interval.

So TTD is the duration in days that corresponds to the value of percent-point function (quantile function), constructed from some distribution with certain parameters, obtained by MLE estimator, equal to 0.98. For example for exponential distribution, 98% quantile can be calculated by formula: $-\frac{\ln(1-p)}{\lambda}$, where lambda is rate, so a customer described in the example has TTD = 161.96 days. We should notice that 98% quantile has been chosen intuitively as a compromise between certainty of customers' death and early churn alert. If for example, we take exponential distribution and 99% quantile, false 'death' error will decrease to 1% but alert caused by customers absence will be raised later (in 190.65 days according to exponential distribution for example).

Death times are used to decide that a client is already gone, but their actual churn time will most often happen before. Intuition tells that from a company's point of view, the 'churn' event could be considered to occur as early as the day (or the next day) of the last transaction after which client will not be making any purchases. Following this logic, the customer's lifetime would be the time interval from first to last purchase and could be directly observed for clients who are considered 'dead'. The lifetime of customers considered still alive is unknown (censored). This logic is in line with churn prevention: proactive retention actions require identifying churners before they leave forever, as early as possible. So, for the client taken as an example, 91.21 days after the last transaction (based on the Weibull distribution) is likely too late. Defining 'churn' event too close to the last transaction may also cause issues for any algorithm that would use time-varying features, as those vary substantially after each transaction due to their nature. Details of the time-varying features are described in the next section).

Therefore, we define a ‘churn’ even, which only occurs for customers whose ‘death’ events were observed. The time to churn (TTC) is the duration in days that corresponds to the median IPI based on the MLE estimator previously discussed. Customer’s lifetime is the duration in days from customer’s ‘birth’ to his ‘churn’. This is equal to the duration between third and last purchases plus TTC.

The median of a Weibull distribution is given by $\lambda(\ln(2))^{\frac{1}{k}}$. For our example, k is shape = 2.0546 and lambda is scale = 46.9612, yielding a median of 39.29 days, the TTC. Note that the other aforementioned five distributions would lead to different yet similar TTCs ranging from 28.7 to 41.4 days. Note that TTD varies more since it is further in the tail. For this particular sample, it ranges from 85.76 to 161.96 days. The choice of distribution might have a significant impact on the results, hence our approach to use goodness-of-fit tests to select wisely.

Both, ‘churn’ and ‘death’ events are obtained from different quantiles (50%, 98%) of the same distribution, but have different purposes. ‘Death’ event specifies whether customers are still alive or if they left forever, but the ‘churn’ event serves as a marker of when the customer actually left. In the context of survival regression, when a client meets the definition of ‘death’, its observed time of death will be defined based on TTC. Otherwise, the customers are assumed alive, and the time from birth to now is censored. Details on survival regression are explained in Chapter 2.

Discrete-time framework

Our model uses discrete time framework within which variables (features) occurring at distinct, separate "points in time" and being unchanged throughout each certain time period (time step). Thus features’ values jump from one value to another as time moves from one time period to the next. The reason we use this framework is that our features cannot be measured directly and must be calculated; this process requires certain time. Time periods where features have values we call ‘States’. **Figure 4** shows the discrete-time framework for CDNOW dataset. Transactions period starts at 1997-01-01 and ends at 1998-06-30. Study period starts at 1997-04-05 and ends at 1998-06-27, therefore, State 0 is at 1997-04-05 and the last state is at 1998-06-27, number of states

equals to 65 (N=64) and time step is equal to 7 days. Dataset provides us a population of 7206 customers that satisfy the eligibility criteria described in previous paragraphs. 5607 of them are censored and 1599 are dead.

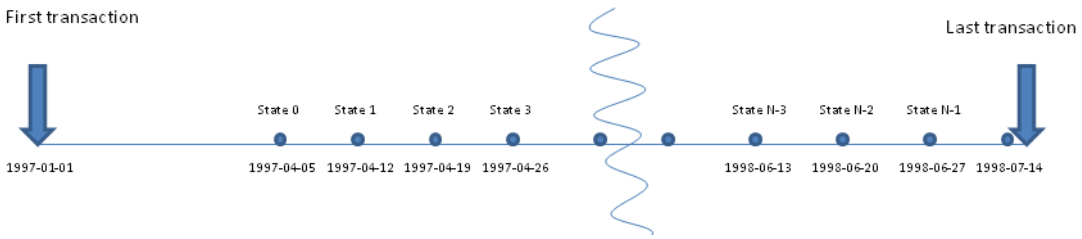

Figure 4. Visual demonstration of discrete-time framework created from CDNOW dataset.

Every state has a subset of customers: each of them must have his 'birth' event somewhere between the first transaction date and a state date. From states' perspective the only knows customer's activity is the one that happened before the states' date. This rule is mandatory for all clients' descriptors except status. Generally speaking, all customers' descriptors (except status) are calculated from states' perspective: future clients' activity (after States date) remains unknown. Customers' status is an exception from this rule; it might be either 'dead' or 'censored'. Customer is 'dead' at State if his TTD (see defining churn section for details) lies within transactions interval ('death' happen before the end of transactions period), and 'censored' otherwise. Every customer at each state has set of descriptors that represent variable, event or are used to calculate features.

Features

Features are the subset of descriptors that are used predictors in machine learning models and can possibly describe quantitatively customers' behavior. Some features may be time-varying. Our data has limitations: we only have purchase history, hence all features may only be engineered from those transactions that contain customer id, purchase date and amount. While defining both the outcome and the predictors from a same source of data may raise concerns with endogeneity, these operations tend to be

frequent in practice. We are also looking for some predictive models, rather than inferential conclusions, a more forgiving endeavour as long as the final black box produces useful predictions. Most predictors we define are similar to the ones used in (Buckinx & Van den Poel, 2005) where a very good summary of many behavioural independent variables is described and supported by past research.

Recency is the time interval from 'last purchase' to the day of study. Customers who recently made purchases are more likely to be active than customers who shopped a long time ago (Wu & Chen, 2000). Previous studies find that the lower the value of recency, the higher the probability that a customer is still loyal. In a non-contractual setting this can be the most important variable to indicate an active or inactive relationship (Reinartz & Kumar, 2002).

Frequency is the number of purchases the customer made since birth (regardless of money spent and number of items bought). The customer's frequency of purchases may be predictive for their future behavior (Schmittlein & Peterson, 1994) because it is positively related to customers' expected future use (Lemon, White, & Wine, 2002). The probability that a customer is alive may be measured by the number of purchases (Reinartz & Kumar, 2002).

Loyalty is the time interval between customer's 'birth' and day of study. The extent to which a customer is able to identify himself with a company is positively related to the period he is willing to continue this relationship (Bhattacharya, 1998). This expectation is confirmed in (Anderson & Weitz, 1989) and indicated that the length of the relationship is positively associated to the perceived future stability of the relationship.

Clumpiness – measure of regularity of intervals between visits or individual-level entropy measure described in the article (Zhang, Bradlow, & Small, 2014) in details and equal to $1 + \frac{\sum_{i=1}^{n+1} \log(x_i) x_i}{\log(n+1)}$ where x_i is ith value in sequence and n is sequence length where sequence is customers IPI. It has range [0 , 1] where values close to zero

correspond to regular buying pattern (all values in sequence are almost identical) and a value closer to 1 represents irregular buying patterns.

Monetary predictors represent the amount of money someone has spent at a company. The monetary value of each customer's past purchase behavior tends to be effective in predicting purchase patterns (Schmittlein & Peterson, 1994) and is used in the literature to determine future patterns. We incorporate three monetary features:

MoneySum – sum of customers' spending.

MoneyDaily – daily average spending which is MoneySum divided by time interval from customer's 'birth state' to date if study.

MoneyMedian – median customers spending.

Poisson features

One assumption can be made that intervals between purchases follow Poisson distribution with certain constant mean rate, individual for every client. However, mean rate can vary from state to state and depends on previous intervals. Equation adapted to rate is:

$$\mathrm{P(k\ events\ in\ time\ t)} = \frac{\mathrm{rt^k e^{-rt}}}{\mathrm{k!}}$$

Probability of one event occurring during [0..recency] period

$$\mathrm{P(1\ or\ more\ events\ in\ time\ t)} = 1 - \mathrm{P(0\ events\ in\ time\ t)} = 1 - \mathrm{e^{-rt}}$$

Where current rate $\mathrm{r} = \frac{\mathrm{n}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{x_i}}$

Where x_i is the i_{th} interval between transactions, n is number of repeating transactions (number of intervals between transactions), t is recency at current state. Also, having current rate, a time interval to the instance where probability of occurring 1 or more events is equal to p can be found as $\mathrm{t} = -\frac{\log(1-\mathrm{p})}{\mathrm{r}}$

Another two features will be used to create predictors for our modeling:

pEventPoisson – probability that client makes transaction at current state

ratePoisson – rate described above

Ranking features (ranks)

Deciles (10 quantiles) are used to divide observations in a sample into continuous intervals with equal probabilities based on method 7 of (Hyndman & Fan, 1996). It is necessary since features like MoneySum can have long range but shifted towards one side distribution. For example mean value for population could be 100 but for few customers cumulative spending can be 100000. There is an issue that appears while number of dead clients is accumulating. For all dead customers, recency is increasing constantly and does not drop since left clients do not make purchases anymore. Therefore, if deciles are calculated including both dead and alive customers, the mean value for each decile shifts towards the higher numbers (in case of recency), which will lead to distortion of real picture that should describe customers' standing relative to other active customers. Having this in mind, for every state, cut points for deciles are calculated using only subset of active clients. Then, for all customers, corresponding features are used to find appropriate decile. On machine learning language, model fits active customers' subset, but predicts values for entire population, that belongs to specified state. As a result, any ranking feature gets integer value from 1 to 10. Highest number 10 corresponds to 'best' customer (the maximum score represents the preferred behavior). This is straightforward for most of features such as frequency, loyalty, and monetary features. Higher value represents better client. However, for recency and clumpiness, order should be reversed, so most recently visited customer has highest rank 10. Other clustering methods can be applied to assign ranks (labels) for customer's features, but usually it takes more computational time and those that have been tried (k-means and agglomerative clustering) did not show any significant improvement in final lifetime predictions. The following ranks will be used in later modeling:

r10_R: recency rank

r10_F: frequency rank

r10_M: moneyDaily rank

r10_L: loyalty rank

It should be mentioned that ranks correspond to their sources (features described above) and therefore are calculated for each state separately.

Complex ranking features

Composite ranks are composition of rank features $\sum_{i=1}^{n} w_i * r_i$ (m.1) or $\prod_{i=1}^{n} w_i * r_i$ (m.2) where n – number of single features, w_i - weight of each feature, r_i – feature itself. Each type of composite feature has the same scale (integer [1 , 10]) as source ranking feature. We will be using (m.2) model it will include interactions between two or more predictors in the regression model.

r10_RF – interaction between recency and frequency.

r10_FM – product of frequency and MoneyDaily ranks. High value will indicate those clients who made many repeating transactions and spent lots of money.

r10_RFM – product of recency, frequency and MoneyDaily. Probably the most common indicator in a way to use data based on existing customer behavior to predict how a new customer is likely to act in the future (RFM analysis).

R10_RFML – product of recency, frequency, MoneyDaily and loyalty.

Trends

Generally speaking, for time series sequences, trend estimation can be used to make and justify statements about tendencies in the data, by relating the measurements to the times at which they occurred and can be used to describe the behavior of the observed data, without explaining it. Trend features are of two types: long and short. Long trends describe global tendency of change of customer's behavior. Linear trends were obtained by simple linear regression and the slope value is used as a feature.

Positive trend value signifies general improvement of customers standing in population among active clients. Short trends are designed to get recent changes. Only most recent states (counting from present) are included in linear regression model that calculates the slope. For this model, data from the most recent 10 time steps were used to find this feature. Since calculating trends for every customer at each state require some time, trends will be prepared only for subset of features described above.

On **Figure 5** and **Figure 6**, two randomly selected customers from CDNOW dataset are presented to illustrate how some predictors change in our discrete-time framework. **Figure 5** shows active customer, **Figure 6** – the churner. There is a timeframe on the x axis, blue vertical dashed lines show transactions that clients made. Top subplot of each two plots shows rank r10_RFM (green) scaled to range from 0 to 1, Clumpiness (C in plots legends in blue) and pPoisson (red). On bottom subplot the corresponding short and long trends are shown (same colors, short trends are dashed curves, long trends - solid). Visual inspection might give an impression that for active customer ranks go up and long trends are positive. For churner the situation is different: ranks go down and long trends have either small positive or negative values. From the other side, short trends could reflect a dynamics in buying pattern. We hope, that those predictors as well as the others described in this section would be informative for machine learning models we are going to implement to predict customers lifetime.

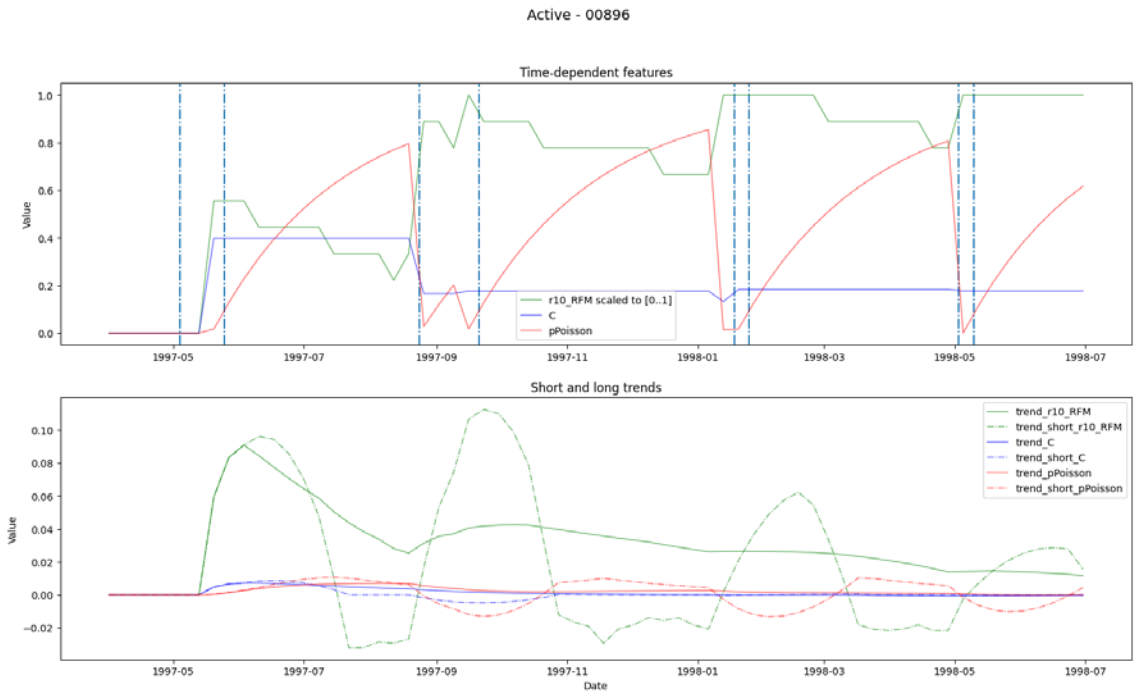

Figure 5. Randomly selected active customer from CDNOW dataset, who made his last purchase not very far from the end of transactions data. Upper subplot shows how features r10_RFM, C and pPoisson change in time. Vertical dashed lines correspond to time instances when customer made purchases. Lower subplot displays short and long trends of corresponding features from top subplot.

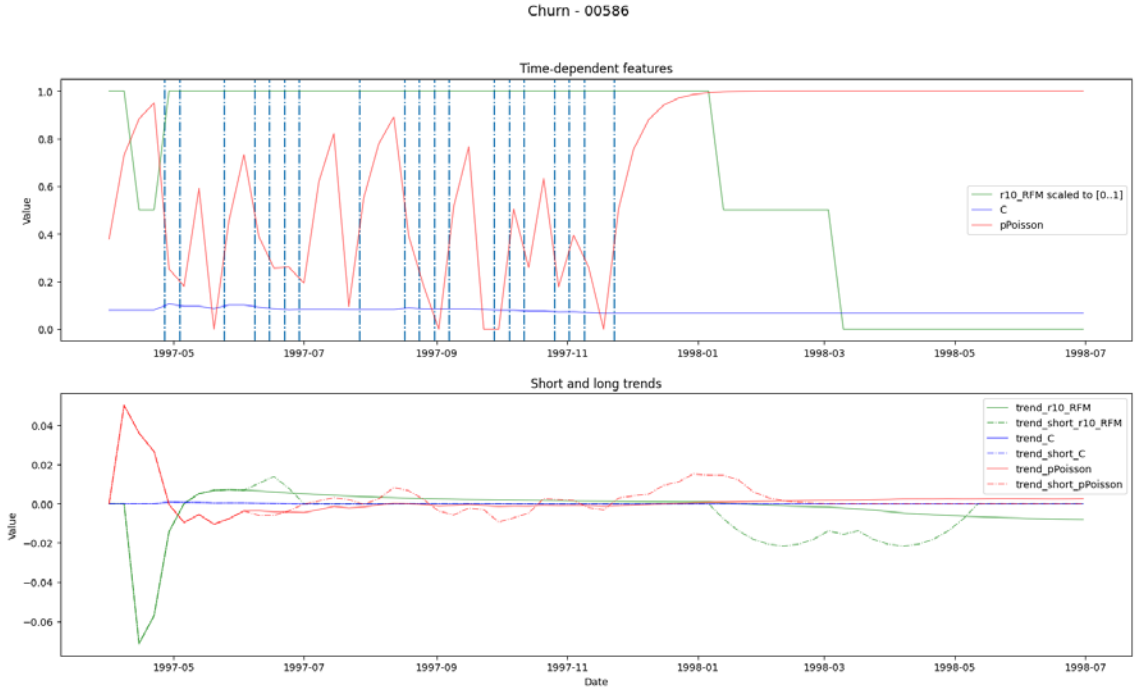

Figure 6. Randomly selected churning customer from CDNOW dataset, who made his last purchase relatively long time ago from the end of transactions data. Upper subplot shows how features r10_RFM, C and pPoisson

change in time. Vertical dashed lines correspond to time instances when customer made purchases. Lower subplot displays short and long trends of corresponding features from top subplot.

Chapter 2
Methods for Estimating Lifetime

Common terms in survival analysis

One way to describe the survival times of members of a group is to model a **survival function** that gives the probability that a patient, device, or other object of interest will survive past a certain time (1). The survival function is defined by:

$$S(t) = P(\{T > t\}) = \int_t^{\infty} f(u)du = 1 - F(t)$$

Equation 1. Survival function.

where F(t) is cumulative distribution function of continuous random variable T.

Another basic notion in survival analysis is the **hazard function**.

$$h(t) = \lim_{\delta>0} \frac{P(t \leq T < t + \delta | T \geq t)}{\delta}$$

Equation 2. Hazard function

It represents the instantaneous risk of experiencing the event at time t given it did not occur before. In our case T is discrete and can take values 1, 2, .. N, so hazard function is defined by:

$$h(t) = P(T = t | T \geq t)$$

and it is a probability that the event occurs at time t given it did not occur before.

Cumulative hazard rate is defined by:

$$H(t) = \int_0^t h(x)dx$$

Equation 3. Cumulative hazard rate.

and can be thought of as the total accumulated risk of experiencing death that has been gained by progressing to time t. While h(t) can increase or decrease with time, the cumulative hazard rate can only increase or remain the same. There is a direct connection between cumulative hazard rate and survival function:

$$S(t) = e^{-H(t)}$$

Equation 4. Survival-hazard connection.

$$H(t) = -\ln\left(S(t)\right)$$

Equation 5. Hazard-survival connection.

Expected survival time is the area under the survival curve:

$$E(T) = \int_0^{\infty} S(t)dt$$

Equation 6. Expected survival time.

Median survival time is the value t_i such that $S(t_i) = 0.5$

If we know that subject survived t^* time, then

Remaining expected survival time is:

$$E(T|T \geq t^*) = \int_{t^*}^{\infty} S(u)du$$

Equation 7. Remaining expected survival time.

Median survival time with condition that subject survived t^* is the value t_i such that:

$$S(t_i) = 0.5 * S(t^*)$$

Equation 8. Median survival time.

Therefore, remaining median survival time is $t_i - t^*$

Univariate survival models

We start with parametric univariate survival models such as Exponential, Weibull, LogNormal, LogLogistic, GeneralizedGamma as well as non-parametric Kaplan–Meier statistics that can estimate survival function from lifetime data. Generally speaking, parametric models have functional forms with parameters that we are to be determined by fitting to the data. Kaplan–Meier statistics will be used for Integrated Absolute Error (IAE) and Integrated Square Error (ISE) metrics since for real data the mathematical expression of the survival function is unknown, we will use an approximation obtained by Kaplan-Meier statistics. It should be mentioned that univariate estimators do not require additional data (features).

Kaplan-Meier statistics is the non-parametric KM estimation method, used to obtain the approximate expression of $S(t)$ defined by:

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

Equation 9. Kaplan-Meier estimate.

Where t_i is time when at least one event happened, d_i is the number of death events at time t and n_i is the number of subjects at risk of death just prior to time t, but not dead yet.

Exponential model has parameterized form with single parameter λ:

$$S(t) = e^{\frac{-t}{\lambda}}, \lambda > 0$$

Which implies the cumulative hazard rate is:

$$H(t) = \frac{t}{\lambda}$$

And hazard rate is:

$$h(t) = \frac{1}{\lambda}$$

Weibull model has two parameters:

$\lambda > 0$(scale), represents the time when 63.2% of the population has died

$\rho > 0$ (shape), controls if the cumulative hazard is convex or concave, representing accelerating or decelerating hazards.

The model has parameterized form:

$$S(t) = e^{-\left(\frac{t}{\lambda}\right)^{\rho}}$$

The hazard and cumulative hazard rates are:

$$h(t) = \frac{\rho}{\lambda}\left(\frac{t}{\lambda}\right)^{\rho-1} \qquad H(t) = \left(\frac{t}{\lambda}\right)^{\rho}$$

Log-normal model has two parameters: $\sigma > 0$ and μ and has a form:

$$S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$$

and cumulative hazard rate:

$$H(t) = -\ln\left(1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)\right)$$

where Φ is the CDF of standard normal random variable.

Log-logistic model has two parameters:

$\alpha > 0$ (scale), has an interpretation as being equal to the median lifetime of the population

$\beta > 0$ influences the shape of the hazard

Models survival function is defined by:

$$S(t) = \left(1 + \left(\frac{t}{\alpha}\right)^{\beta}\right)^{-1}$$

and corresponding hazard and cumulative hazard rates:

$$h(t) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{1+\left(\frac{t}{\alpha}\right)^{\beta}} \qquad H(t) = \ln\left(\left(\frac{t}{\alpha}\right)^{\beta} + 1\right)$$

Generalized Gamma has three parameters: μ, σ, λ. Its survival function is:

$$S(t) = \begin{cases} 1 - \Gamma_{RL}\left(\frac{1}{\lambda^2}; \frac{e^{\lambda\left(\frac{\ln t - \mu}{\sigma}\right)}}{\lambda^2}\right), & \lambda > 0 \\ \Gamma_{RL}\left(\frac{1}{\lambda^2}; \frac{e^{\lambda\left(\frac{\ln t - \mu}{\sigma}\right)}}{\lambda^2}\right), & \lambda \leq 0 \end{cases}$$

where Γ_{RL} is the regularized lower incomplete Gamma function. It should be mentioned that Exponential ($\lambda = 1, \sigma = 1$), Weibull ($\lambda = 1$) and Log-normal ($\lambda = 0$) are sub-models of Generalized gamma model.

Survival regression

The name implies the model regress covariates against another variable - duration. There are a few popular models in survival regression: Cox's model, accelerated failure (AFT) models. Some of them we are going to explore and we will try to use them to predict customers' lifetime.

Cox proportional hazard model (CoxPH)

The idea behind CoxPH is that the log-hazard of an individual is a linear function of their covariates and a population-level baseline hazard that changes over time. Cox hazard rate is the following:

$$\mathrm{H(t|x)} = \mathrm{h_0(t)e}^{\sum_{\mathrm{i=1}}^{\mathrm{n}} \beta_{\mathrm{i}}(\mathrm{x_i} - \bar{\mathrm{x}}_{\mathrm{i}})}$$

where $h_0(t)$ is time-dependent baseline hazard (the only time component) and exponential part is partial hazard, a time-invariant scalar factor that only increases or decreases the baseline hazard. Thus changes in covariates will only inflate or deflate the baseline hazard. Model does not specify entirely the survival function and the hazard function ($h_0(t)$ is left unspecified). The baseline hazard is modeled using Breslow (1975) method (non-parametrically) and the entire model is the traditional semi-parametric Cox model.

The survival function is:

$$S(t|x) = S_0(t)^{e^{\beta x}}$$

where $S_0(t)$ is baseline survival function that corresponds to zero vector β.

Accelerated failure time (AFT) regression models

All AFT models have general form:

$$T = \phi(x, \beta)T_0$$

where ϕ is a positive function that links the predictors x to an unknown vector of parameters β, T_0 is the random event time with all covariates equal to 0. Covariates increase or decrease through $\phi(x, \beta)$ the survival time compared to the reference.

Weibull AFT has parameterized form of survival function:

$$S(t; x; y) = e^{-\left(\frac{t}{\lambda(x)}\right)^{\rho}}$$

where

$\lambda(x) = e^{(\beta_0+\beta_1 x_1+..+\beta_n x_n)}$

Cumulative hazard rate is:

$$H(t; x; y) = \frac{t}{\lambda(x)}^{\rho}$$

Log-normal AFT has parameterized cumulative hazard rate:

$$H(t; x; y) = -\ln\left(1 - \Phi\left(\frac{\ln T - \mu(x)}{\sigma}\right)\right)$$

where:

$$\mu(x) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_n x_n$$

Log-logistic AFT has parameterized cumulative hazard rate:

$$\mathrm{H(t; x; y)} = \ln\left(1 + \left(\frac{\mathrm{t}}{\alpha(\mathrm{x})}\right)^{\beta}\right)$$

where

$$\alpha(\mathrm{x}) = \mathrm{e}^{(\alpha_0 + \alpha_1 \mathrm{x}_1 + \cdots + \alpha_n \mathrm{x}_n)}$$

Survival tree

Gradient-boosted Cox proportional hazard loss with regression trees as base learner (GB). The loss function is the partial likelihood loss of CoxPH model. The objective is to maximize the log partial likelihood function

$$\underset{f}{\operatorname{argmin}} \sum_{i=1}^{n} \delta_i \left[f(x_i) - \ln\left(\sum_{j \in R} e^{f(x_i)} \right) \right]$$

but instead of linear $\mathrm{x}^{\mathrm{T}}\beta$, the additive model

$$\mathrm{f(x)} = \sum_{\mathrm{m=1}}^{\mathrm{M}} \beta_{\mathrm{m}} \mathrm{g(x; \theta_m)}$$

is used. Here, M is the number of base learners, β_m is a weighting term and function g refers to a base learner parameterized by vector θ.

Metrics

To compare survival models the following metrics will be used:

Concordance index (CI) or the C statistic or Harrell's index is the number of concordant pairs of observations divided by the number of comparable pairs (Harrell, Califf, Pryor, Lee, & Rosati, 1982). The mathematical expression for CI is:

$$CI = \frac{\sum_{i,j} I\left(\tilde{T}_i > \tilde{T}_j\right) * I\left(\eta_j > \eta_i\right) * \Delta_j}{\sum_{i,j} I\left(\tilde{T}_i > \tilde{T}_j\right) * \Delta_j}$$

It is designed to estimate the concordance probability $P\left(\eta_j > \eta_i | T_i > T_j\right)$ which compares the rankings of two independent pairs of survival times T_i, T_j and predictions η_i, η_j. The concordance probability evaluates whether large values of η_i are associated with small values of T_i and vice versa.

Concordance index for right-censored data based on inverse probability of censoring weights (CI_IPCW) - is an alternative to the CI estimator and does not depend on the distribution of censoring times in the test data (Uno H. , Cai, Pencinac, D'Agostino, & Wei, 2011).

Akaike information criterion (AIC) – an estimator of prediction error and relative quality of models for a given set of data.

Bayesian information criterion (BIC) - a criterion for model selection among a finite set of models.

Integrated (time-dependent) Brier Score (IBS) provides an overall calculation of the model performance at all available times $t_1 \leq t \leq t_{max}$. IBS over the interval [t_1; t_{max}] is defined as

$$IBS = \int_{t_1}^{t_{max}} BS^c(t) dw(t)$$

where the weighting function is $w(t) = \frac{t}{t_{max}}$. The integral is estimated via the trapezoidal rule.

And BS is **time-dependent Brier score (BS)** for right censored data (the mean squared error at time point t):

$$BS^c(t) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \le t \wedge \delta_i = 1)\frac{(0 - \hat{\pi}(t|x_i))^2}{\hat{G}(y_i)} + I(y_i > \square)\frac{(1 - \hat{\pi}(t|x_i))^2}{\hat{G}(t)}$$

Where $\hat{\pi}(t|x)$ is the predicted probability of remaining event-free up to time point t for a feature vector x, and $\frac{1}{\hat{G}(t)}$ is an inverse probability of censoring weight, estimated by the Kaplan-Meier estimator.

Integrated Absolute Error (IAE) and Integrated Square Error (ISE) defined by:

$$IAE = \int_t \left|S(t) - \hat{S}(t)\right| dt \qquad ISE = \int_t \left(S(t) - \hat{S}(t)\right)^2 dt$$

where S(t) and $\hat{S}(t)$ represent the true survival function and the predicted survival function, respectively. However, in our case, the mathematical expression of the survival function is unknown, so the non-parametric Kaplan-Meier estimation method will be used to obtain the approximate expression of S(t).

Mean absolute error (MAE) will be used in simulated data only where 'true' survival time is known even for censored data. Is defined by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right|$$

where, N is size of censored test data, Y_i is 'true' remaining life and $\hat{Y}_i$ is expected or median predicted remaining life.

Chapter 3
Analysis of Real Datasets

CDNOW dataset

Choice of distribution

For better understanding of customers buying patters, the density histogram shown on **Figure 7** might be useful. It shows distribution of IPI of CDNOW dataset for the entire customers population. Obviously, it is right skewed, which means that short intervals are dominating. Also, this implies that distribution that can be used as representative for each customer should be right skewed as well.

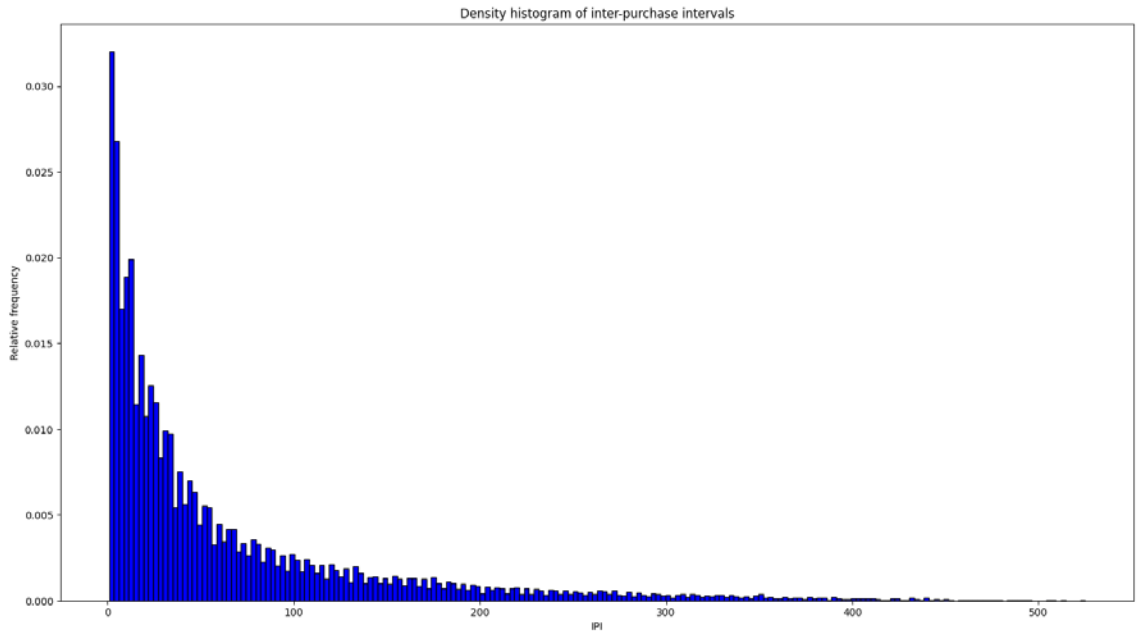

Figure 7. Density histogram of inter-purchase intervals for all customers made three and more purchases from CDNOW dataset.

Therefore, for the analysis we take five well-known distributions: 'Lognormal', ‘Fisk’, 'Weibull', 'Exponential' and 'Gamma' that might have right-skewed shape with long tail on the right side and 'Normal' to verify our assumption. Then, from transactions we filter only regular clients that made purchases 3 or more times each (if same client makes two or more transactions during one day we count them as one purchase). For each of those 7,473 customers we create a sequence of IPI (similar to that what was

described in the example above). Using MLE we estimate parameters for all of 6 mentioned distributions for all 7,473 clients. Then, we find TTD and TTC according to definitions described above for all 7,473 customers. The box plot on **Figure 8** shows distributions of TTD and TTC for 6 distributions.

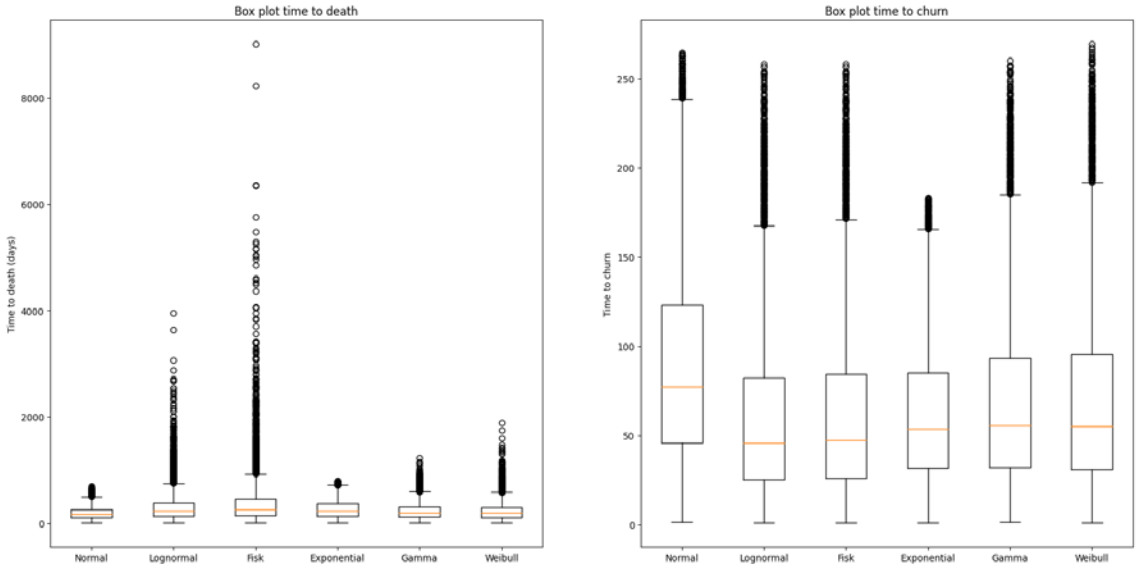

Figure 8. Box plot of distribution of time to death (left) and time to churn (right) for Normal, Lognormal, Fisk, Exponential, Gamma and Weibull distributions.

TTC distributions (median of each IPI sample) look similar for all models. However, TTD (98% quantile) looks different: Fisk distribution which has heavy tail, often gives large numbers TTD followed by lognormal and Weibull. Since variance of TTC is relatively low, long tails of Fisk and lognormal distributions (98% quantile is situated on tail) might signify that many customers have non-regular buying patterns [5, 10, 100] for example. The density histogram on **Figure 9** shows that distribution of number of transactions made by each customer.

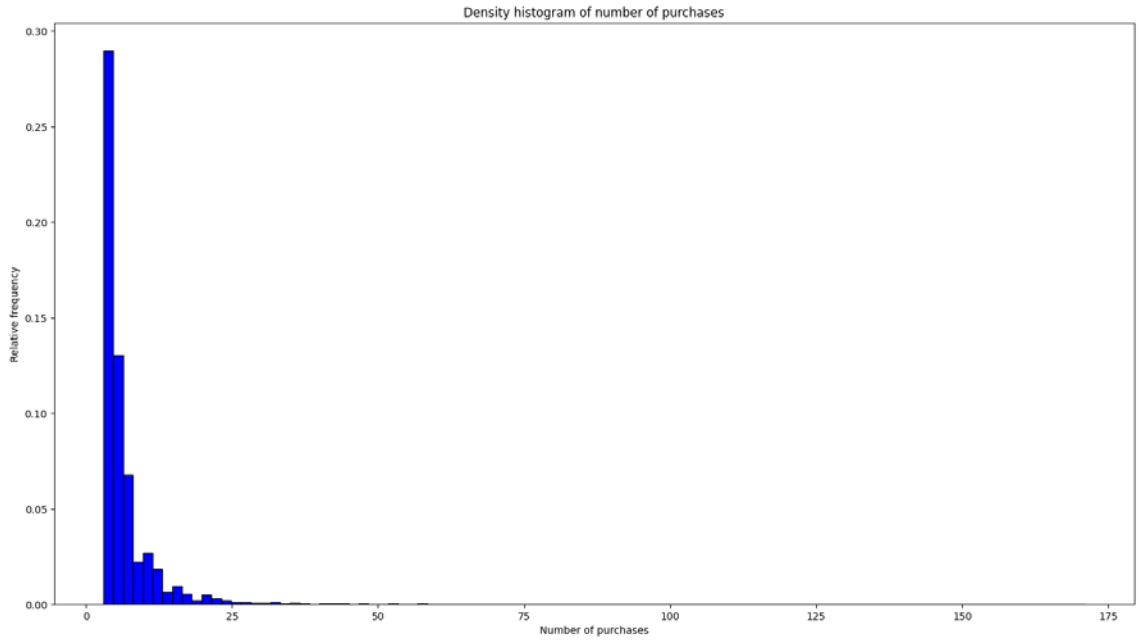

Figure 9. Histogram of number of purchases (frequency) made by regular customers, CDNOW dataset.

It shows that most of clients make only few purchases dusing the transactions perod. The **Figure 10** shoes that values of TTD and TTC go down as number of transactions increases. This is not surprising since transactions interval is limited: first date - 1997-01-01, last date - 1998-06-30, interval length: 545 days. So, customers who made many transactions during this period did them frequently having short IPI. This also means that most of clients have large TTD and for many of them we will not be able to see their ‘death’ even if it has happened (again, due to limited transactions period).

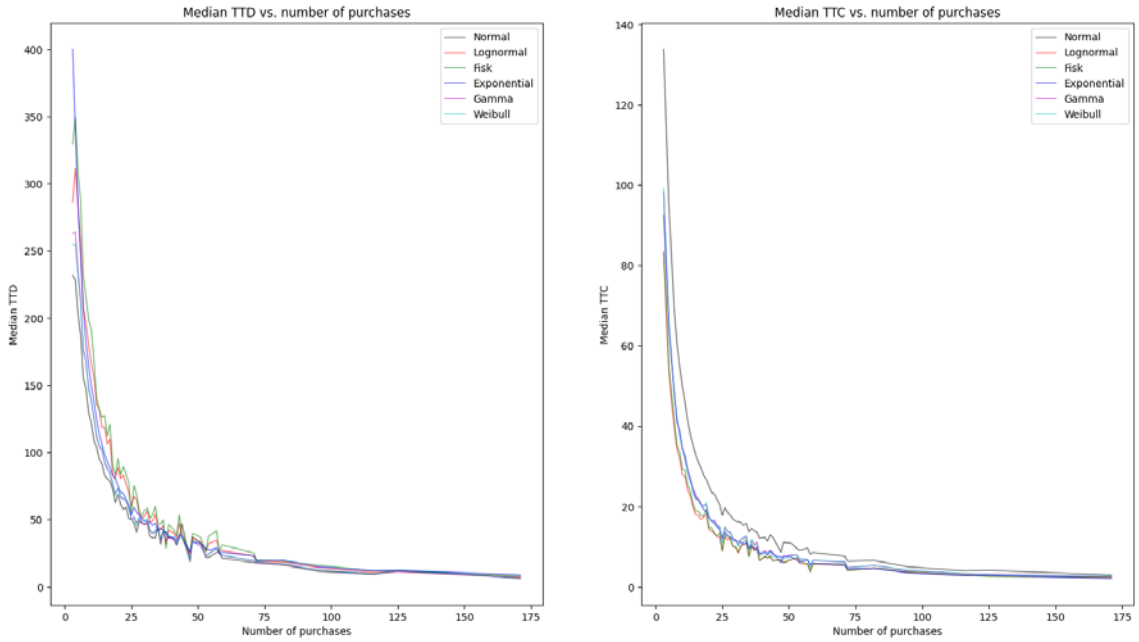

Figure 10. Plot median TTD (left) and median TTC (right) versus number of purchases. There are six plots; each corresponds to one of distributions: Normal, Lognormal, Fisk, Exponential, Gamma, Weibull.

In order to find a distribution that might be used for best representation of customers buying pattern we will be using Kolmogorov–Smirnov test (KS) and Cramér–von Mises criterion (CvM). For each regular customer's IPI sequence we already estimated parameters for all 6 distributions. Now by comparing a sample (customer's sequence of IPI) with a reference probability distribution (each of 6 mentioned above one by one) we get six p-values (for each sample). Tests (either of KS or CvM) might answer the question "what is the probability that sample could have been drawn from the probability distribution". P-value, can be used as such a metric, so the highest p-value obtained by tests corresponds to distribution that is representative for the sample. The histogram on **Figure 11** summarises results obtained by KS and CvM tests. It shows that about 60% of population samples could have been drawn from Fisk (log-logistic) distribution.

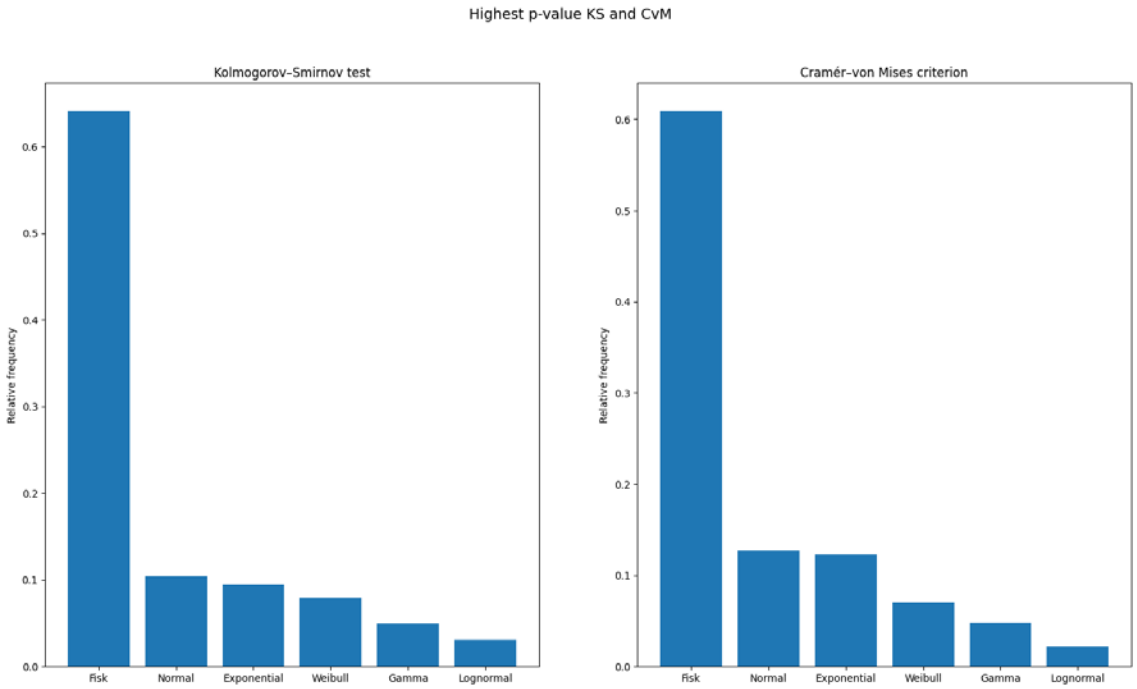

Figure 11. Histogram of the distribution of largest p-value obtained by KS (left) and CvM (right). All regular customers from CDNOW dataset.

Figure 12 and **Figure 13** show in details the distributions of p-values for each of two test statistics for all 6 distributions. Looking at peaks near the point where p-value is high (more than 0.9), we can see that the difference in p-values for all test statistics (except exponential) is not very large and any of them might be taken for TTD and TTC calculation. However, we will be using Fisk distribution having in mind the fact, that according to tests, log-logistic distribution is the most representative, it has long tail which might be a good feature to model irregular buying patterns.

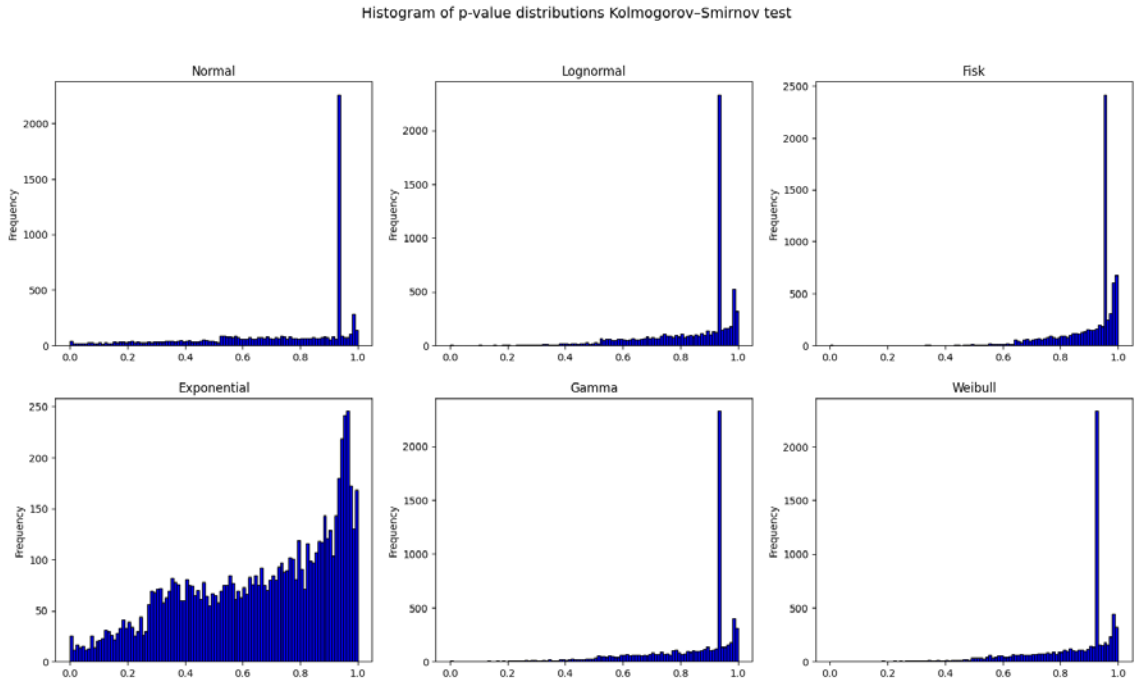

Figure 12. Histogram of p-values obtained by KS test for each of six distributions fitted on regular customers IPI from CDNOW dataset.

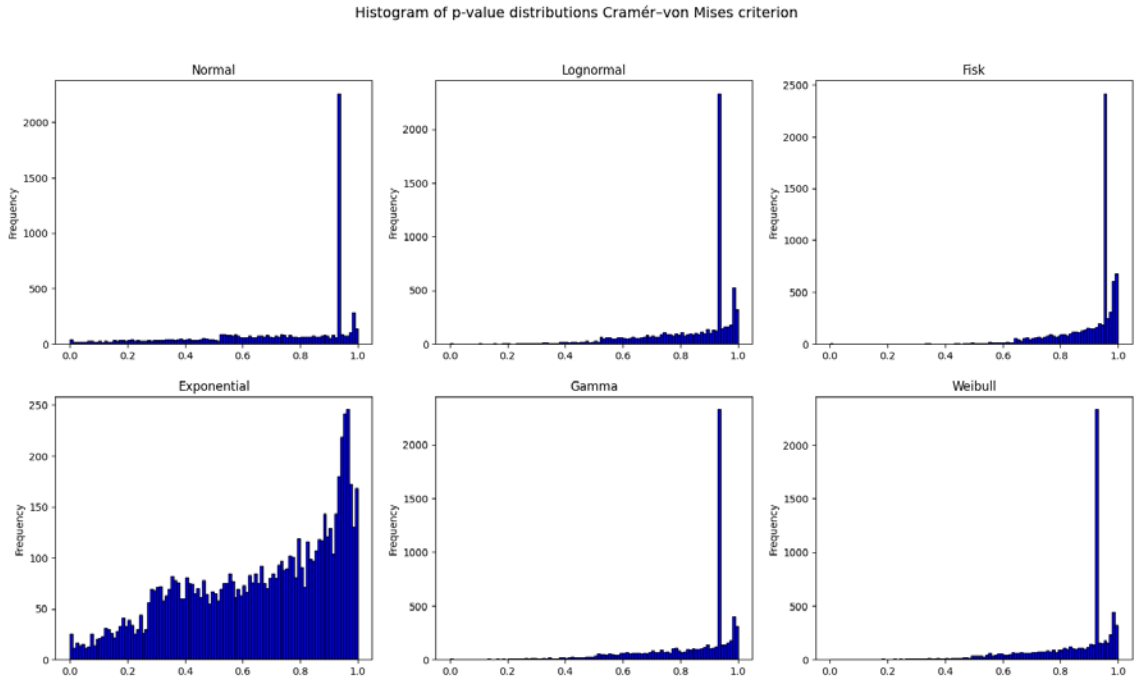

Figure 13. Histogram of p-values obtained by CvM test for each of six distributions fitted on regular customers IPI from CDNOW dataset.

Univariate models

First of all we will try to estimate the survival function for the population by univariate models. To be more precise we will try to fit five parametric models (Exponential, Weibull, Lognormal, log-logistic and generalized gamma) and non-parametric Kaplan-Meier statistics. All mentioned parametric models have their functional forms with parameters we are going to determine by fitting to the training data. We take a fraction equal to 0.2 from both censored and dead observations to form a test subset of size 1,440 observations; the remaining 5,766 observations would be used as training data. All described parametric models as well as Kaplan–Meier statistics are implemented in 'lifelines'[25] python library. We fit each model to training data; AIC and BIC criterions are provided within models' class after convergence. BS and IBS metrics is implemented in another python library named 'scikit-survival'[26].

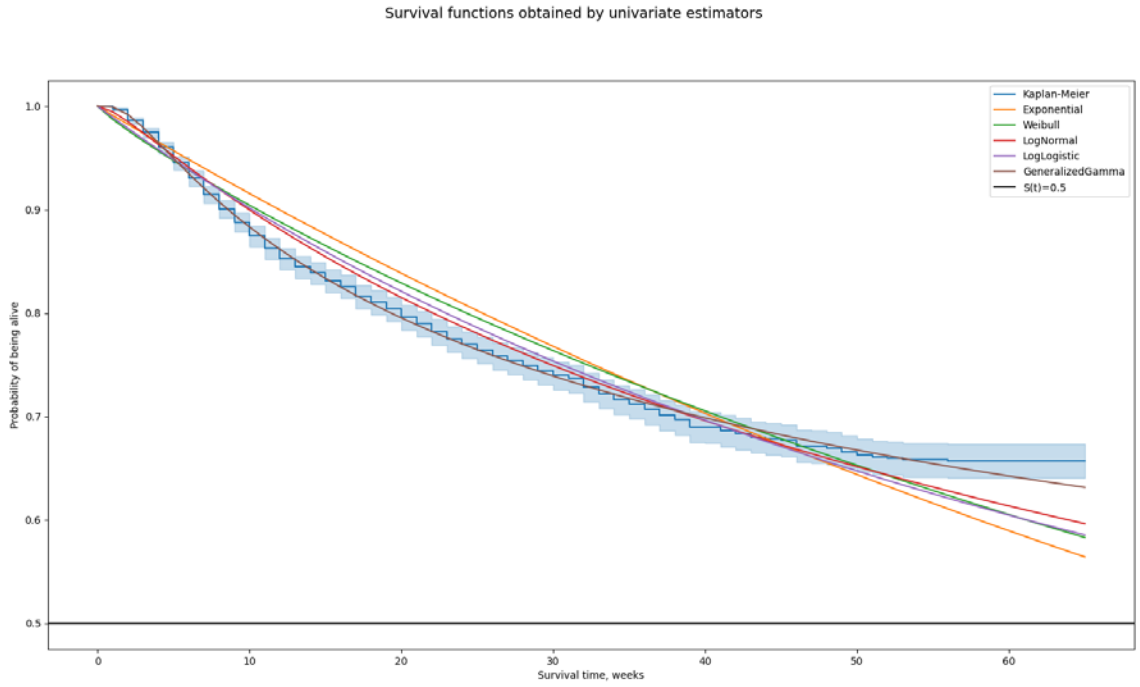

Figure 14. Survival functions obtained by five parametric models and Kaplan-Meier estimatior on training subset of CDNOW dataset.

From the **Figure 14** it can be observed that survival curves do not cross horizontal line that corresponds to value equal 0.5. This might occur when the largest observed

[25] https://lifelines.readthedocs.io/en/latest/Survival%20Analysis%20intro.html (accessed on December 21, 2022)
[26] https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html (accessed on December 21, 2022)

time is censored and implies that for the churn rate of the population from CDNOW data, available study interval of 64 weeks is too short since it covers approximately one third of median lifetime of customers population (best model predicts median lifetime equals to 197 weeks). In ideal case, study interval should be long enough to allow the Kaplan-Meier curve to drop to the value 0.05-0.1 or even better to the value close to zero. In **Table 1**, determined parameters of five univariate models can be found as well as 95% confidence intervals.

Model	CoefName	coef	se(coef)	coef lower 95%	coef upper 95%
Exponential	lambda_	112.8335	3.1537	106.6522	119.0149
Weibull	lambda_	130.7778	5.7958	119.4182	142.1374
	rho_	0.8844	0.0210	0.8431	0.9258
LogNormal	mu_	4.6147	0.0483	4.5199	4.7096
	sigma_	1.8218	0.0394	1.7445	1.8991
LogLogistic	alpha_	92.1252	3.9726	84.3390	99.9114
	beta_	0.9864	0.0230	0.9413	1.0315
GeneralizedGamma	mu_	3.3843	0.1187	3.1517	3.6170
	ln_sigma_	0.8124	0.0216	0.7699	0.8549
	lambda_	-1.9496	0.1616	-2.2663	-1.6328

Table 1. Parameters and confidence intervals obtained by five univariate models by fitting on train subset of CDNOW data.

For those models we used a subset of metrics described previously, particularly AIC, BIC, IBS, IAE, ISE. Since CDNOW is real data and we do not know true lifetime, we used the Kaplan-Meier estimator to obtain the approximate expression of S(t) for IAE and ISE calculations. Those metrics as well as expected and median survival times are summarized in **Table 2**.

	AIC	BIC	IBS	IAE Median	ISE Median	tExpected	tMedian
Exponential	14660.3401	14666.9999	0.1773	2.2954	0.1131	112.83	78.21
Weibull	14634.5699	14647.8894	0.1768	1.7564	0.0653	138.94	86.41
LogNormal	14442.5641	14455.8835	0.1765	1.2802	0.0387	433.67	100.97
LogLogistic	14566.6883	14580.0077	0.1767	1.5399	0.0545	404.51	92.13
GeneralizedGamma	14307.7362	14327.7154	0.1764	0.3613	0.0036	1931.7	197.25

Table 2. Summary of available metrics to describe the goodness of fit of five univariate parametric models. tExpected and tMedian are expected and median survival times obtained by corresponding models for population from training set of CDNOW data.

Among univariate parametric models Generalized Gamma seems to be the best representation: all shown metrics indicate that this estimate is the closest to real data. IAE and ISE scores of generalized gamma model are significantly better than other models have meaning that survival function modeled by GGM mimics very well the survival curve obtained by Kaplan-Meier estimator.

Survival regression models

Usage of regression models implies having some predictors in possession. From the described list of covariates for survival regression we will be using the following: 'C', 'trend_C', 'trend_short_C', 'moneySum', 'moneyDaily', 'r10_F', 'r10_RF', 'r10_FM', 'r10_RFM', 'trend_r10_RF', 'trend_short_r10_RF', 'trend_r10_FM', 'trend_short_r10_FM', 'trend_r10_RFM', 'trend_short_r10_RFM', 'trend_r10_RFML', 'trend_short_r10_RFML'. They are all numeric, so we will standardize them before fitting procedure (zero mean, unit variance). Due to the popularity of semi-parametric Cox proportional hazard model (Cox, 1972), we will try two different implementations from two python libraries: 'lifelines' and 'scikit-survival'. The most common way of CoxPH model usage is its standalone version in accompany with the estimator of baseline hazard (both selected models use Breslow estimator). Another interesting approach to use Cox which we are going to try is proposed by 'scikit-survival' library is implemented in model GradientBoostingSurvivalAnalysis. It is a week learner that has a gradient-boosted Cox proportional hazard loss with regression trees as base learner. AFT parametric models provide an alternative to CoxPH, so we will fit to data three AFT models from 'lifeline' library: 'WeibullAFT, LognormalAFT' and 'LoglogisticAFT'. The heat map on **Figure 15** shows Pearson correlation between our predictors. As expected, composite features (or their trends) are correlated with single features that were used as components to construct complex ranks. For example the correlation between r10_F (frequency rank) and r10FM (frequency – daily spending product) is 0.92. High correlation can be also observed between composite predictors such as trend_r10_RF and trend_r10_RFML for example. Existence of correlated predictors might cause problems for convergence of some algorithms (generalized gamma model is the most vulnerable among those we tried), but since we do not know

yet the predictive power for each of them, we would keep this feature subset for our modeling and will use small regularization penalty to let algorithms converge.

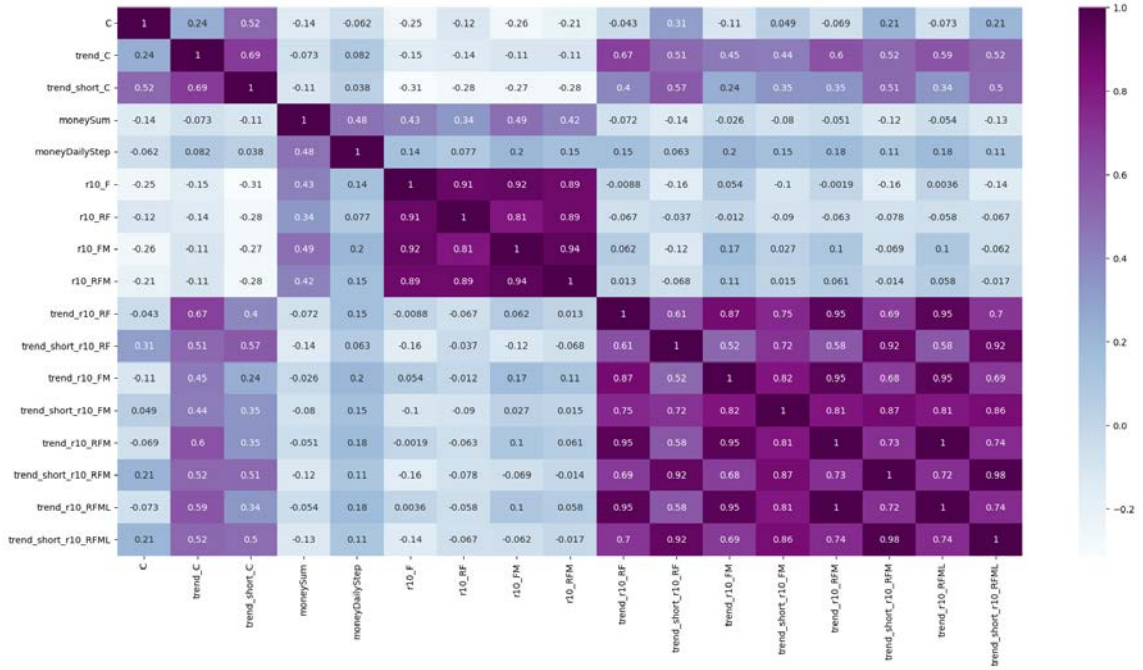

Figure 15. Pearson correlations between features we are going to use as predictors in survival regression.

On **Figure 16** coefficients obtained by two CoxPH models after fitting to train data are shown. The one from lifelines provides confidence intervals as well (black color on plot). The red dots represent coefficients determined by COX model from 'scikit-survival'. Parameters from two models that correspond to same features have slightly different magnitudes, but same sign (except those that are very close to zero). It should be remained that proportional model in this case means that increasing a covariate xi by 1 scales the baseline hazard by e^{β_i}.

Summary on **Table 3** shows coefficients, determined by 'lifelines' COX model. They have different magnitude and therefore features should have different influence on the survival function.

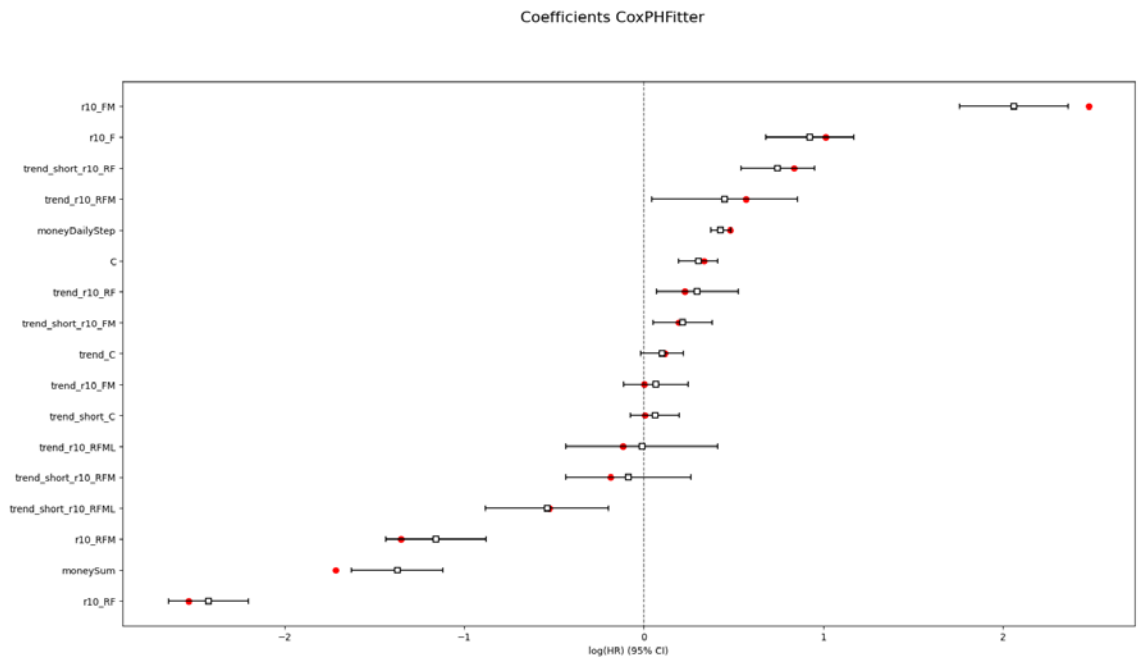

Figure 16. Coefficients of two different implementations of CoxPH. Black color corresponds to coefficients and confidence intervals for the model from 'lifelines', red dots show coefficients from 'scikit-survival'.

covariate	coef	exp(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%
C	0.3033	1.3543	0.1944	0.4122	1.2146	1.5102
trend_C	0.1021	1.1075	-0.0175	0.2217	0.9825	1.2483
trend_short_C	0.0614	1.0633	-0.0727	0.1956	0.9298	1.2160
moneySum	-1.3735	0.2531	-1.6295	-1.1176	0.1960	0.3270
moneyDailyStep	0.4274	1.5333	0.3723	0.4824	1.4512	1.6200
r10_F	0.9245	2.5208	0.6807	1.1684	1.9754	3.2168
r10_RF	-2.4241	0.0885	-2.6458	-2.2024	0.0709	0.1105
r10_FM	2.0604	7.8496	1.7568	2.3641	5.7940	10.6347
r10_RFM	-1.1583	0.3140	-1.4382	-0.8784	0.2373	0.4154
trend_r10_RF	0.2975	1.3465	0.0694	0.5255	1.0719	1.6914
trend_short_r10_RF	0.7458	2.1081	0.5419	0.9497	1.7193	2.5849
trend_r10_FM	0.0680	1.0704	-0.1122	0.2483	0.8938	1.2819
trend_short_r10_FM	0.2149	1.2398	0.0498	0.3801	1.0511	1.4624
trend_r10_RFM	0.4486	1.5662	0.0426	0.8547	1.0435	2.3507
trend_short_r10_RFM	-0.0847	0.9187	-0.4335	0.2640	0.6482	1.3021
trend_r10_RFML	-0.0106	0.9894	-0.4335	0.4122	0.6482	1.5102
trend_short_r10_RFML	-0.5395	0.5830	-0.8825	-0.1964	0.4137	0.8216

Table 3. Coefficients and corresponding confidence intervals determined by 'lifeline' CoxPH model.

Let's take a closer look on small subset of four features: r10_FM, C, r10_RF and r10_RFM. First two features have positive coefficients and e^{β} greater than 1 which means that if an individual has large values of those predictors, his hazard will be large as well and therefore, his survival time will be low. Last two features have negative coefficients and corresponding e^{β} less than 1, so large values of features will make a smaller hazard rate and longer survival time. **Figure 17** shows in details the impact of mentioned features on survival time according to CoxPH model. Varying original feature values are shown on plots legends; they cover the range each feature might take (1 to 10 for ranks and 0 to 1 for clumpiness). Baseline survival function is provided for comparison and represents the survival time for a subject with median value of each feature. Different survival curves from **Figure 17** are concordant with magnitudes of coefficients from table: if exp(coefficient) has magnitude greater than one, the impact of corresponding covariate on hazard is the following: larger original values of covariate increases hazard and decreases survival time, predictors r10_FM and C for example.

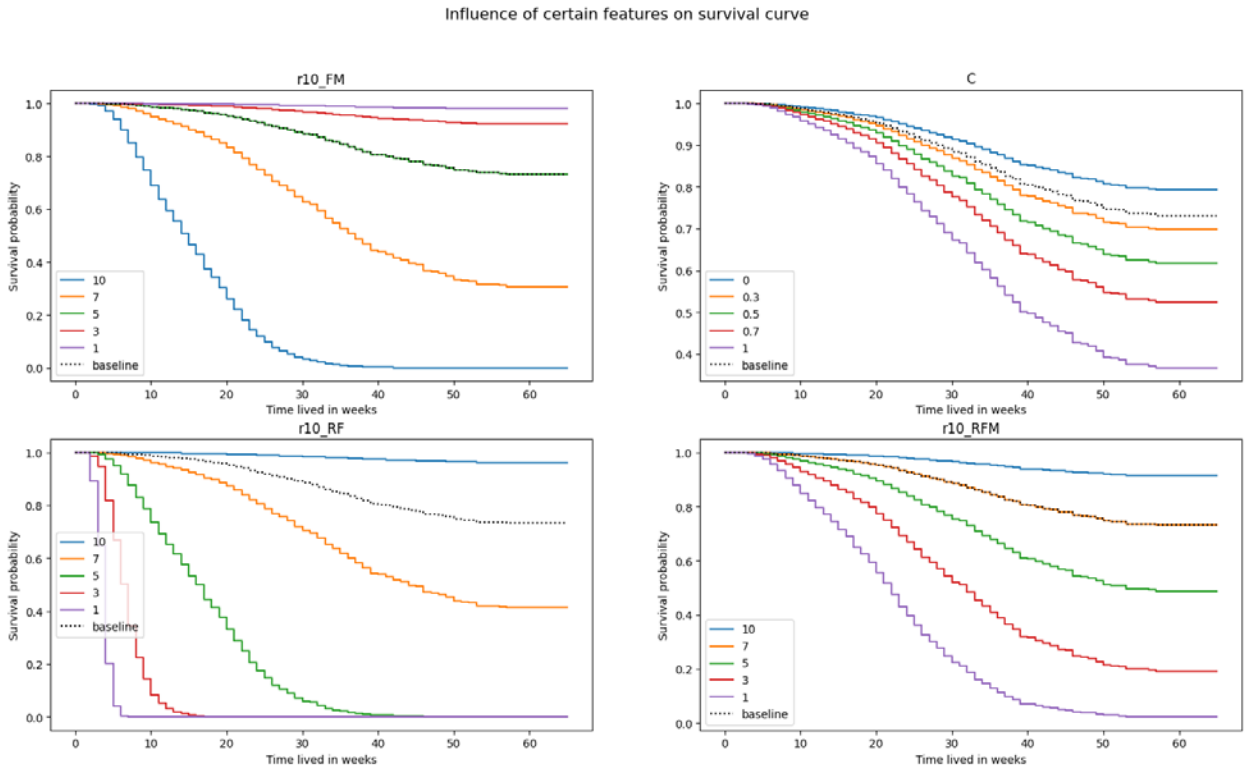

Figure 17. Partial impact of certain features on survival time of CoxPH model from 'lifelines'. Top two plots demonstrate negative impact on survival time, bottom ones – positive impact.

AFT models differ from COX proportional model in the sense that the covariates have the multiplicative effect directly on survival time. **Figure 18** shows Log-normal model fitted coefficients: this time coefficient of r10_RF that had the largest negative value in CoxPH model, has largest positive value (except intercept) as expected since parameters of CoxPH influences directly the hazard function, but coefficients of AFT model have similar direct impact on survival time (see Equation 4 and Equation 5).

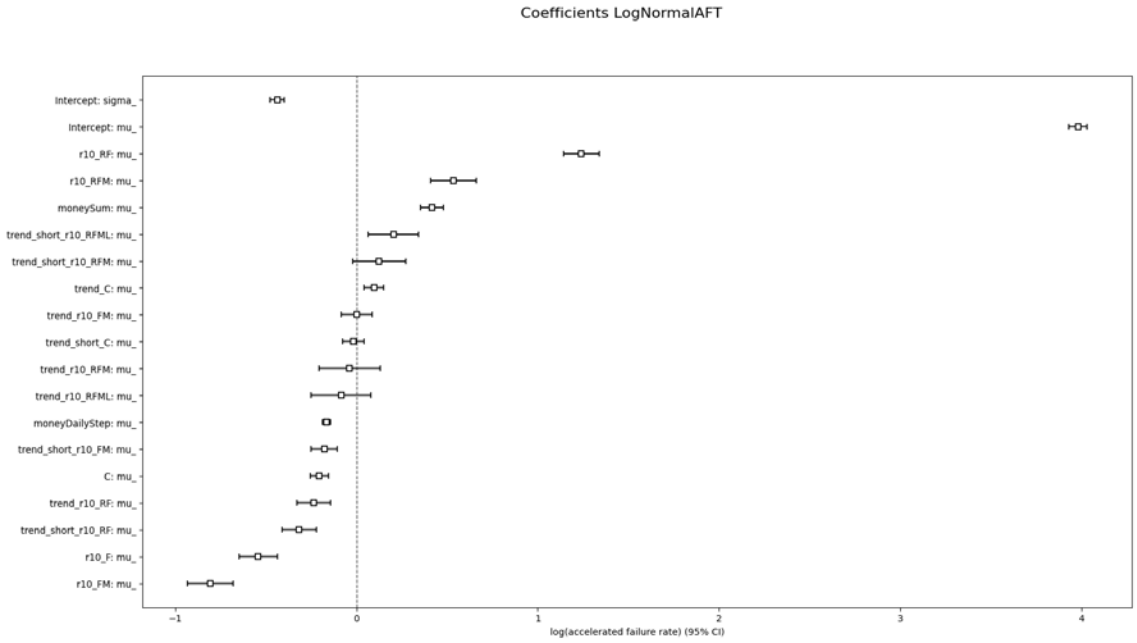

Figure 18. Coefficients determined by Lognormal AFT model on training data from CDNOW.

Table 4 shows all fitted coefficients and intercept, determined by Log-normal AFT model by fitting training data as well as confidence intervals.

param	covariate	coef	exp(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%
mu_	C	-0.2050	0.8145	-0.2559	-0.1542	0.7741	0.8570
	moneyDailyStep	-0.1647	0.8480	-0.1862	-0.1432	0.8300	0.8664
	moneySum	0.4153	1.5148	0.3521	0.4785	1.4220	1.6137
	r10_F	-0.5426	0.5812	-0.6484	-0.4368	0.5228	0.6460
	r10_FM	-0.8054	0.4469	-0.9306	-0.6801	0.3942	0.5065
	r10_RF	1.2408	3.4585	1.1416	1.3401	3.1317	3.8194
	r10_RFM	0.5338	1.7055	0.4084	0.6592	1.5045	1.9333
	trend_C	0.0970	1.1019	0.0438	0.1503	1.0447	1.1622
	trend_r10_FM	0.0023	1.0023	-0.0825	0.0872	0.9207	1.0911
	trend_r10_RF	-0.2359	0.7897	-0.3280	-0.1438	0.7202	0.8660

	trend_r10_RFM	-0.0387	0.9620	-0.2072	0.1298	0.8127	1.1386
	trend_r10_RFML	-0.0853	0.9181	-0.2499	0.0791	0.7788	1.0824
	trend_short_C	-0.0171	0.9830	-0.0774	0.0431	0.9254	1.0441
	trend_short_r10_FM	-0.1774	0.8374	-0.2492	-0.1056	0.7793	0.8997
	trend_short_r10_RF	-0.3153	0.7295	-0.4103	-0.2203	0.6634	0.8022
	trend_short_r10_RFM	0.1251	1.1332	-0.0202	0.2704	0.9799	1.3106
	trend_short_r10_RFML	0.2035	1.2257	0.0650	0.3421	1.0671	1.4079
	Intercept	3.9794	53.4885	3.9295	4.0294	50.8816	56.2289
sigma_	Intercept	-0.4360	0.6465	-0.4746	-0.3975	0.6220	0.6719

Table 4. Coefficients and confidence intervals determined by fitting Lognormal AFT model to training data of CDNOW dataset.

Figure 19 shows plots of effect of some of varying covariates obtained from Lognormal AFT. It can be observed that this parametric model produces the similar survival curves for the subset of features we used for CoxPH. Here large values of features 'r10_FM' and 'C' decrease survival time similarly to CoPH model, but their exponents have values smaller than one. Same concordance can be remarked by observing plots of features r10_RF and r10_RFM. **Figure 17** and **Figure 19** illustrate that two models with different learning approaches (proportional hazard and accelerated failure) find appropriate feature impacts on survival function. Also, coefficient plots of three AFT models (**Figure 18**, **Figure 20** and **Figure 21**) show that often, relative impact of features (related to other covariates) is similar: for example, r10_RF followed by r10_RFM have greatest positive impact on survival function.

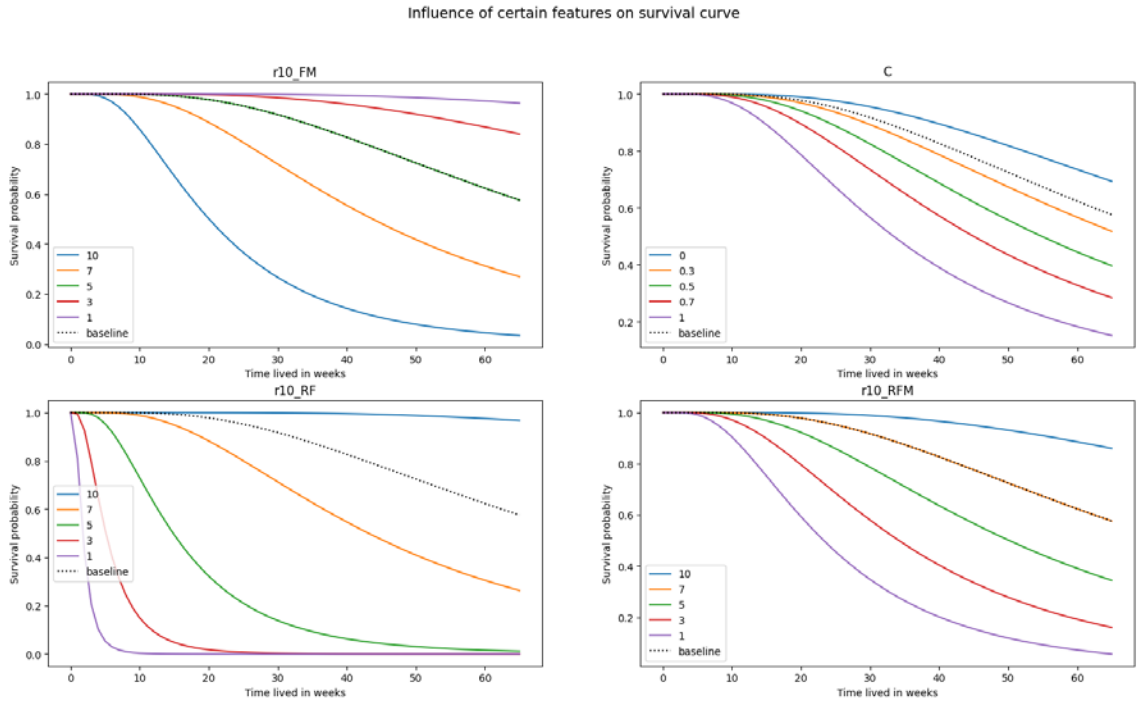

Figure 19. Partial effect of certain features on survival time of Lognormal AFT model. Top two plots demonstrate negative impact on survival time, bottom ones – positive impact.

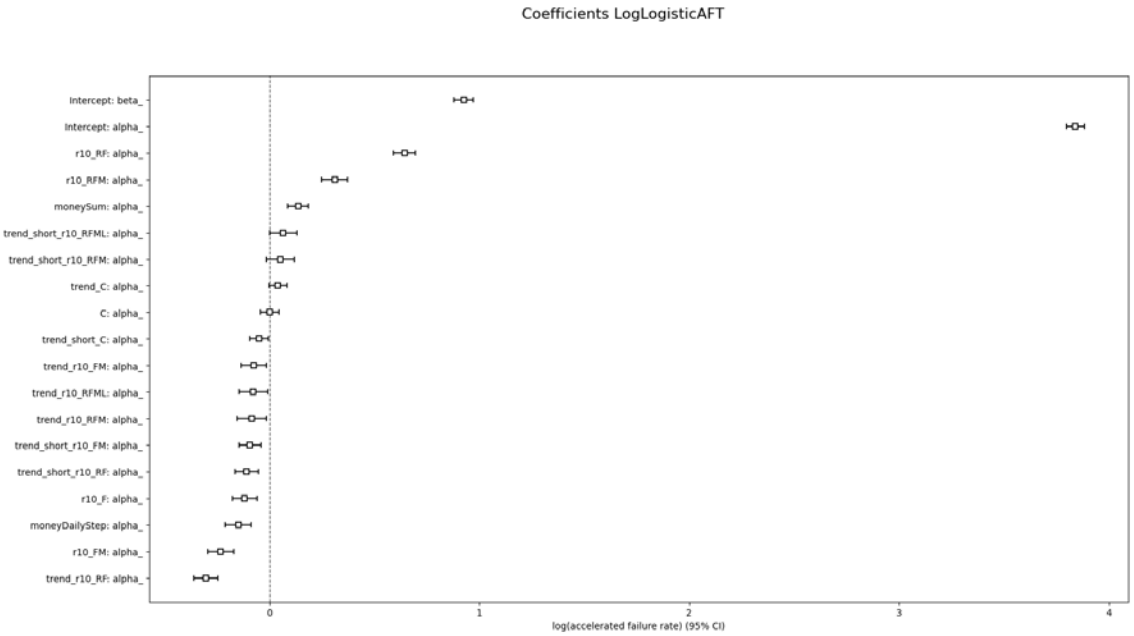

Figure 20. Coefficients determined by Log-logistic AFT model on training data from CDNOW.

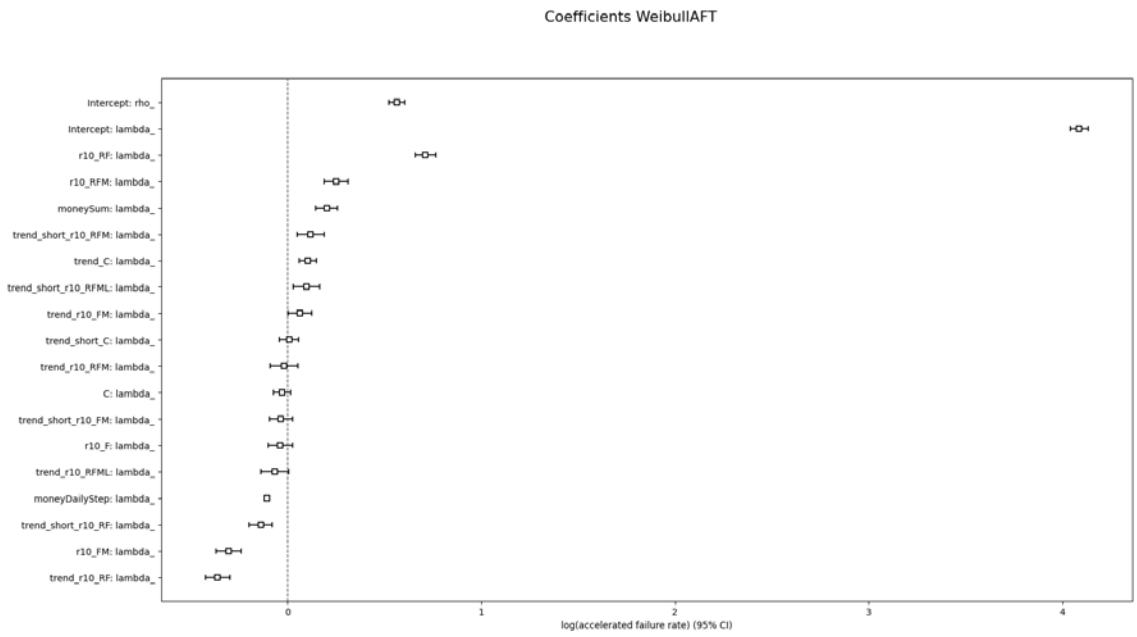

Figure 21. Coefficients determined by Weibull AFT model on training data from CDNOW.

GB model is tree based; it has a possibility to show feature importance obtained from out of bag observations during the training process: predictors that have high values are more important than the ones with low values. **Figure 22** illustrates relative feature importance on decision made by GB algorithm.

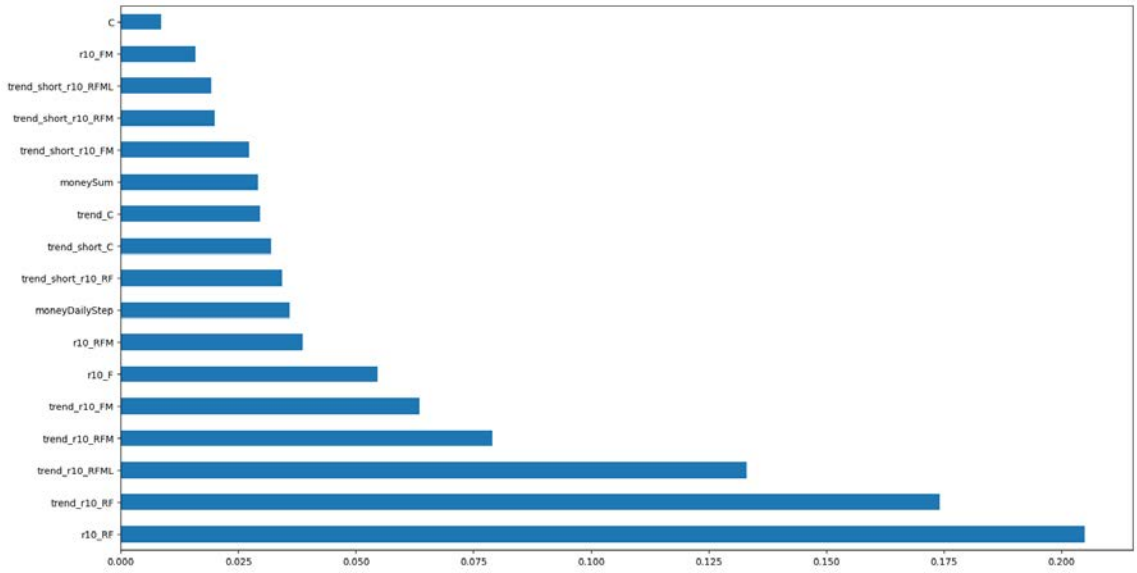

Figure 22. Feature importance by GB on out of bag observations of training data.

Unfortunately, it does not show the sign of impact, but results from this plot partially correlate with results from CoxPH and AFT models: features r10_RF and some of trend features have relatively large (in magnitude) coefficients and significant in GB model.

The summary of metrics of regression models is presented in **Table 5**. Results were obtained by making predictions by each of six regression models on test data that has 1440 observations. If model cannot estimate the expected or median survival time for a subject, the observation is not taken into account. As we described before, AFT models can provide the survival function of any desired length regardless the duration of study period. On the contrary, COX and GB models cannot extrapolate the survival function beyond the duration of study period (64 weeks). Therefore, if survival curve do not cross horizontal line that corresponds to value equal 0.5 median survival time becomes infinity. To estimate the expected survival time, survival function must approach to zero at reasonably small distance (we counted 0.05 is close enough to zero), otherwise, expectation is infinity again. Row tExpected shows average expected survival time for the population except those who have been predicted infinity. Row sizeMedian tells us how many test observations we were able to predict. Similarly, tMedian and size Median are median survival time for population and number of eligible subjects. For this particular dataset, COX models are able to predict median survival time for only about one third of test observations and GB less than one fourth. The number of subjects that models could estimate expected survival time is much smaller. Obviously, COX and GB models require longer study period or larger churn rate. Row rankCI shows models rank from best to worst according to CI IPCW metric only. From scores in **Table 5**, for CDNOW dataset, Lognormal AFT seems to be the most promising algorithm for lifetime estimation: CI, AUC scores are close to the leading model (GB), but the ability to predict survival time for all subjects makes this model more favorable than GB.

	GradientBoosting	LogNormalAFT	CoxPHFitter	CoxPHSurvival	LogLogisticAFT	WeibullAFT
CI	0.9804	0.9426	0.9433	0.9419	0.9251	0.9089
CI IPCW	0.9656	0.9071	0.9013	0.8993	0.8812	0.8612

IBS	0.0307	0.0548	0.0501	0.0502	0.0575	0.0687
AUC	0.9943	0.9723	0.9731	0.9732	0.9649	0.9532
IAE Median	11.3245	12.6572	12.0433	12.2155	11.2482	9.1168
ISE Median	2.1671	2.8477	2.5786	2.6508	2.2315	1.5246
tExpected	10.81	94.46	10.2	9.96	83.12	70.43
sizeExpected	223	1437	182	177	1440	1439
tMedian	14	75	25	24	63	65
sizeMedian	331	1438	498	463	1440	1440
rankCI	1	2	3	4	5	6

Table 5. Summary of metrics and results obtained by six regression modes on training subset of observations of CDNOW datased.

Box plots on **Figure 23** show the distribution of IAE (left) and ISE (right) for each of mentioned regression models.

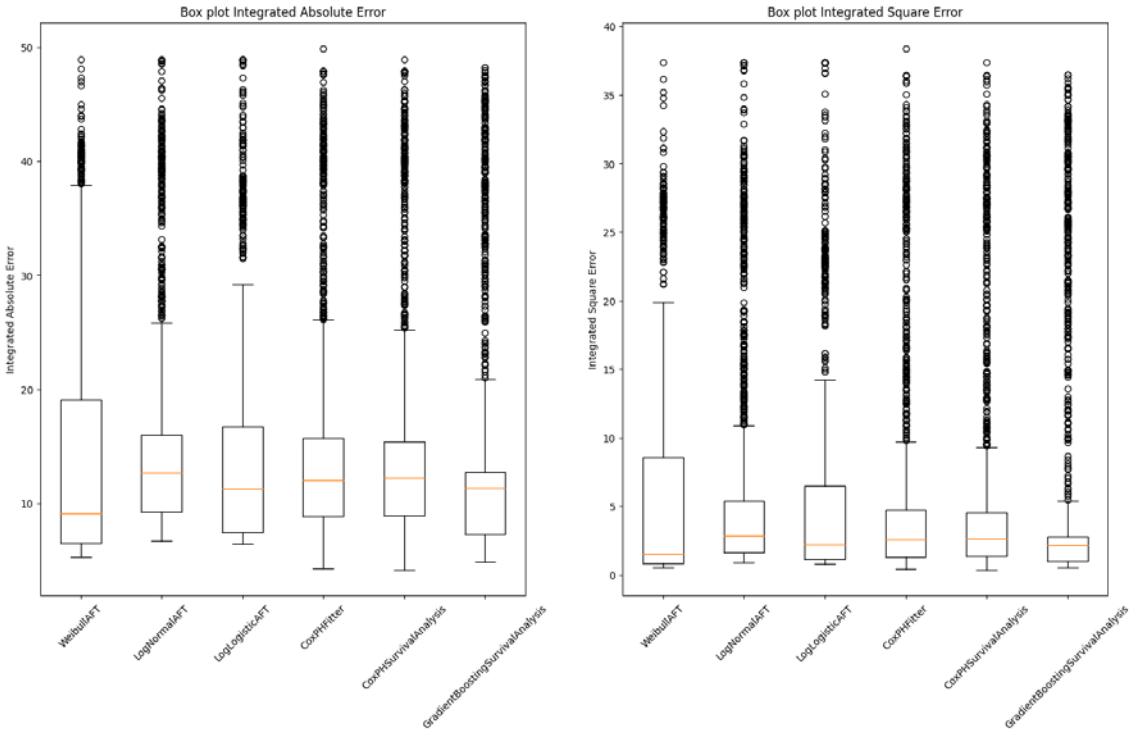

Figure 23. Box plot of the distribution of IAE (left) and ISE (right) of six regression models.

Distributions of expected and median survival times for same six models are shown on **Figure 24**. Below the each box, there is an indicator of the number of test observations that specified model was able to predict. Small boxes of COX and GB are consequences of small study period restrained by the duration of 64 weeks.

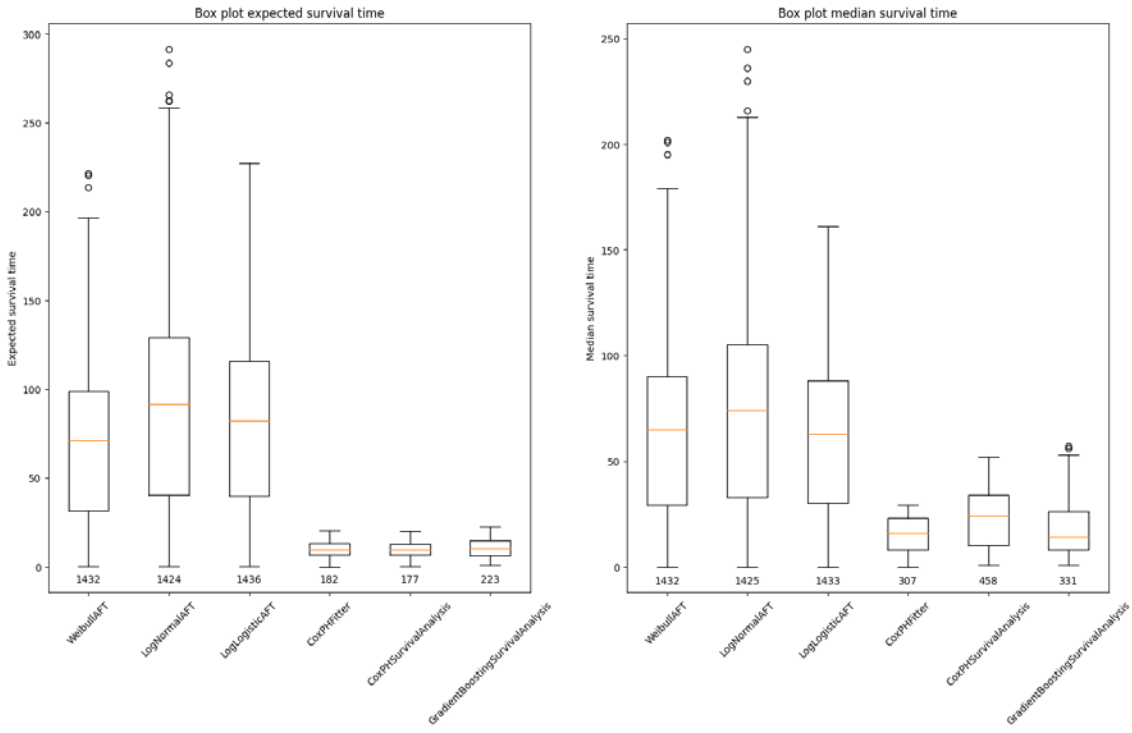

Figure 24. Box plots of distributions of expected (left) and median (right) survival time obtained by six models on test subset. Numbers below boxes indicate the quantity of observations corresponding model was capable to estimate.

Since our study period is limited to 64 weeks Cox model cannot estimate lifetime beyond this period. If survival curve is undefined before it reaches 0.5, median lifetime cannot be estimated, and therefore becomes infinity. The situation is even worse for expected value; to evaluate it properly, estimated survival curve from the Cox model must reach zero value, which is in fact very rare case for our dataset. We could integrate up to the maximum available value but it's not clear that this is a good strategy with severely censored data. On **Figure 24** all predictions beyond the limit were replaced by the highest observed lifetime (64 weeks). As a result, this value became a median for the population. This situation was expected after analyzing results from previous (univariate) models where all mean and median predicted values for population lifetime excided study interval. This is not exactly the Cox's models problem: more precisely this complication arises from $h_0(t)$ term which in fact is been estimated by Breslow's method. If we had a possibility to extrapolate baseline hazard, Cox model would be applicable for our data. Unfortunately, existing python packages for survival analysis ('lifeline' and 'scikit-survival') do not have other options. In the contrary, fully

parametric models allow us to construct survival function of the desired length. On **Figure 25** survival curves of 5 randomly selected subjects, obtained by CoxPH model are shown with corresponding expected and median survival times. Baseline survival function is shown in bold black. In this particular sample, median survival time could be estimated by Cox model for only two subjects. Expected lifetime cannot be estimated by Cox at all since survival curves do not approach to zero.

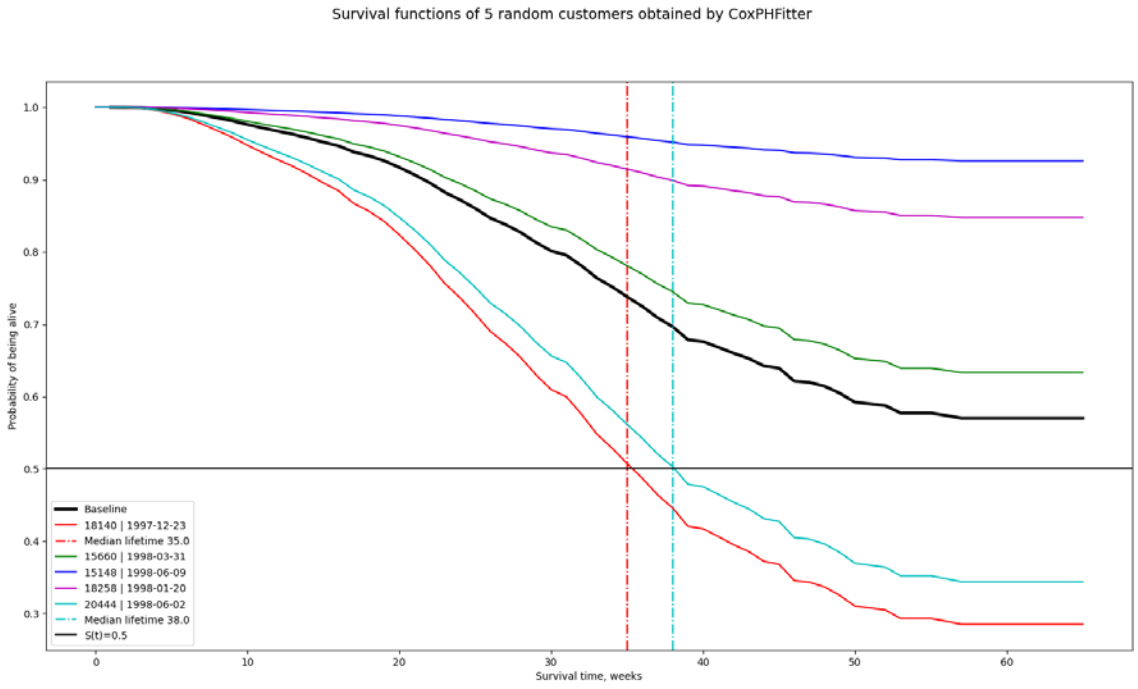

Figure 25. Baseline survival curve and five survival functions predicted by CoxPH for 5 randomly chosen individuals.

However, parametric model such as log-normal (shown on **Figure 26**) could extend survival function as long as necessary to determine expected and median survival times. We cut the plot at about 160 weeks for better illustration: all median survival times are clearly seen on the chart.

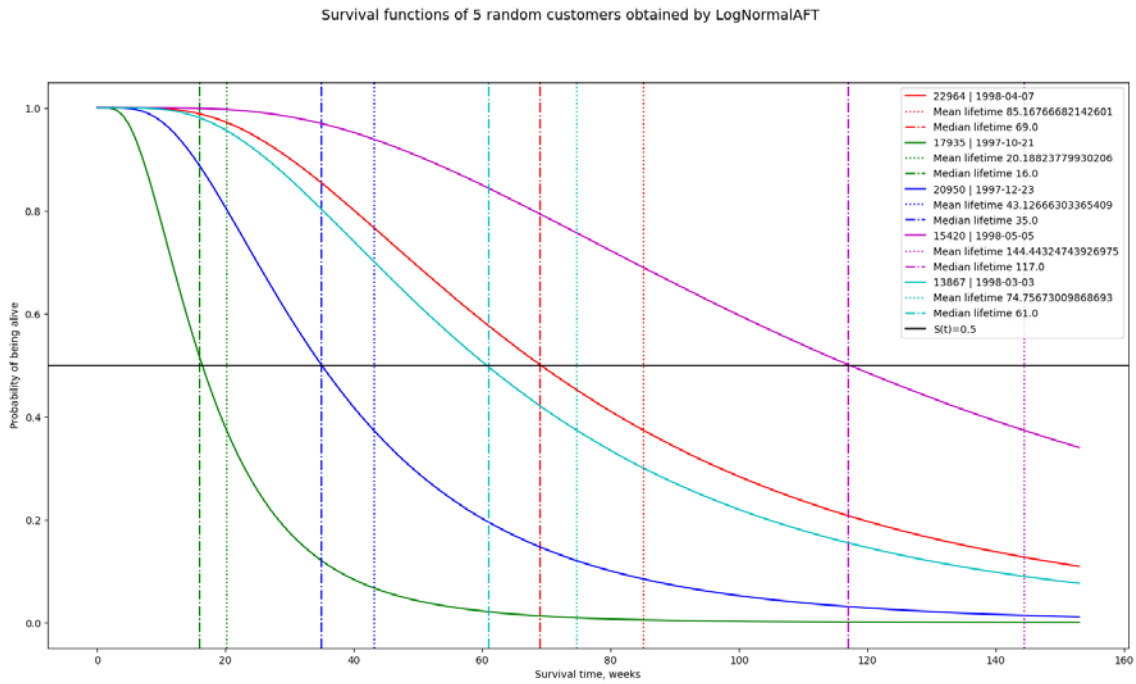

Figure 26. Five survival functions predicted by Lognormal AFT for 5 randomly chosen individuals.

Another example of survival curves for different 5 random customers produced by log-logistic AFT model is shown on **Figure 27**.

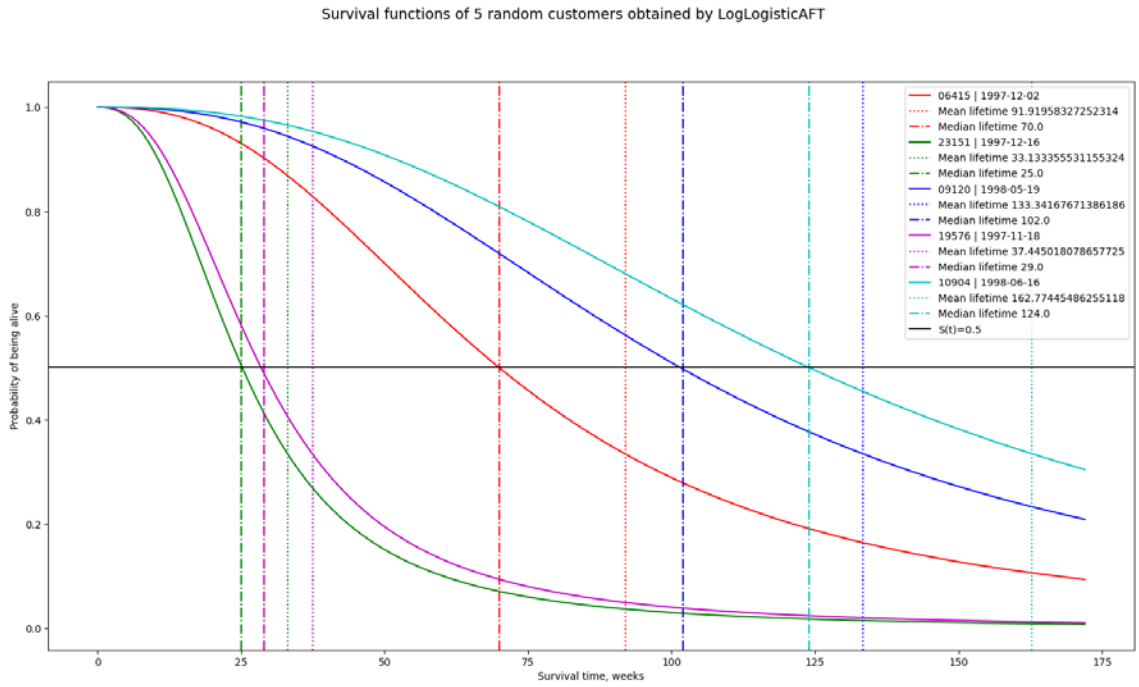

Figure 27. Five survival functions predicted by Log-logistic AFT for 5 randomly chosen individuals.

The highest values of CI, CI IPCW and AUC belong to GradientBoostingSurvivalAnalysis as well as the lowest IBS score. However, it experiences the same problem as Cox proportional hazard model or Kaplan-Meier estimator: survival curves often do not end up at zero or even undefined before it reaches 0.5.

Retail dataset

We will try our models on another publicly available dataset. It is similar to CDNOW but there are some differences. Transactions period starts at 2009-12-01 and ends at 2011-12-09. Study period starts at 2010-03-01and covers 92 weeks which is about 30% longer than CDNOW dataset has. We will be using the same discrete-time model described in previous parts with time unit equals to one week. After splitting data, training population size is 2,526 (380 dead and 2,146 censored) and test size – 631 (95 dead, 536 censored) observations, so total number of customers is three times smaller than in CDNOW dataset.

Univariate survival models

Following the same procedure as described for CDNOW dataset, we fit five univariate models and Kaplan-Meier statistic on training subset of retail dataset. Obtained survival functions are shown on **Figure 28**. Again, similarly to previous dataset, survival curves do not even cross the horizontal line y=0.5. This fact implies we will experience similar problems with survival time estimation for COX and GB models.

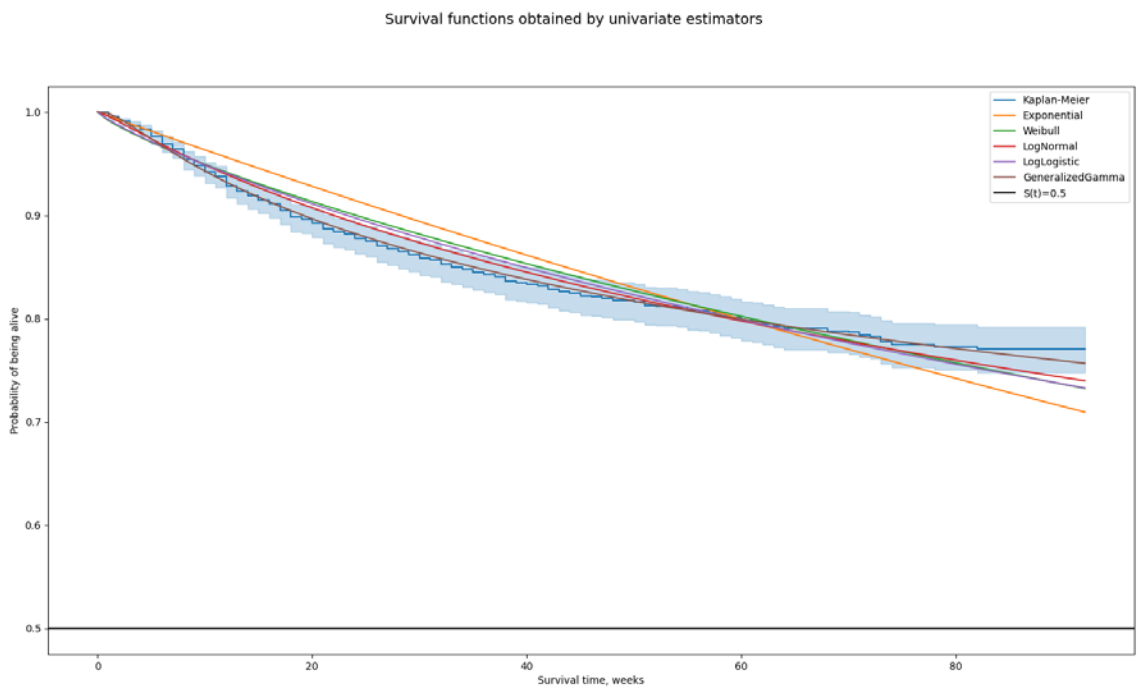

Figure 28. Survival functions estimated by five univariate models and Kaplan-Meier estimator on test subset of retail data.

According to metrics presented in **Table 6**, generalized gamma model is the best representation among univariate parametric models. Unfortunately, its median survival time for the population is 1,247 weeks which is far behind the value of study period (92 weeks) and much larger than in CDNOW dataset (197 weeks). Therefore, even if we have longer study period than in previous dataset (92 vs. 64 weeks), median lifetime seems to be longer as well.

	AIC	BIC	IBS	IAE Median	ISE Median	tExpected	tMedian
Exponential	5011.1352	5016.9696	0.1313	2.2066	0.0728	267.99	185.76
Weibull	4989.016	5000.6848	0.1309	1.3329	0.0267	437.79	247.41
LogNormal	4948.0844	4959.7532	0.1308	0.9323	0.0139	1705.3	387.84
LogLogistic	4978.3657	4990.0345	0.1309	1.2064	0.0227	1242.26	295.74
GeneralizedGamma	4929.2575	4946.7607	0.1308	0.3216	0.0017	3923.04	1247.1

Table 6. Summary of metrics obtained by five univariate models on training subset of retail data.

Heat map of feature correlations presented on **Figure 29** looks similar to CDNOW data, but there is at least one big difference: correlations between long and short trends of same features are much smaller than in previous experiment; one possible reason might be longer study period of retail data.

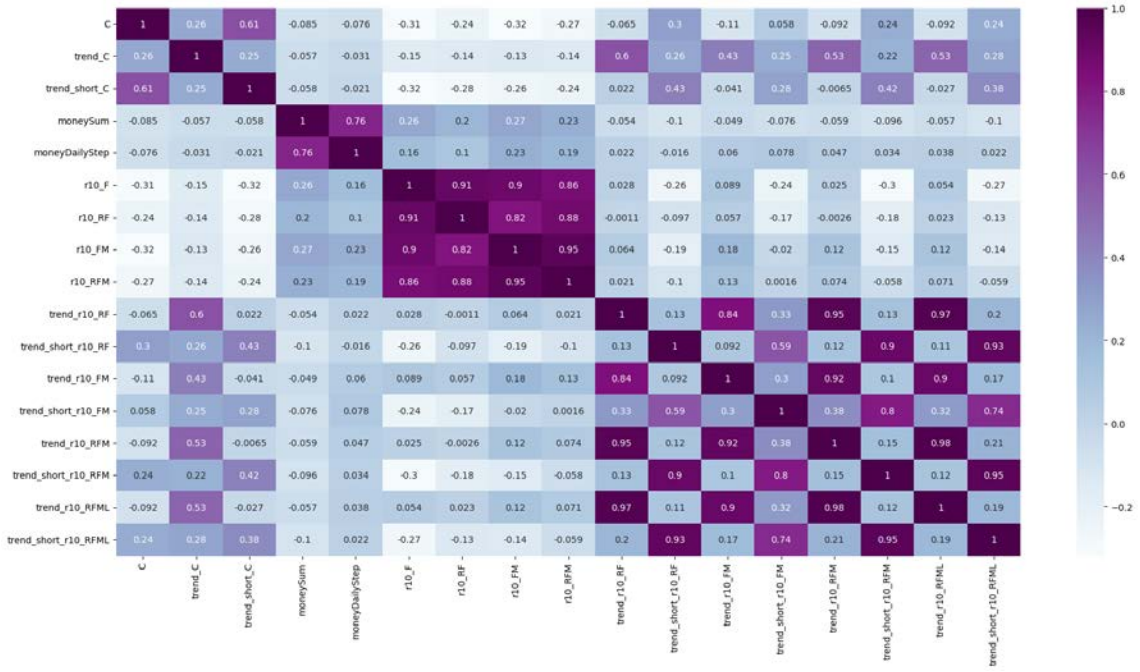

Figure 29. Pearson correlations calculated for selected features engineered from training subset of transactions of retail data.

Regression** survival **models

Following exactly the same procedure as for CNOW data, we fit same six regression models to retail training subset. Visualization of coefficients obtained by CoxPH models is presented on **Figure 30**. Their numerical values as well as lower and upper confidence intervals are in **Table 7**. Definitely, they are different than those that were obtained for CDNOW data. However, the following pattern can be observed: the group of features that have significant positive impact on hazard (by increasing the value of feature hazard becomes larger) remains the same: r10_FM, r10_F, trend_r10_RFM. We can say that the subset of features {r10_RF, moneySum, trend_short_r10_RFML} have large negative impact on hazard (decrease hazard rate) in CoxPH models for either Retail or CDNOW data. Coefficients determined by three parametric models Weibull, Lognormal and log-logistic AFT are shown on **Figure 31**, **Figure 32** and **Figure 33** respectively.

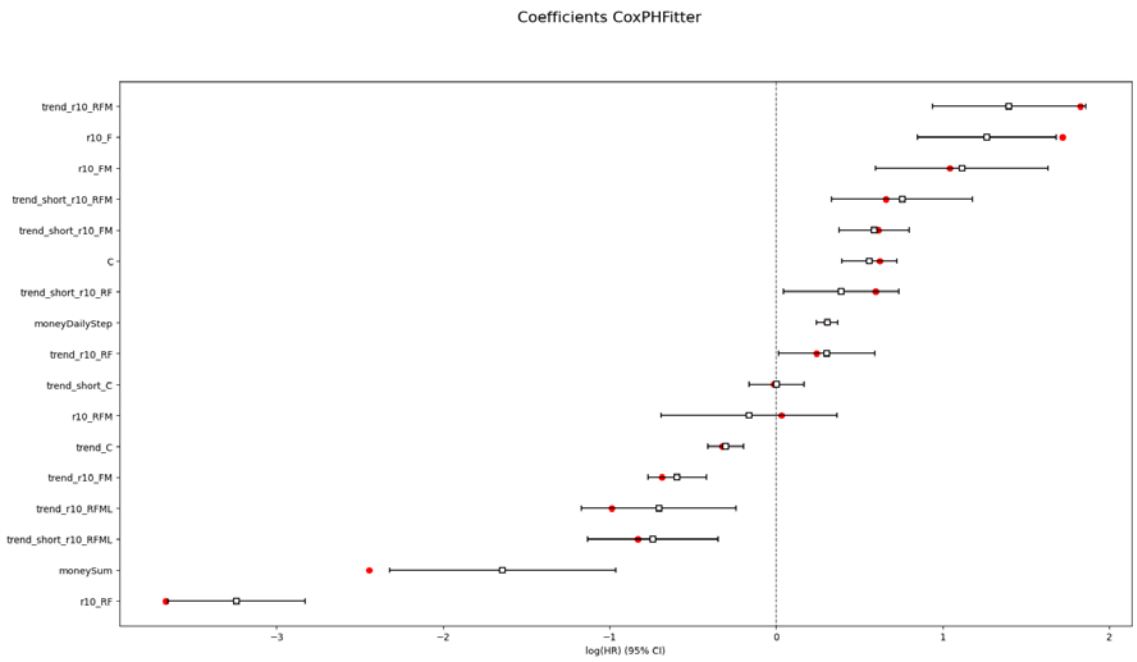

Figure 30. Coefficients of two different implementations of CoxPH. Black color corresponds to coefficients and confidence intervals for the model from 'lifelines', red dots show coefficients from 'scikit-survival'

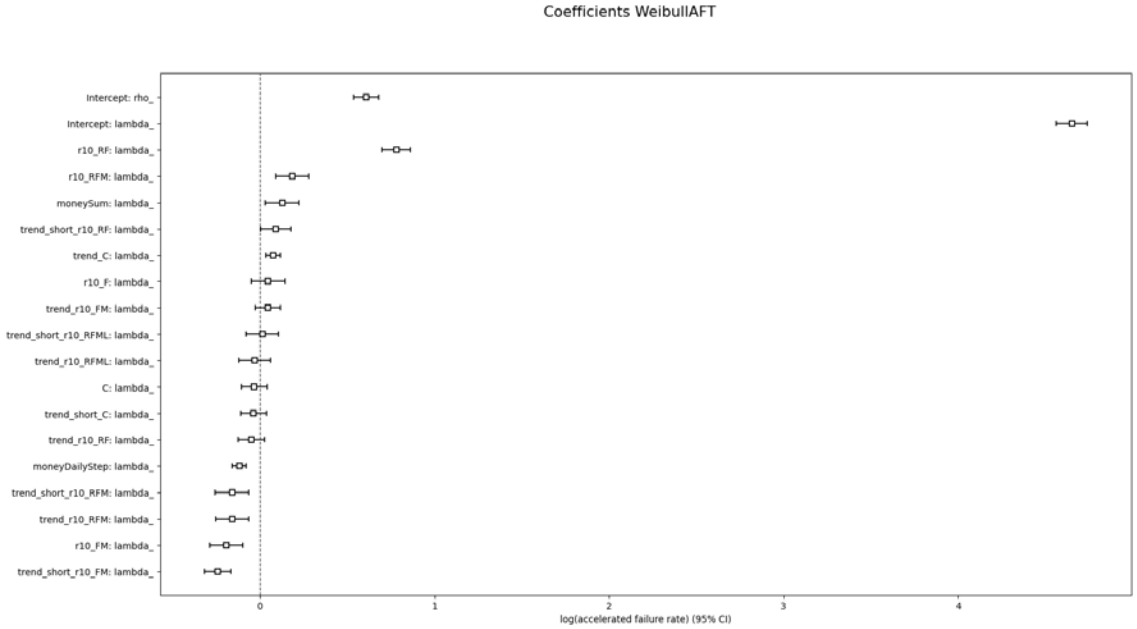

Figure 31. Coefficients determined by Weibull AFT model on training set of retail data.

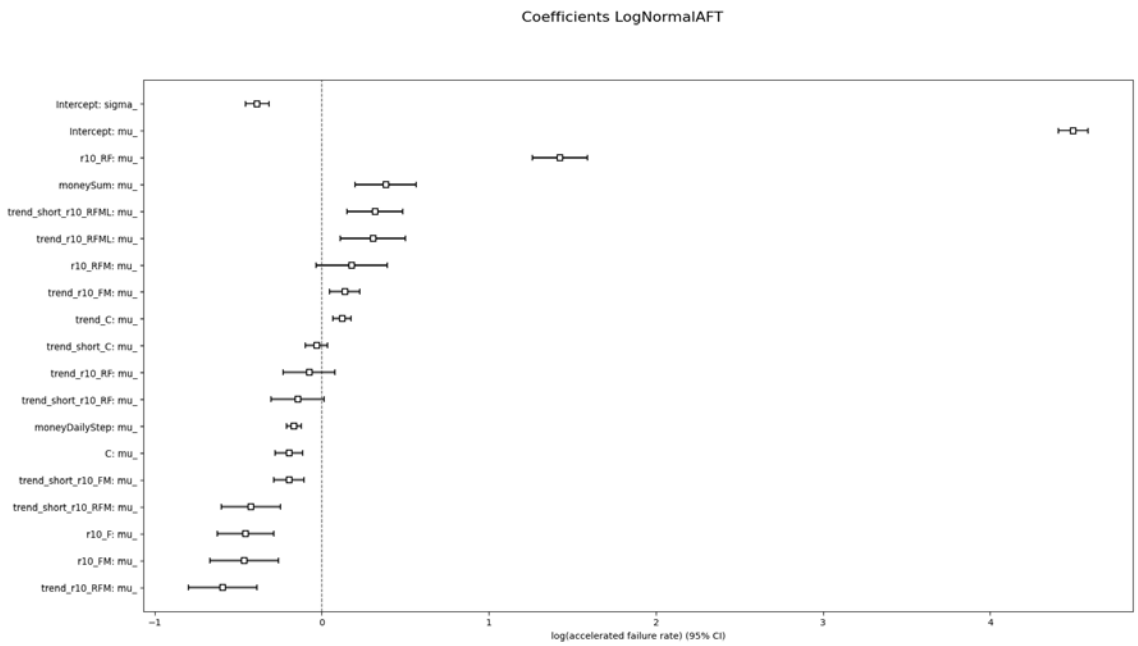

Figure 32. Coefficients determined by Lognormal AFT model on training set of retail data.

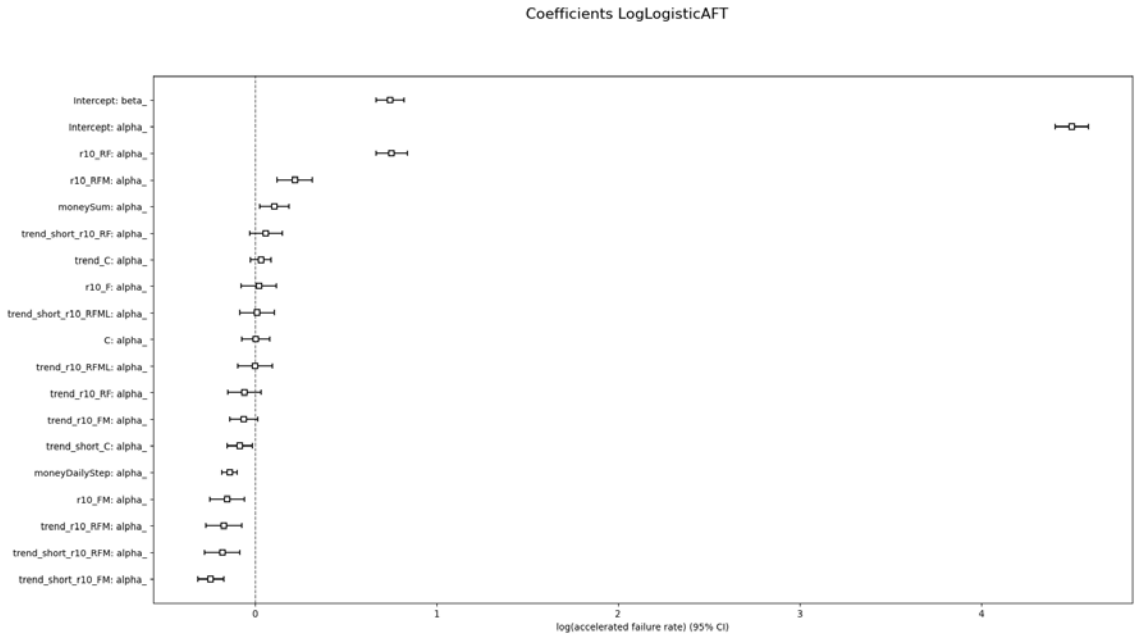

Figure 33. Coefficients determined by Log-logistic AFT model on training set of retail data.

covariate	coef	exp(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%
C	0.5600	1.7507	0.3936	0.7264	1.4824	2.067
trend_C	-0.3029	0.7386	-0.4089	-0.1969	0.6643	0.8212
trend_short_C	0.0036	1.0036	-0.1619	0.1691	0.8505	1.1842
moneySum	-1.6422	0.1935	-2.3204	-0.9641	0.0982	0.3813

moneyDailyStep	0.3056	1.3574	0.2417	0.3695	1.2734	1.4470
r10_F	1.2646	3.5419	0.8483	1.6810	2.3356	5.3710
r10_RF	-3.2405	0.0391	-3.6523	-2.8288	0.0259	0.0590
r10_FM	1.1152	3.0503	0.5979	1.6326	1.8183	5.1171
r10_RFM	-0.1628	0.8497	-0.6916	0.3660	0.5007	1.4419
trend_r10_RF	0.3015	1.3520	0.0128	0.5903	1.0129	1.8046
trend_short_r10_RF	0.3893	1.4760	0.0425	0.7362	1.0434	2.0880
trend_r10_FM	-0.5944	0.5518	-0.7709	-0.4178	0.4625	0.6584
trend_short_r10_FM	0.5876	1.7997	0.3762	0.7990	1.4568	2.2233
trend_r10_RFM	1.3974	4.0449	0.9369	1.8580	2.5521	6.4109
trend_short_r10_RFM	0.7555	2.1287	0.3315	1.1795	1.3930	3.2530
trend_r10_RFML	-0.7043	0.4944	-1.1679	-0.2407	0.3110	0.7860
trend_short_r10_RFML	-0.7403	0.4769	-1.1310	-0.3495	0.3226	0.7050

Table 7. Coefficients and upper and lower confidence intervals obtained by CoxPH from 'lifelines' library on training subset from retail data.

GB model does not have explicit parameters, but it provides the variable importance chart. One remarkable thing, that can be observed from results: feature r10_RF have the largest impact (or importance) for lifetime prediction in all regression models. Bar plot in **Figure 34** shows the importance of each feature that GB model used.

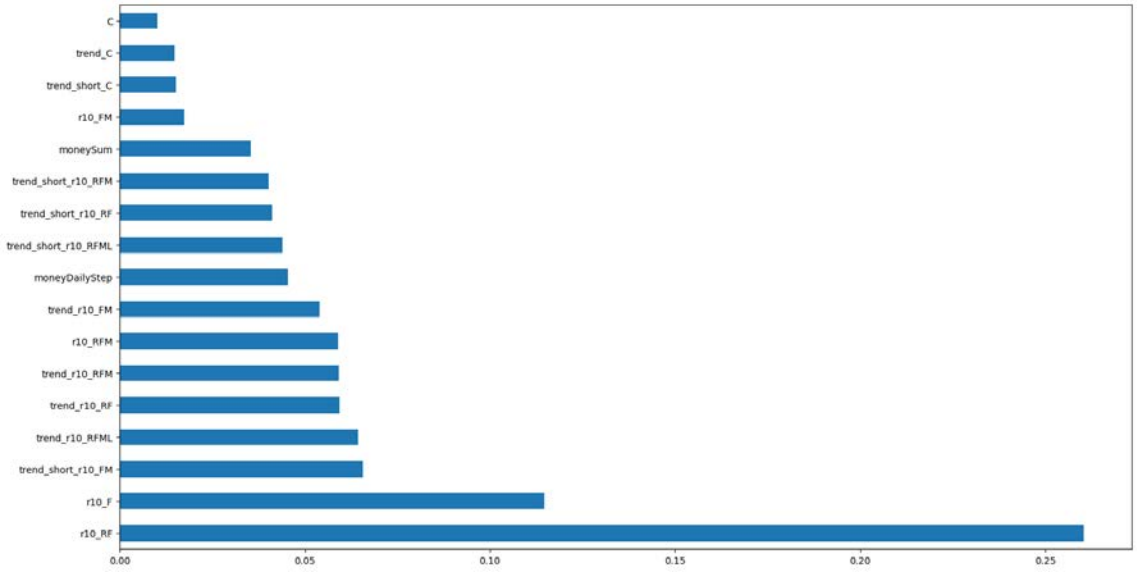

Figure 34. Bar plot of feature importance obtained by GB model on out of bag data after fitting the training subset of retail data.

The summary of metrics obtained by six regression models on Retail data is shown on **Table 8**. According to CI IPCW, the best model becomes GB again. However, it

experiences the same issue as described in CDNOW dataset analysis: GB as well as CoxPH models can predict lifetime for only small fraction on test subjects due to relatively short study period. This issue can be clearly seen on **Figure 35** that displays the distribution of predicted expected and median survival times obtained by each of six models. Boxes that correspond to CoxPH and GB models are very little, but those that are related to AFT models are much larger. Distributions of IAE and ISE scores are shown on **Figure 36**.

	GradientBoosting	CoxPHSurvival	CoxPHFitter	LogNormalAFT	LogLogisticAFT	WeibullAFT
CI	0.9773	0.9602	0.958	0.9576	0.9352	0.9278
CI IPCW	0.9769	0.9507	0.9462	0.9458	0.9028	0.8909
IBS	0.0273	0.0356	0.0367	0.0366	0.0427	0.0435
AUC	0.9878	0.9821	0.9813	0.9814	0.9732	0.9676
IAE Median	10.4663	13.4712	13.1956	13.8247	12.8929	12.1199
ISE Median	1.3086	2.2826	2.1565	2.3732	2.0686	1.8636
tExpected	15.25	19.04	18.85	180.08	207.32	140.02
sizeExpected	64	122	126	629	631	631
tMedian	20	30	29	124	99	98
sizeMedian	90	245	246	630	631	631
rankCI	1	2	3	4	5	6

Table 8. Summary of metrics obtained by six survival regression models on training subset of retail data. Column tExpexted contains mean expected survival time for population. Column sizeExpected shows number of individuals that corresponding algorithm succeeded to estimate. Column tMedian has median survival time for population estimated by corresponding algorithm. sizeMedian has numbers of individuals each algorithm was capable to predict.

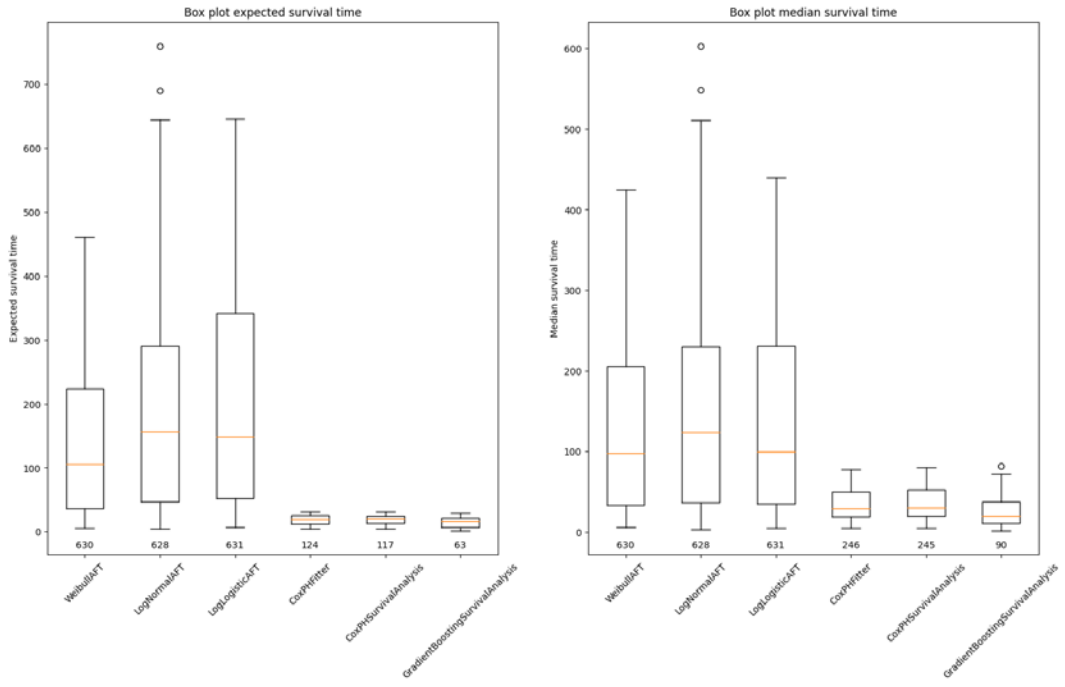

Figure 35. Box plot of distributions of expected (left) and median (right) survival times computed by six survival regression models fitted on training subset of retail data.

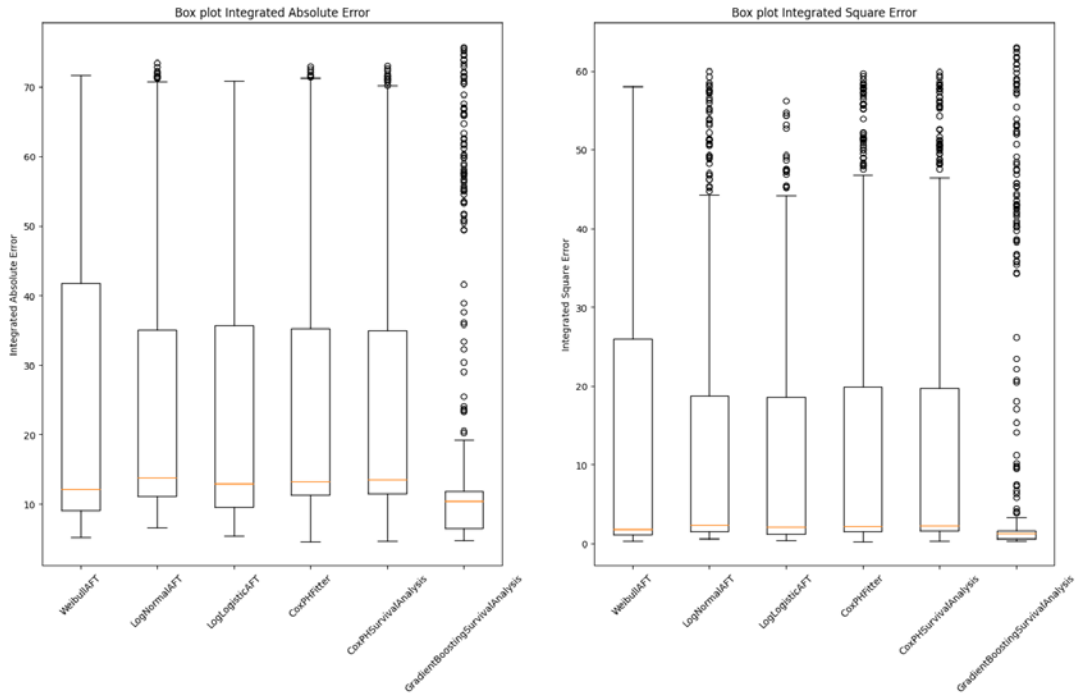

Figure 36. Box plot of distributions of IAE (left) and ISE (right) calculated from results obtained by six survival regression models fitted on training subset of retail data.

Chapter 4
Monte Carlo Simulations

Since both of our datasets have relatively small churn rate and short study period, we can try to simulate transactions data with relatively large study period and greater churn rate. We want a longer follow-up time, which would normally be available to a company, but was not released in the publicly available CDNOW and retail data, which will allow us to extend study period. Ideally, we expect to have a study period longer that the largest censored individual has. In this case, Kaplan-Meier survival curve should approach to zero, which means, we will be able to calculate both: expected and median survival times.

We hope that even if simulation cannot reflect all aspects of true data, it would be a good approximation and therefore would allow us to study models behavior on larger study period. Another useful aspect in simulation that we will have 'true' survival time even for 'censored' customers. Having this in mind we can use an additional metric such as mean (median) absolute error to compare how predicted remaining life for 'censored' customers differs from 'true' remaining life. The algorithm for simulation is written in. Using the synthetic data, we will also verify our assumption about customers 'death' event described in 'defining churn' sub-section.

Synthetic data generation

To simulate our synthetic data we will be using CDNOW dataset as a template. Three crucial characteristics: time between purchases, spending and loyalty (time interval from first to last purchase) were collected for all 7,296 customers of CDNOW dataset. The time between purchases looks like a nested list of the length, equals to number of customers from CDNOW data, 7,296 in our case. Each inner list corresponds to one individual and contains time intervals between consecutive purchases. Spending is a nested list of the length equal to 7,296; every inner list contains the money each customer spent at one cumulative daily transaction (if few purchases were made during one day by one client, it is considered as one buy and money sums). Therefore, the

length of every inner list is equal to frequency of its client. Loyalty is a sequence of the length equal to 7,296; each value corresponds to the time interval from customers' first to last transaction.

First we define some key parameters for our synthetic data:

a) *dStartTransactions* is a first date of transactions, equals to 2000-01-01.
b) '*avgLifetime'* is calculated as mean value of loyalty of all customers.
c) '*maxBirthDelay'* equals to 365 * 5 and determines the upper margin for dime lag that cohort appear uniformly on a period from *dStartTransactions* to *maxBirthDelay'*. It is designed to make new customers to appear not at one time instance, but distributed in time during 5 year period.
d) *nPoints* – max number of new customers we make. Must be smaller than 7,296.

We are going to model customers lifetimes by using exponential distribution which is in fact is the probability distribution of the time between events in a Poisson point process, where events occur continuously and independently at constant rate. This approximation might not reflect all aspects present in real data, but it might serve as a good approximation for clients' lifetimes in synthetic data. It has been found that exponential distribution naturally occurs in models that describe inter-arrival times in Poisson process. Therefore, we create 7,296 values for lifetime (7,296 equals to the number of customers from CNOW used as a templates for synthetic data), calculated as *round(rexp(7,296, 1/avgLifetime))* where *rexp()* is R's function for exponential distribution and 1/ *avgLifetime* is its only parameter (rate) which corresponds in our case to average rate of transactions' appearances within certain period. Those lifetime values are stored in array *churntime* of size equal to 7,296.

Then we define a birth lag for each of 7,296 data points calculated as *round(runif(7,296, min=0, max=maxBirthDelay))*, where *runif*() is R's function for uniform distribution.

To accomplish simulation of transactions for every new customer we need to simulate his IPI and spending for each purchase. As a template we are going to use the existing regular customers from CDNOW. Using log-normal distribution and MLE, for every CDNOW regular customer we determine two parameters for log-normal distributions, fitted in IPI and spending sequences (for one customer one model for IPI, another model for spending). Parameters μ and θ for IPI are stores in *muPurchase and sigmaPurchase* respectively: both arrays are of the same size equals to 7,296. Similarily parameters μ and θ for spending are stored in *muSpending* and *sigmaSpending.*

We take a sample of the size equals to *nPoints* from our 7,296 templates. In loop from 1 to *nPoints* we try to make IPI for new synthetic clients in the following way:

Get i_{th} lifetime previously drawn from exponential distribution

loyalty1 = churntime[i]

Get i_{th} parameters for IPI log-normal distribution

muPurchase1 = muPurchase[i]

sigmaPurchase1 = sigmaPurchase[i])

Get i_{th} parameters for spending log-normal distribution

muSpending1 = muSpending[i]

muSpending1 = muSpending[i])

Generate IPI for i_{th} observation by function genInterevents

IPI_1 = genInterevents(loyalty1, muPurchase1, sigmaPurchase1)

Where *genInterevents* is the following user-defined function:

```
genInterevents = function(span, mu, sigma){
  purchases=NULL
```

```
    newtime=NULL
    while(sum(c(purchases, newtime)) < span){
      purchases=c(purchases, newtime)
      newtime=ceiling(rlnorm(1, mu, sigma))
    }
    return(purchases)
  }
```

The function *genInterevents* takes three parameters as inputs:

1) span is customers lifetime taken from *churntime*
2) μ and θ are two parameters for IPI lognormal distribution, taken from *muPurchase* and *sigmaPurchase* arrays respectively.

genInterevents basically generates (if it can) the sequence of IPI for a new observation within time interval equals to *span.* Each IPI is drawn from log-normal distribution with μ and θ taken from i_{th} values of *muPurchase* and *sigmaPurchase.*

Having a sequence of IPI we need to generate the amount spent at each transaction by newly generated client. We take parameters from i_{th} records of previously created arrays *muSpending* and *sigmaSpending* and draw values from log-normal distribution to fill spending for a new customer. The R's function *round(rlnorm(length(purchases) + 1, muSpending[i], sigmaSpending[i]), 2)* does this job.

Following the described procedure we simulate one transactions dataset using *nPoints* equals to 5,000 and repeating the procedure five times. All produced records are merged into one data. We need this set to compare the structure of the simulated data with the original CDNOW. Then we simulate twenty transactions by described procedure: each simulated data will be used to fit / predict survival times by same models we described in Chapter 3. Twenty repetitions is low number compared to usual simulations, but the procedure of feature engineering is very time consuming. Each repetition taking close to two hours, even producing twenty of them took more than a

day. We found that twenty were enough to have a sense of variability and provide a realistic picture to describe our approach to solve the main problem of this work.

While using any synthetic data in this experiment for the analysis we split transactions interval of simulated data on two parts: training and hold out periods separated by date 2004-07-01. Data from holdout period is unseen for most of modeling parts except the one when we compare ‘true’ and predicted remaining life. We will be training our models only using training period before 2004-07-01 like the following after it transactions do not exist. So, training period starts at 2000-01-01 and ends at 2004-07-01. Study period starts at 2000-07-01 and has duration of 156 weeks. Hold out period starts just after 2004-07-01 and ends at 2013-09-12.

One simulation

We simulate one dataset with extended number of customers to verify how simulated data resembles the original one and what the difference is. During the training period (before 2004-07-01) 13,970 customers made at least three purchases (became regular and eligible for our model). We split this population into training (7,072 dead, 4,105 censored, 11,177 total train) and test (1,767 dead, 1,026 censored, 2,793 total test) subsets. Even if we simulated data that covers much longer time period and population has higher churn rate, it should resemble somehow the original one. For example, density histograms of inter-purchase intervals on **Figure 37**, as well as box plots of TTC and TTD on **Figure 38**, look very similar to the ones from CDNOW described in the section ‘choice of distribution’ of Chapter 3 in details.

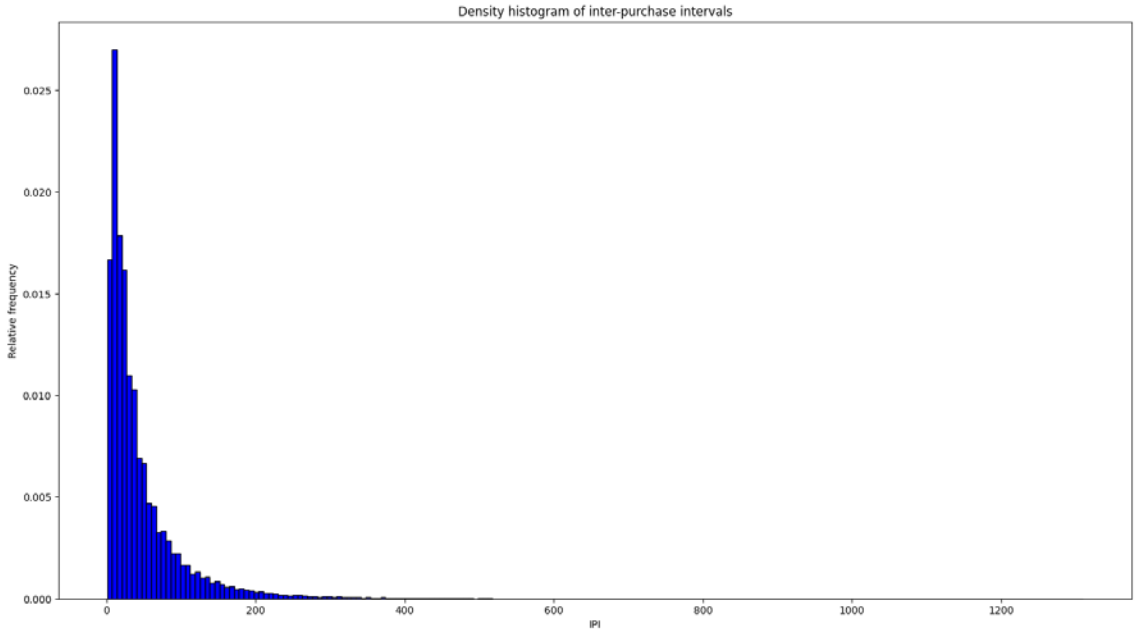

Figure 37. Histogram of IPI distribution for synthetic data.

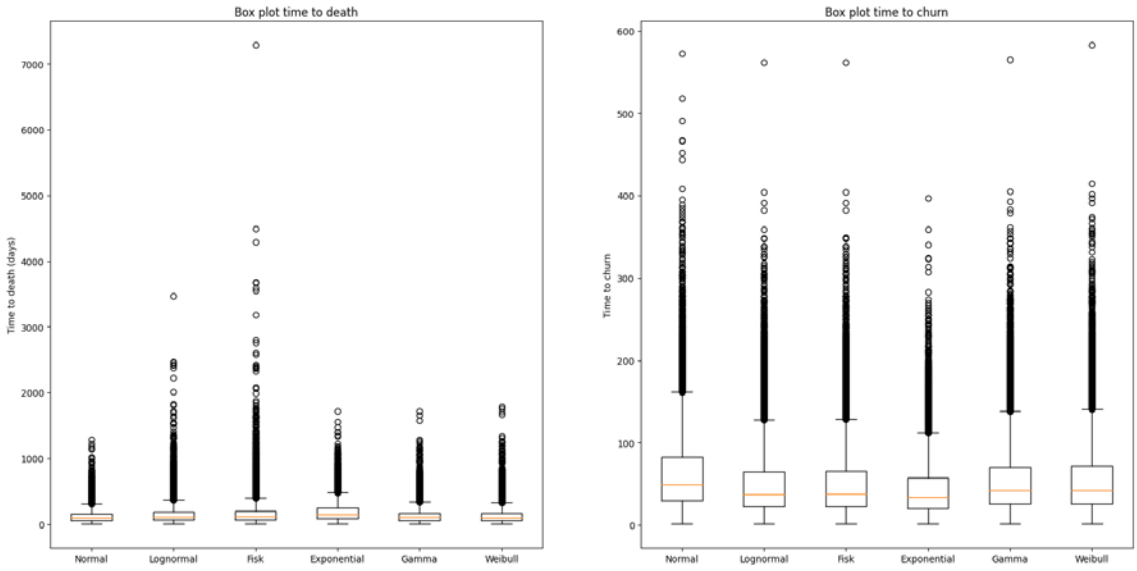

Figure 38. Distribution of TTD (left) and TTC (right) obtained by modeling critical events by six distributions described at the beginning of Chapter 3.

As for CDNOW data we use KS test and CvM criterion to determine the most suitable distribution (from set of six) for customers buying pattern. Box plot of highest p-values calculated by majority voting is shown on **Figure 39**. It looks like same plot from CDNOW data; Fisk distribution has about 60% of observations, so the choice becomes obvious.

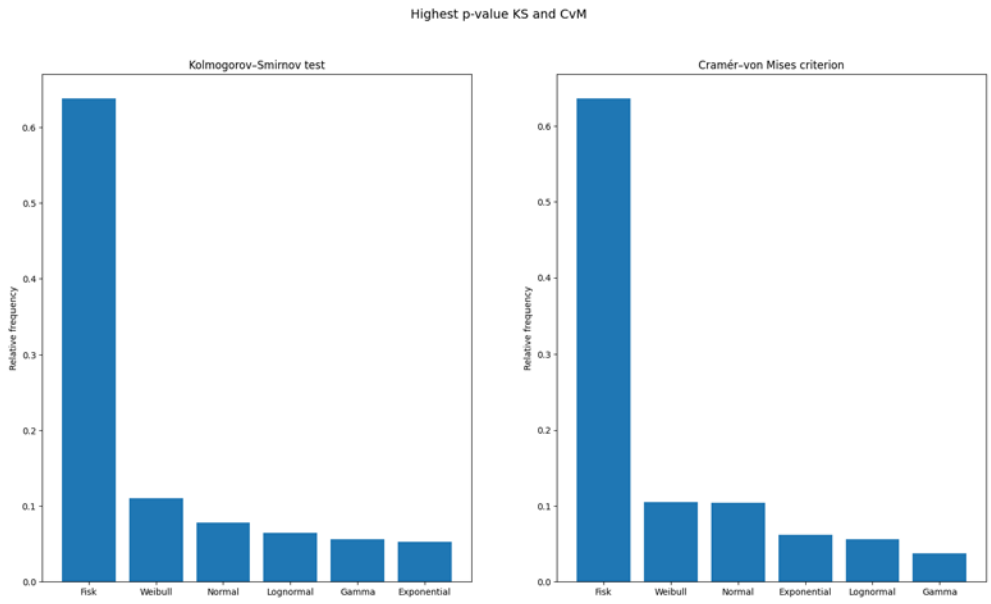

Figure 39. Bar plot of results of majority voting according to best p-value calculated by KS (left) and CvM (right) test statistics.

The interesting fact is that the correlation heat map shown on **Figure 40** looks more like the one from Retail dataset even if we used CDNOW as a template. Particularity, correlations between long and short trends are small; this might happen because Retail and simulated dataset have longer study intervals than CDNOW.

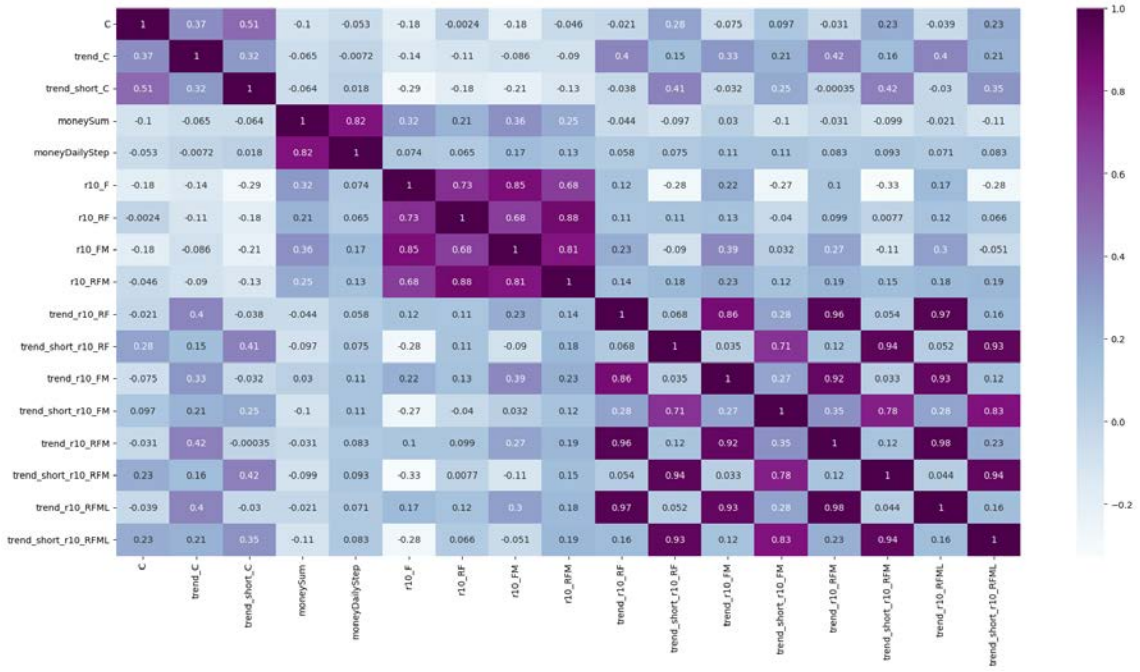

Figure 40. Heat map of Pearson correlations between features from synthetic data.

Univariate survival models

Starting with Kaplan-Meier estimator and univariate models we fit the training subset of our simulated data to see median survival functions for the population. Hopefully, this time all survival curves cross the horizontal line $y = 0.5$ and survival functions values approaches the value 0.1 at the end of study period. As can be seen on **Figure 41**, all six survival curves are close to each other. The summary of the coefficients of five univariate models are presented in **Table 9**.

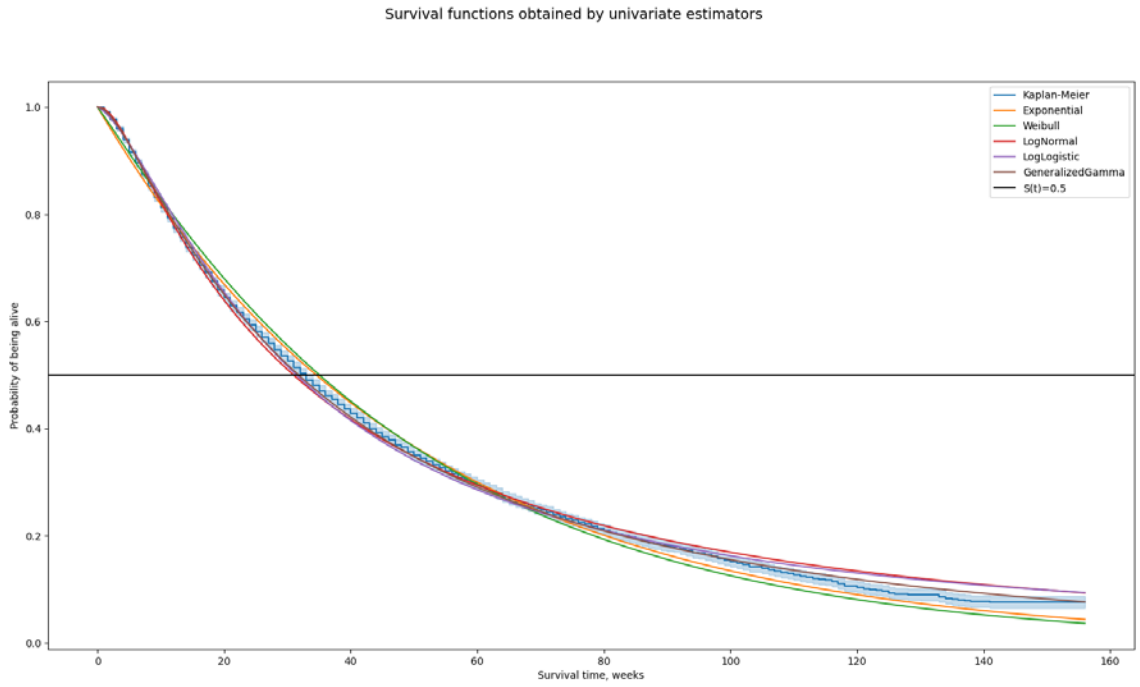

Figure 41. Survival functions obtained by five univariate models and Kaplan-Meier statistics for synthetic data.

Model	CoefName	coef	se(coef)	coef lower 95%	coef upper 95%
Exponential	lambda_	49.8415	0.5926	48.6798	51.003
Weibull	lambda_	49.7585	0.5656	48.6499	50.8671
	rho_	1.0462	0.0098	1.0270	1.0654
LogNormal	mu_	3.4328	0.0131	3.4071	3.4586
	sigma_	1.2236	0.0105	1.2029	1.2443
LogLogistic	alpha_	31.4204	0.4015	30.6333	32.2075
	beta_	1.4159	0.0137	1.3889	1.4428
GeneralizedGamma	mu_	3.5512	0.0226	3.5069	3.5955
	ln_sigma_	0.1557	0.0119	0.1323	0.1790
	lambda_	0.2377	0.0381	0.1630	0.3123

Table 9. Parameters of five univariate models obtained by fitting to synthetic data.

According to metrics from **Table 10**, generalized gamma model has lowest values for AIC, BIC, IBS, IAE and ISE and median survival time is near 32 weeks. This number is much lower comparing to the one from CDNOW dataset (197 weeks).

	AIC	BIC	IBS	IAE Median	ISE Median	tExpected	tMedian
Exponential	69432.7422	69440.0638	0.1568	2.4507	0.0460	49.84	34.55
Weibull	69411.9342	69426.5774	0.1571	3.3986	0.0877	48.87	35.05
LogNormal	69069.01	69083.6532	0.1563	2.3107	0.0506	65.47	30.97

LogLogistic	69151.2613	69165.9045	0.1562	2.1811	0.0444	81.74	31.42
GeneralizedGamma	69032.7416	69054.7064	0.1561	1.0731	0.0119	57.67	31.76

Table 10. Summary of metrics obtained by five univariete models on test subset of synthetic data.

Regression survival models

As in the experiment for real data, we fit six survival regression models to training subset of synthetic survival dataset. Coefficients for two CoxPH, Weibull, Lognormal and Log-logistic regression models are presented on **Figure 42**, **Figure 43**, **Figure 44** and **Figure 45**.

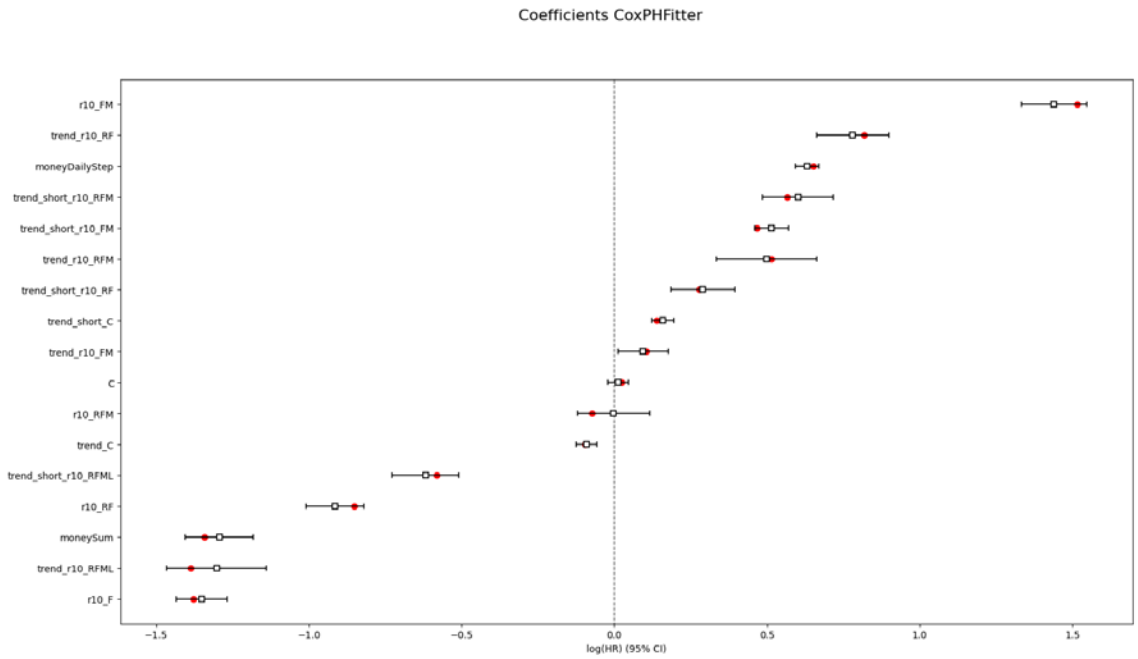

Figure 42. Coefficients of two different implementations of CoxPH. Black color corresponds to coefficients and confidence intervals for the model from 'lifelines', red dots show coefficients from 'scikit-survival'

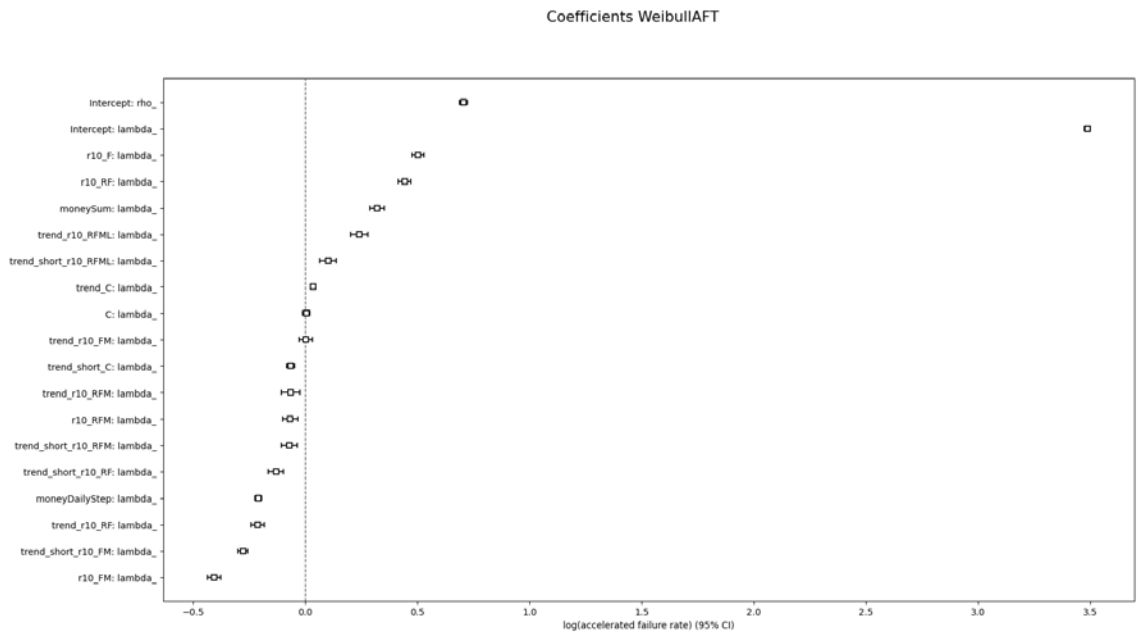

Figure 43. Coefficients determined by Weibull AFT model on training set of synthetic data.

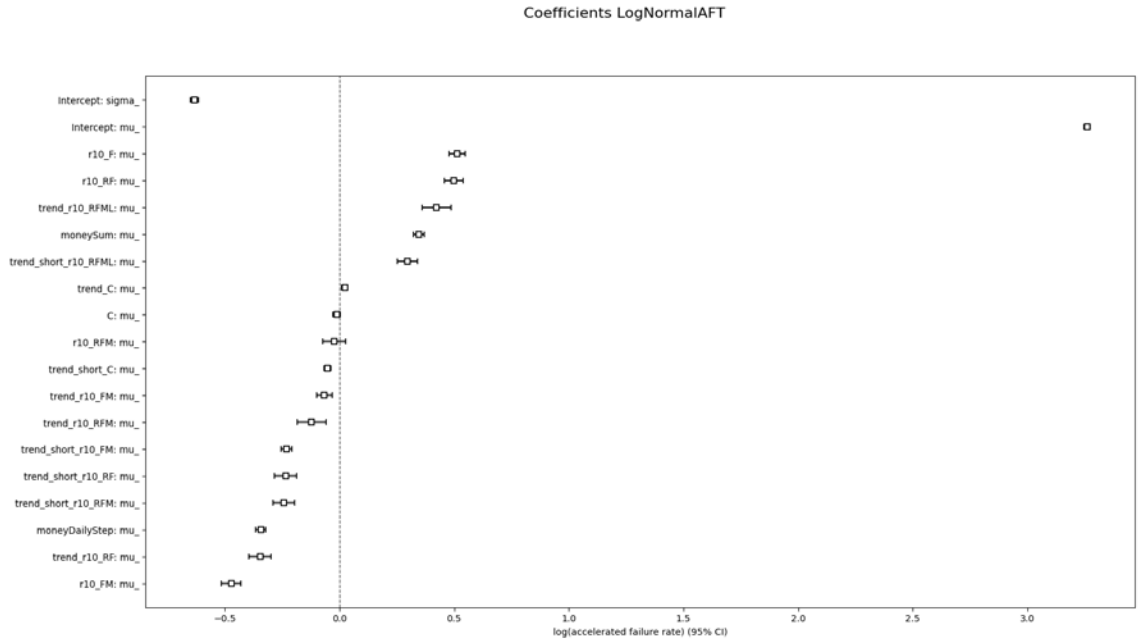

Figure 44. Coefficients determined by Lognormal AFT model on training set of synthetic data.

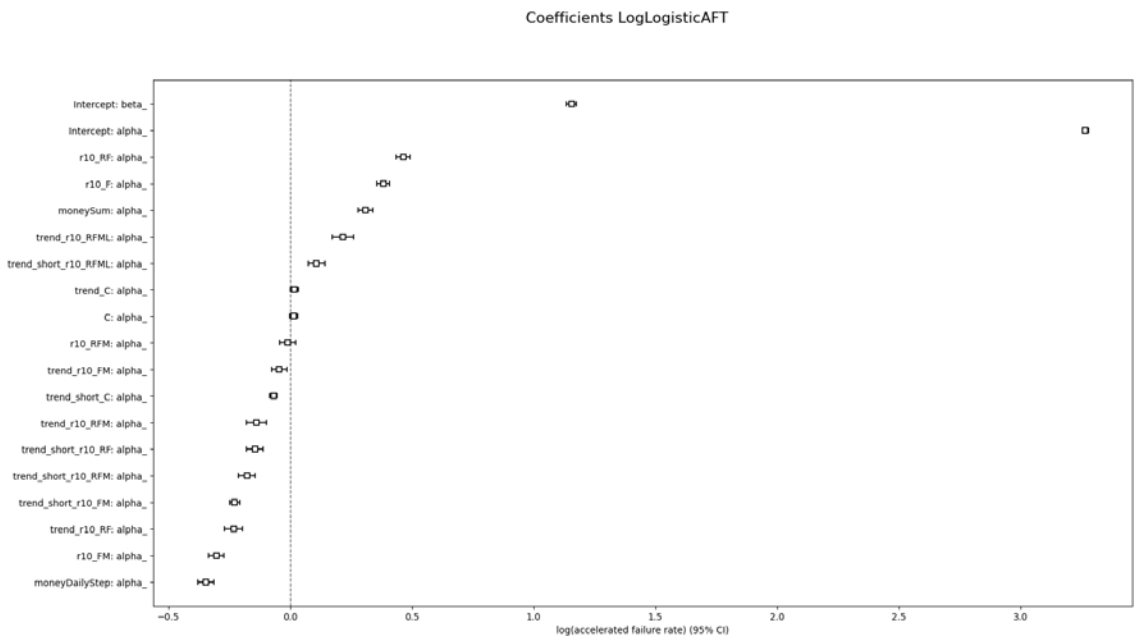

Figure 45. Coefficients determined by Log-logistic AFT model on training set of synthetic data.

Surprisingly, two predictors that we used as examples to illustrate how they make an influence on survival time in previous section do not show significant impact in simulated data. **Figure 46** and **Figure 47** show that features ‘C’ and ‘r10_RFM’ have no or tiny impact on survival time of subjects from population.

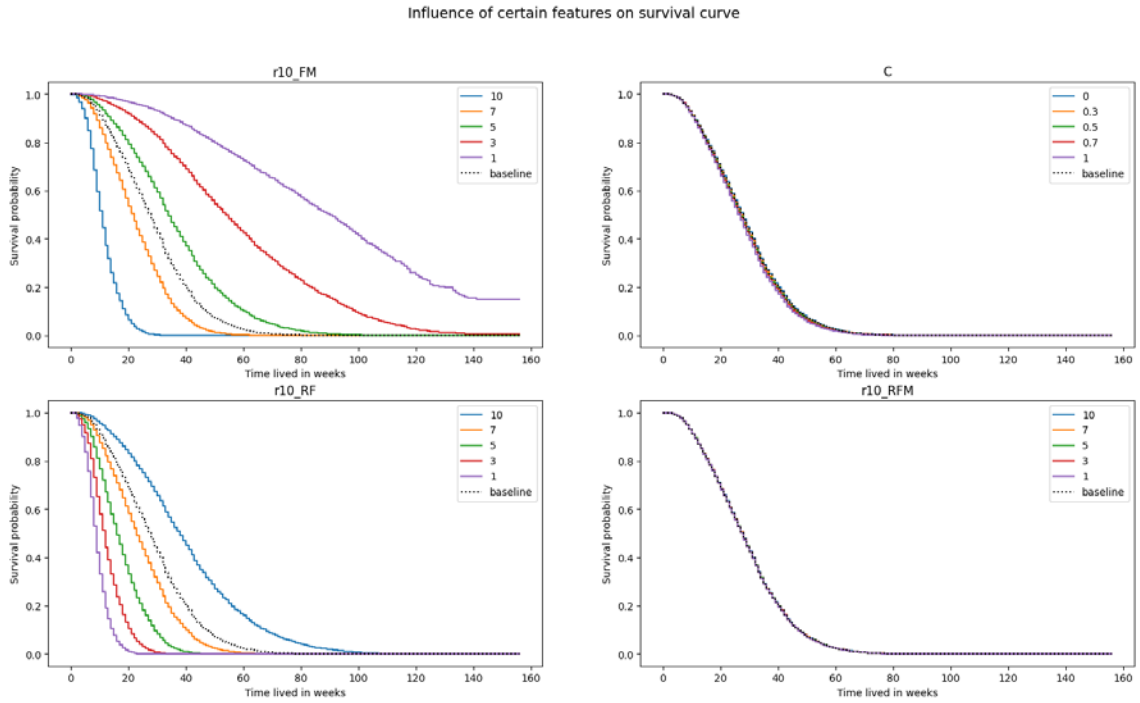

Figure 46. Partial impact of certain features on survival time of CoxPh model. Top left plot demonstrates negative impact on survival time, bottom left – positive impact. Features shown on two right plots do not make significant influence on survival function.

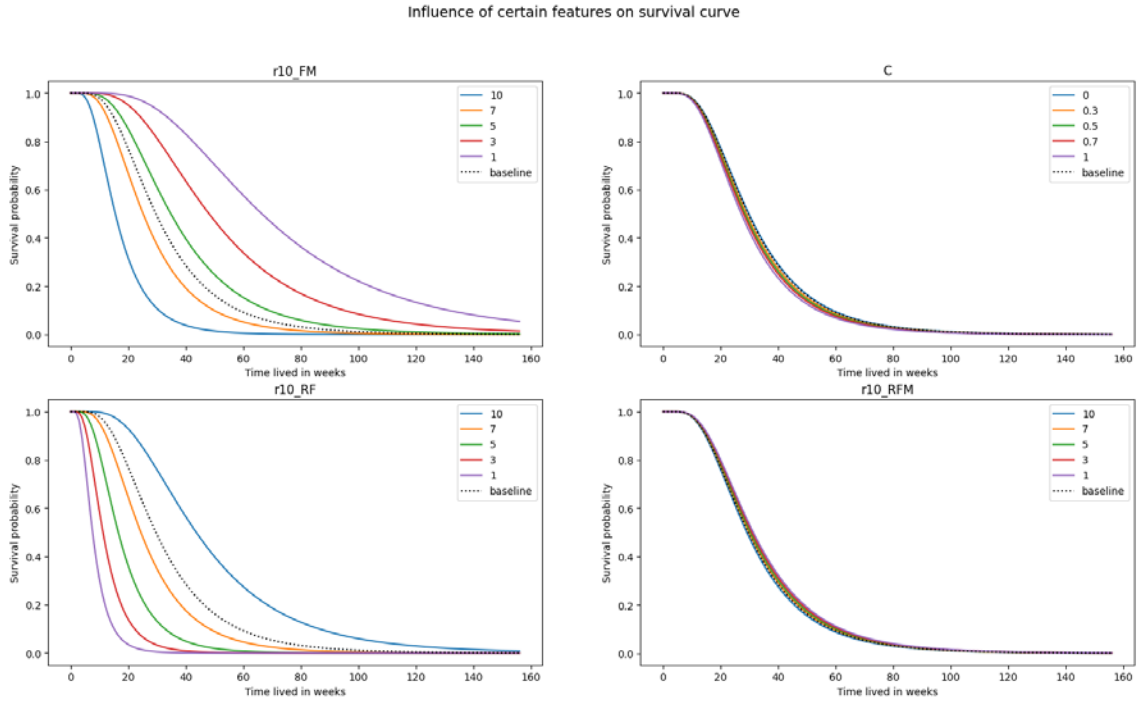

Figure 47. Partial impact of certain features on survival time of Lognormal AFT model. Top left plot demonstrates negative impact on survival time, bottom left – positive impact. Features shown on two right plots do not make significant influence on survival function.

Summary of metrics for six survival regression models are shown on **Table 11**. The remarkable difference from any of real data is that CoxPH and GB models are capable of predicting survival time for almost all censored customers. This is a result of longer study period which is one of our purposes to perform this simulation.

	GradientBoosting	LogLogisticAFT	LogNormalAFT	CoxPHSurvival	CoxPHFitter	WeibullAFT
CI	0.9602	0.8865	0.8839	0.8786	0.8781	0.8728
CI IPCW	0.9512	0.8648	0.8597	0.853	0.8524	0.8468
IBS	0.0341	0.0696	0.0722	0.0765	0.0765	0.08
AUC	0.9906	0.9513	0.9495	0.9449	0.9447	0.9426
IAE Median	30.0173	26.6142	27.7178	28.6376	28.8477	26.4081
ISE Median	8.3600	6.3593	6.9167	7.5105	7.5618	6.4327
tExpected	27.21	42.54	41.9	25.05	25.13	39.36
sizeExpected	2476	2793	2793	2172	2170	2793
tMedian	27	28	28	27	27	29
sizeMedian	2777	2793	2793	2732	2731	2793
rankCI	1	2	3	4	5	6

Table 11. Summary of metrics obtained by six survival regression models on training subset of synthetic data. Column tExpexted contains mean expected survival time for population. Column sizeExpected shows number of individuals that corresponding algorithm succeeded to estimate. Column tMedian has median survival time for population estimated by corresponding algorithm. sizeMedian has numbers of individuals each algorithm was capable to predict.

Highest CI IPCW, AUC and lowest IBS has GB model. IAE / ISE values are slightly better for AFT models. All six models estimate median survival time in very close range from 27 to 29 weeks. **Figure 48** presents the distributions of expected and median survival times for test subjects predicted by six survival regression models. Numbers below boxes indicate the size of test individual for which survival time was successfully estimated by the model. Definitely, we have higher numbers for median survival time than for expected since survival curve has more chances to reach the value 0.5 than to drop up to conditional zero. However, we can conclude that study interval for this data is reasonably long to allow predictions for the majority of test customers.

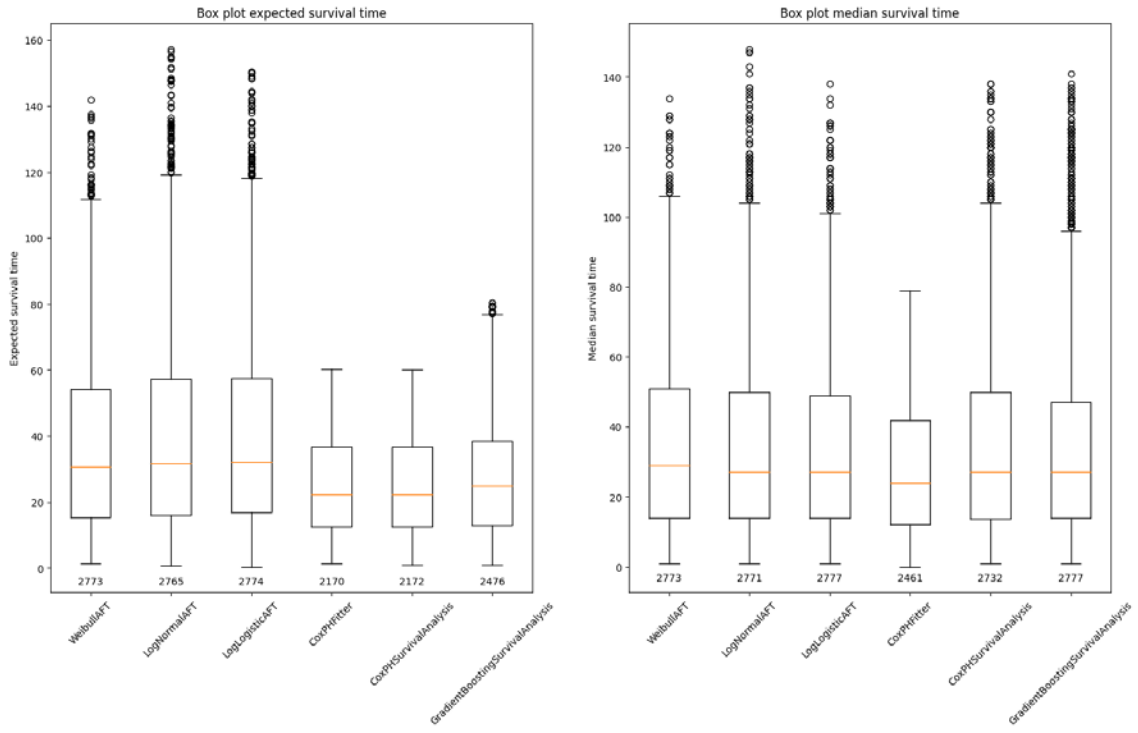

Figure 48. Box plots of distributions of expected (left) and median (right) survival time obtained by six regression survival models on test subset. Numbers below boxes indicate the quantity of observations corresponding model is capable to estimate.

Simulation data allows us to accomplish one additional task: measure the difference between true and predicted remaining life for test subjects since we 'know' the 'true' lifetime for everyone in population. We used MAE metrics to measure the difference since it provides an additional impression on how wrong the models are (it is easy to compare MAE with individual lifetime itself as well as with populations' lifetime and to see how reasonable the results are). For this metrics from the test subset we separate only subjects that were alive by the end of study period (1,026 individuals). **Figure 49** presents the distribution of absolute error for predictions of expected and median remaining life of 'alive' clients from the test subset.

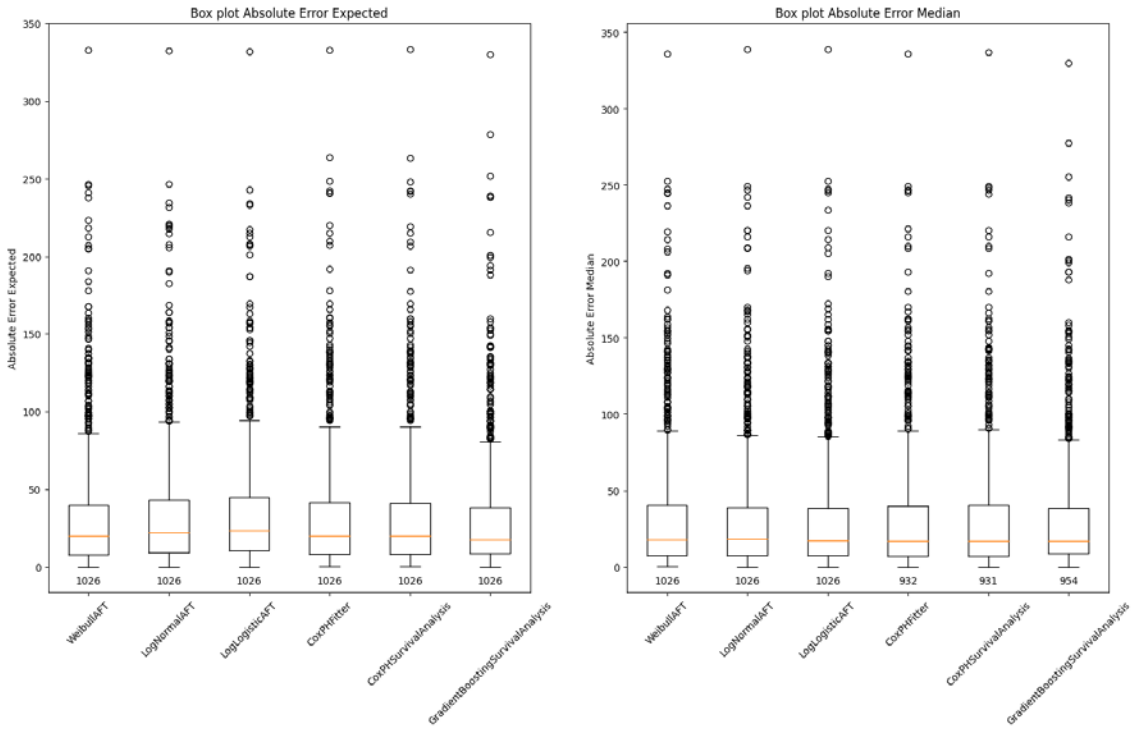

Figure 49. Box plot of distribution of absolute error between remaining expected life and 'true' values for six regression models obtained on test subset of synthetic data.

Table 12 summarizes results for all models we tried including univariate ones. It can be clearly seen that any regression model have lower MAE for median remaining time than any univariate, so we can conclude that our features (at least some of them) are valuable as predictors in this scope. It should be mentioned that for this experiment while calculating expected value we did not drop subjects whose survival curve did not reach zero value and performed integration to get the area under the survival curve up to the point where survival function ends regardless its last value, which should lead to underestimation of survival time. First three models (two COX and GB) have very close MAE, so 'best' model according to CI is still on top of ranking.

	MAE Expected	sizeExpected	MedianAE	sizeMedian	rankMedianAE
CoxPHSurvivalAnalysis	32.694699	1026	16.9047	931	1
CoxPHFitter	32.688452	1026	16.9627	932	2
GradientBoosting	31.581848	1026	16.9652	954	3
LogLogisticAFT	34.337111	1026	17.4313	1026	4
WeibullAFT	32.833293	1026	17.8611	1026	5
LogNormalAFT	34.00198	1026	18.3707	1026	6

Weibull	34.709638	1026	22.4704	1026	7
Exponential	35.704911	1026	23.3501	1026	8
GeneralizedGamma	47.072602	1026	24.7446	1026	9
LogLogistic	96.696889	1026	25.7	1026	10
LogNormal	60.257127	1026	26.3920	1026	11

Table 12. MAE and median absolute error between estimated and 'true' values of remaining life of test individuals obtained by five univariate and six regression survival models from synthetic dataset.

Simulation of twenty datasets.

To be more certain about accuracy of our estimations it is reasonable to create multiple simulated datasets and apply the same procedure described above for each simulation. Then, we can use an average of twenty simulations for each metric to have more accurate and less biased results.

Churn model assumption

In section 'definition of churn' from Chapter 1 we made few assumptions how to model customers' death for non-contractual business. For real data with short study interval it is hard to verify how our assumptions work and is there any violations of them (for example customer tagged as dead after certain period of time makes another purchase and becomes alive). For simulated data with longer study period and especially having hold-out period this verification becomes trivial. **Figure 50** plot shows the distribution of number of customers (left) that violate our churn definition described in previous sections. Corresponding distributions of fractions of violating customers (middle) and the populations size (right) are presented as well. In average, only about 27 customers that became 'dead' during the training period made at least one transaction during the hold-out period. Those clients are considered to be 'dead' according to our churn definition described in previous sections, but in fact, they are not. However, this is a relatively small fraction comparing to all customers in training period (less than 0.9% in average). We hope that this error will not significantly affect the accuracy of our modeling.

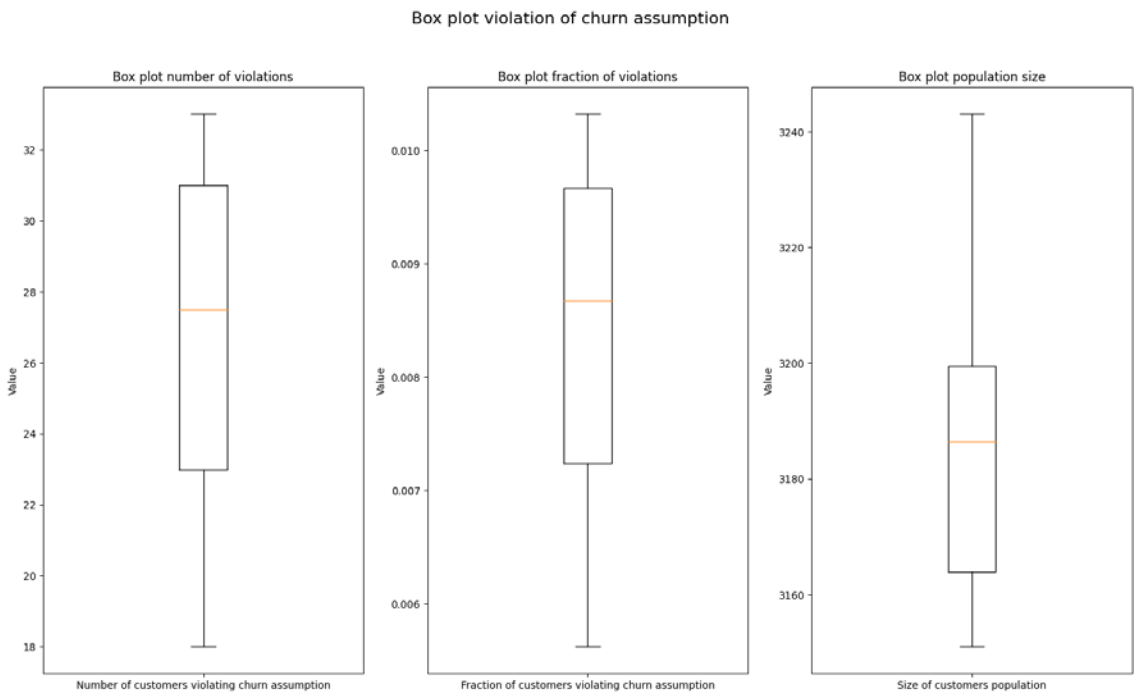

Figure 50. Distribution of number of customers that violate churn assumption (left). Distribution of fraction of customers that violate churn assumption (middle). Distribution of population size along 20 simulated datasets.

Univariate survival models

Starting from the simplest, we fit five univariate models to training subset of each of 20 simulated datasets and collect metrics obtained by making predictions on test subset. Average values for each of survival metrics we used in this work for univariate models are presented in **Table 13**.

Model	IBS	IAE Median	ISE Median	tExpected	sizeExpected	tMedian	sizeMedian
Exponential	0.1727	3.1243	0.0834	49.6645	393.05	34.425	393.05
Weibull	0.1731	3.6853	0.1126	49.03	393.05	34.7565	393.05
LogNormal	0.1717	2.2356	0.0517	65.923	393.05	30.639	393.05
LogLogistic	0.1717	2.0420	0.0462	83.2745	393.05	31.075	393.05
GeneralizedGamma	0.1718	1.3959	0.0236	59.6215	393.05	31.245	393.05

Table 13. Summary of average metric obtained by five univariate models on 20 simulated data.

Model that provides the survival curve closest to KM is generalized gamma. However, Lognormal has slightly better IBS. It should be mentioned that median survival time predicted by all univariate models lie within range from 30 to 35 weeks.

Results from the procedure of fitting six survival regression models are presented in **Table 14**. Highest CI and AUC and lowest IBS has GB model. Median survival time predicted by all regression models lie within the range from 27 to 30 weeks. Comparing to the range from univariate models it is shifted towards lower values.

	GradientBoosting	LogLogisticAFT	CoxPHSurvival	WeibullAFT	LogNormalAFT	CoxPHFitter
CI	0.9546	0.9079	0.9022	0.9015	0.8992	0.8816
CI IPCW	0.9373	0.8785	0.8696	0.8692	0.8691	0.8516
IBS	0.0486	0.0711	0.0775	0.0772	0.0753	0.0909
AUC	0.9855	0.9617	0.9562	0.9571	0.9575	0.9394
IAE Median	30.9445	29.5167	29.9632	30.0793	28.8173	31.5407
ISE Median	8.8052	8.1221	8.1537	8.3314	7.6053	9.3906
tExpected	29.0485	44.1365	24.1175	43.3645	44.0925	26.4135
sizeExpected	356.6	392.85	298.35	392.6	392.95	293.35
tMedian	27.325	27.4	27.2	27.9	28.25	29.625
sizeMedian	391.8	392.9	381.2	392.65	393	376.5

Table 14. Summary of average metric obtained by six survival regression models on 20 simulated data.

Distributions of CI IPCW and AUC are presented on **Figure 51**.

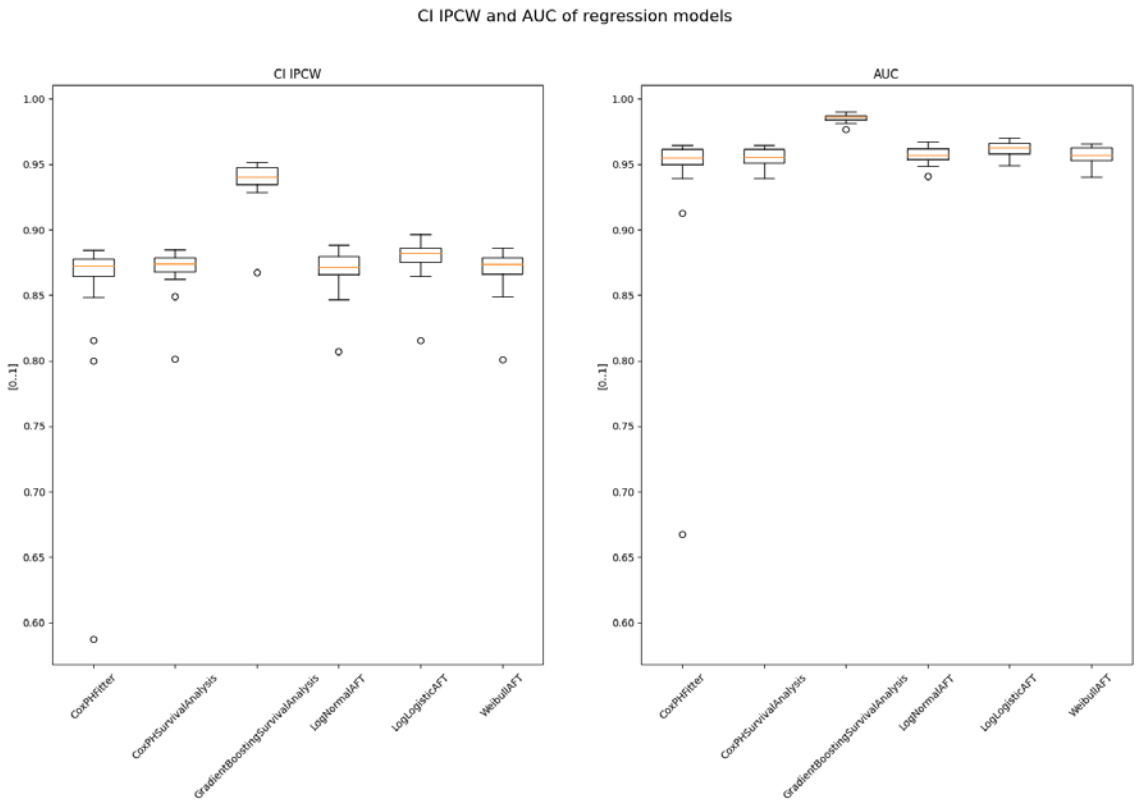

Figure 51. Distributions of CI IPCW (left) and AUC (right) of six survival regression models obtained on test subsets of 20 simulated data.

Distributions of expected remaining life obtained by five univariate models and six regression models on test subset over 20 simulations are presented in the form of box plots on **Figure 52**. Generally, expectations survival regression models are shorter than from univariate models. Among regression models, AFT group gives predictions of remaining life longer than the Cox group. Similar picture can be observed for median remaining life on **Figure 53**. Another observation that can be made from mentioned figures that expected remaining life usually greater in magnitude than median.

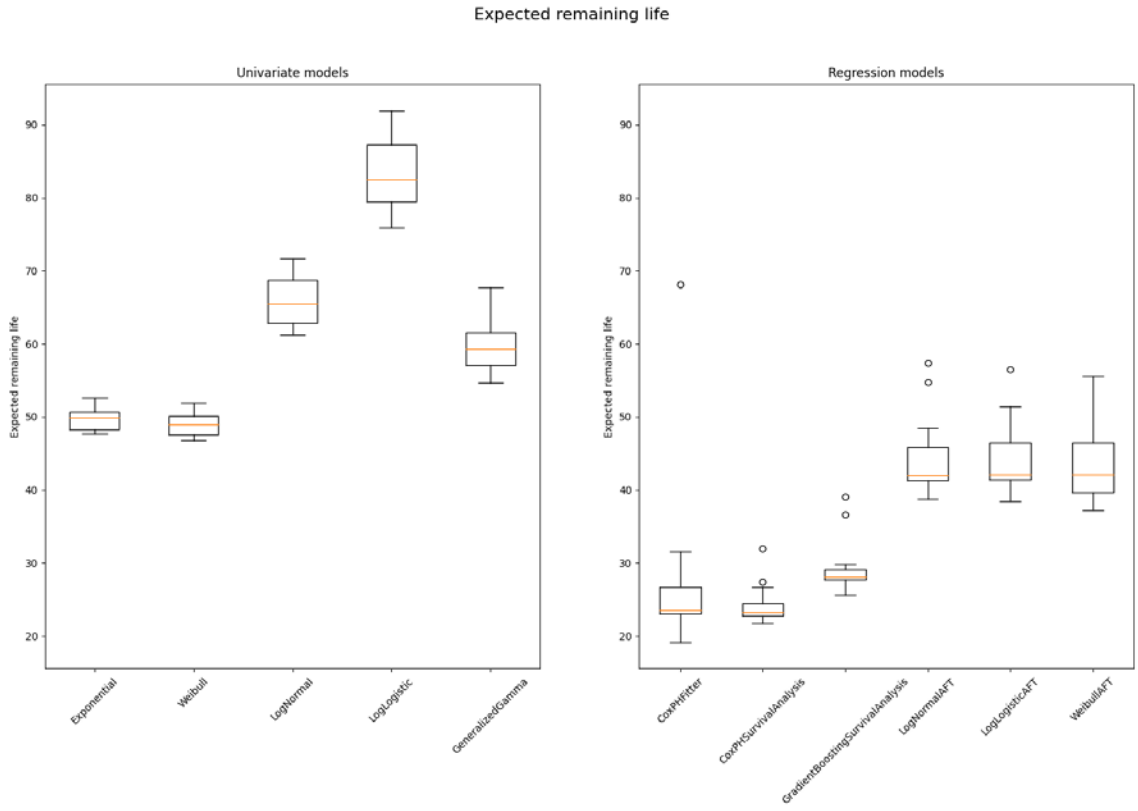

Figure 52. Distributions of expected remaining life obtained by five univariate models (lfet) and six survival regression models (right) on test subsets of 20 simulated datasets.

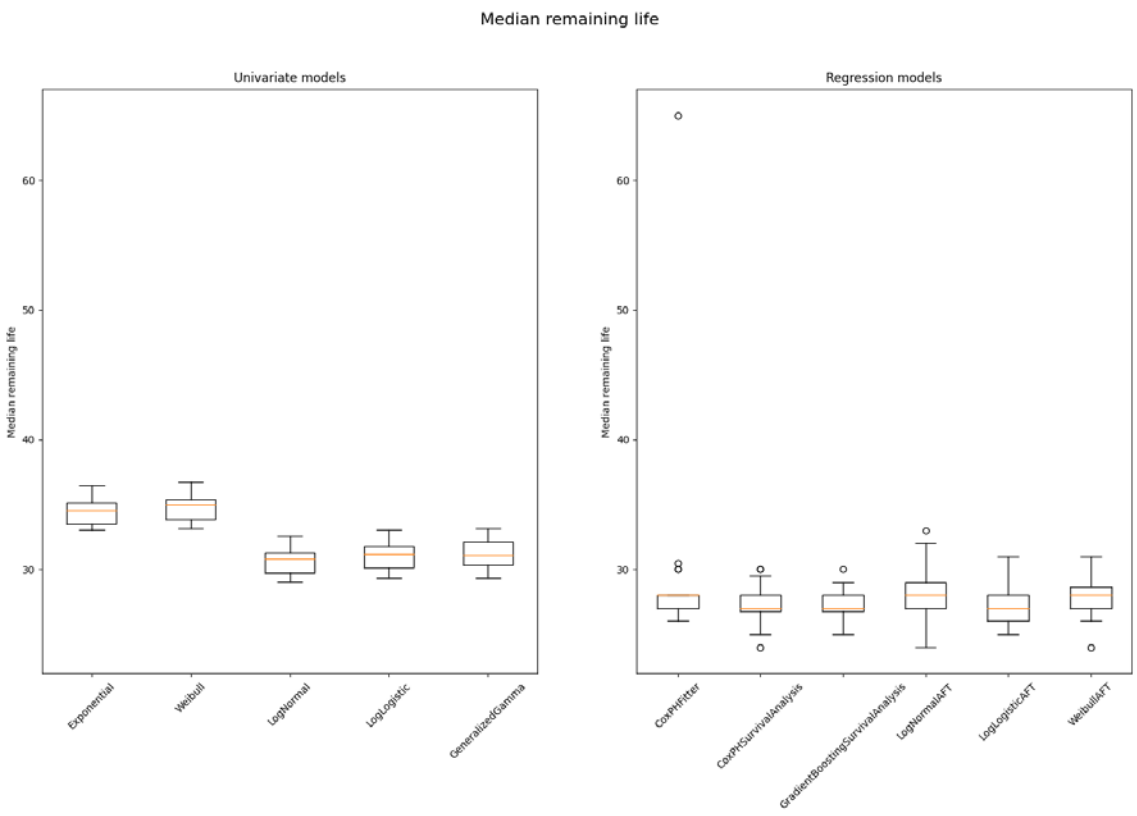

Figure 53. Distributions of median remaining life obtained by five univariate models (lfet) and six survival regression models (right) on test subsets of 20 simulated datasets

Predictions of remaining life from all models we use in our work are summarized in **Table 15**.

Model	MAE Expected	sizeExpected	MedianAE	sizeMedian
GradientBoosting	32.7486	143.85	16.8024	135.7
LogLogisticAFT	52.8302	144.1	17.5141	143.95
CoxPHSurvivalAnalysis	34.2882	143.85	18.5845	127.25
WeibullAFT	69.8723	144.1	18.7664	143.8
LogNormalAFT	44.7511	144.1	19.0490	144.05
CoxPHFitter	36.1321	144.1	19.9068	124.45
Weibull	36.1807	144.1	22.6170	144.1
Exponential	36.8842	144.1	23.1588	144.1
GeneralizedGamma	52.8980	144.1	26.2657	144.1
LogLogistic	105.2549	144.1	27.0938	144.1
LogNormal	64.2434	144.1	27.5028	144.1

Table 15. MAE and median absolute error between estimated and 'true' values of remaining life of test individuals obtained by five univariate and six regression survival models from 20 simulations.

Results are very similar to the ones obtained from first large synthetic dataset. However, order of models, sorted by median absolute error in descending order is slightly different. The top model becomes GB followed by Log-logistic AFT. Distributions of MAE between expected and 'true' remaining life are shown on **Figure 54**. Similarly, median absolute errors between median remaining life and true values are on **Figure 55**.

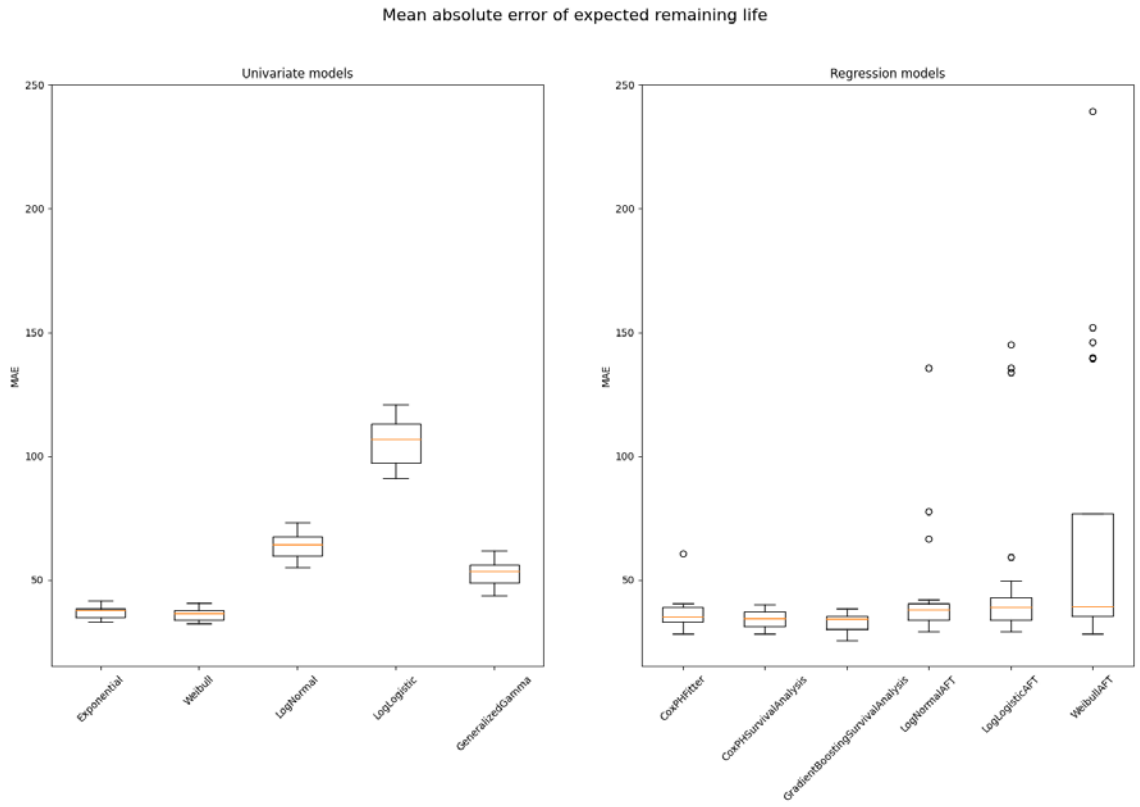

Figure 54. Distributions of MAE between expected and 'true' remaining life obtained by five univariate models (left) and six survival regression models (right) on test subsets of 20 simulated datasets.

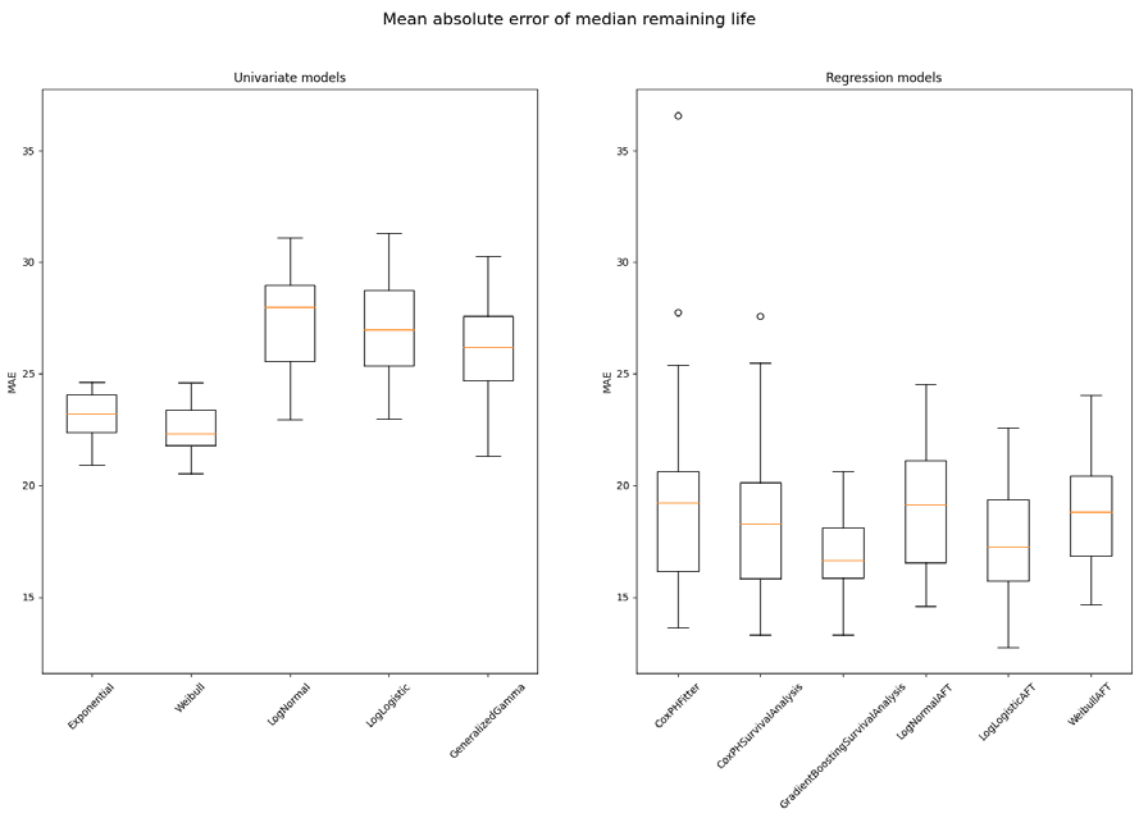

Figure 55. Distributions of median absolute error between median and 'true' remaining life obtained by five univariate models (lfet) and six survival regression models (right) on test subsets of 20 simulated datasets

Prediction of future profit

Our model can be extended and provide an estimation of future profit of the company. Having estimated remaining life for existing customers and knowing their purchase history we can also estimate their daily future spending and therefore, companies profit until all existing customers will churn. With our simulated data we can compare two values: total 'true' spending and a sum of estimated spending of censored customers. The last value can be obtained by multiplying clients' daily spending on his estimated remaining life in days. In **Table 16** results obtained over 20 simulations are presented where expected remaining life is used to calculate CLV. Surprisingly, predictions made by univariate Weibull model appear to be the most accurate. However, error seems to be large relatively to the average value of 'true' CLV. Remarkable observation is that all AFT models significantly overestimate CLV, therefore, remaining survival life. GB model underestimates remaining life the most among all models.

Model	Predicted	True	AbsError
Weibull	213,815.	165,383.	492,29.
Exponential	223,778.	165,383.	583,94.
CoxPHSurvivalAnalysis	138,363.	165,383.	720,32.
CoxPHFitter	153,127.	165,383.	805,25.
GradientBoosting	62,998.	165,383.	102,385.
LogNormalAFT	290,153.	165,383.	184,494.
LogLogisticAFT	305,037.	165,383.	185,468.
GeneralizedGamma	372,577.	165,383.	207,193.
LogNormal	456,431.	165,383.	291,048.
WeibullAFT	446,353.	165,383.	312,416.
LogLogistic	713,950.	165,383.	548,566.

Table 16. Predicted furure profit from existing active customers. Expected remaining life is used as lifetime variable for CLV calculation. Average values over 20 simulations. Column Predicted contains predicted values of CLV of the population of censored customers. Column True contains average 'true' value of the sum of CLV of all censored customers. Column AbsError is the difference between sums of predicted and 'true' CLVs.

If we use median remaining life as lifetime measure for CLV calculations, results becomes better. As we can see from **Table 17**, lowest mean absolute error belongs to univariate generalized gamma model and this error is more than twice lower than the best one from **Table 16.** Both Cox models few times overestimate lifetimes.

Model	Predicted	True	AbsError
GeneralizedGamma	171,423.	165,383.	19,977.
Exponential	140,910.	165,383.	25,828.
LogNormal	187,212.	165,383.	27,001.
Weibull	136,609.	165,383.	28,982.
LogLogistic	190,511.	165,383.	29,426.
GradientBoostingSurvivalAnalysis	240,731.	165,383.	96,168.
LogNormalAFT	256,496.	165,383.	183,568.
LogLogisticAFT	270,332.	165,383.	184,994.
WeibullAFT	434,518.	165,383.	307,148.
CoxPHSurvivalAnalysis	687,348.	165,383.	521,965.
CoxPHFitter	729,850.	165,383.	564,466.

Table 17. Predicted furure profit from existing active customers. Median remaining life is used as lifetime variable for CLV calculation. Average values over 20 simulations. Column Predicted contains predicted values of CLV of the population of censored customers. Column True contains average 'true' value of the sum of CLV of all censored customers. Column AbsError is the difference between sums of predicted and 'true' CLVs.

Box plots on **Figure 56** and **Figure 57** show distributions of profit calculated by using expected and median remaining life.

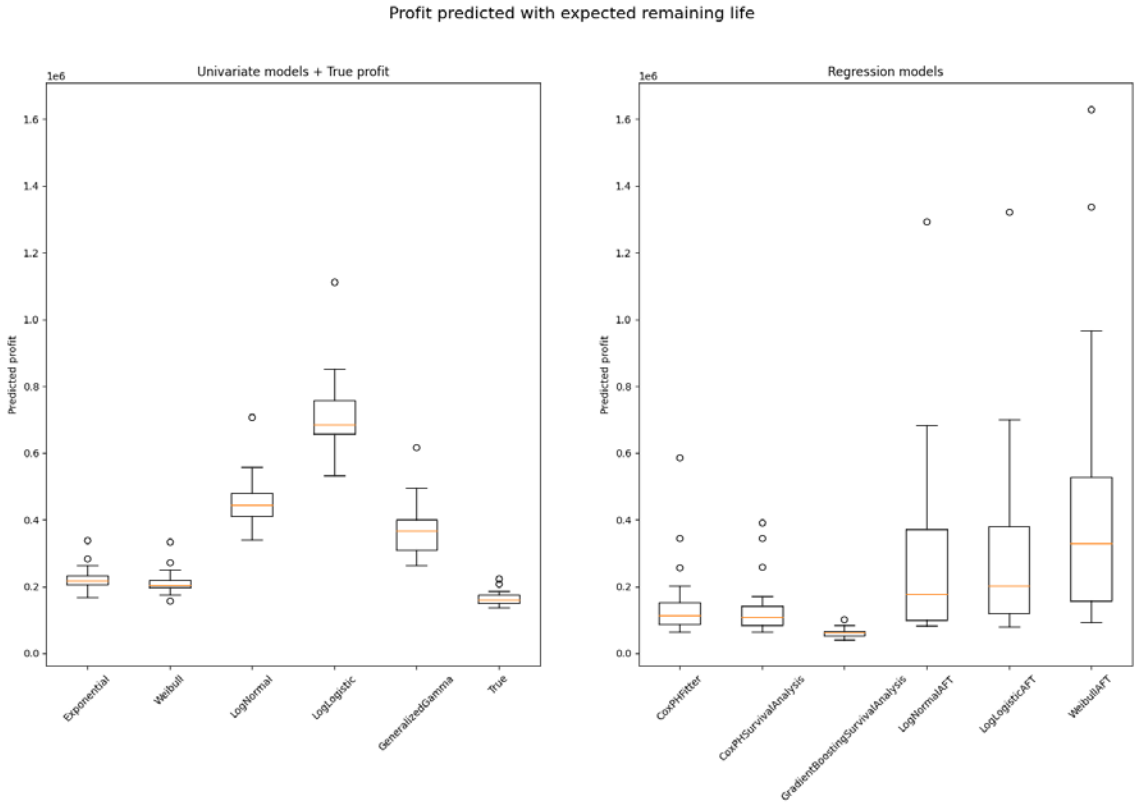

Figure 56. Box plot of profit distribution calculated from expected remaining life by five univariate models (left) and six regression models (right). On left plot, Label True corresponds to distribution of 'true' profit.

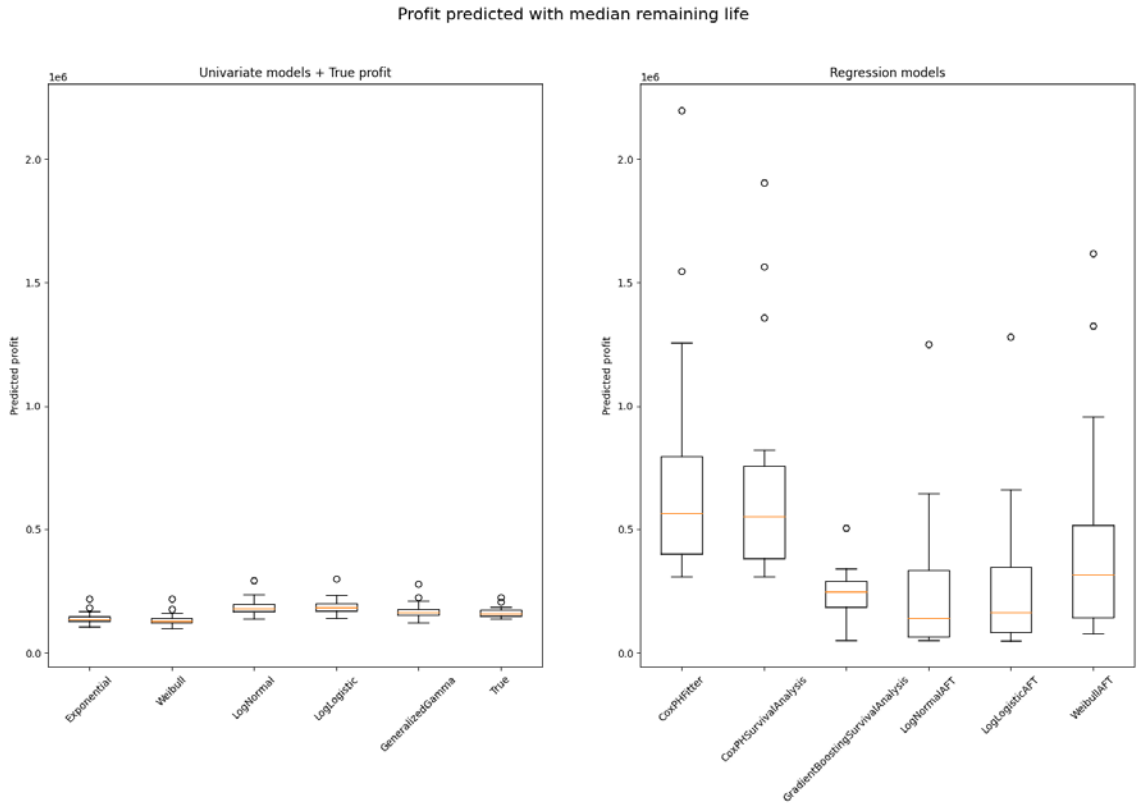

Figure 57. Box plot of profit distribution calculated from median remaining life by five univariate models (left) and six regression models (right). On left plot, Label True corresponds to distribution of 'true' profit.

Dazzle effect

In the introduction we mentioned that within our discrete-time framework features are time-varying. For every customer each feature actually is a sequence of values, each of them corresponds to different time instance within our framework. **Figure 4** from section discrete-time framework shows possible time instances where values for every feature can possibly be extracted. The length of longest sequence is equal to total number of time steps (states) within the observation period. This is the case when the long-living customer has his third purchase before the start of the observation period and 'alive' by the end of study. We have an intuition that features' history of changes could make an impact on customers remaining life.

However, usage of time-varying predictors in survival regression models is fraught with certain complications: in order to make predictions we need to know future values of time-dependent features. This time we try to step from survival concept and to

create artificial neural network regression model that predict survival time without taking into consideration the censorship. Plots presented on **Figure 5** and **Figure 6** show the history of change of few selected time-series predictors for one customer. Each feature on plots looks like time series data and in fact, it is. Having in mind, that we are going to use many features, one observation of the input data will be a 2D array of the shape [number of time steps, number of features]. Therefore, the input array has three dimensions and the shape [number of observations, number of time steps, number of features]. The outcome for an ANN model is the lifetime defined in the duration and censorship section, but for modeling we take only individuals that experience death within the study period.

It should be mentioned that customers have various observed lifetimes, therefore, number of time-steps can vary from one to number of time steps within the observation period (65 for CDNOW for example). One option is to use the entire sequences for predictions. However, depending on the duration of study period, the sequence length might be quite long, resulting the high memory usage and significant computational time. The question 'how long should customer event history be for customer churn prediction?' is explored in details by (Ballings & Van den Poel, 2012). We used similar methodology and found experimentally that the most recent fifty values for each feature that correspond to approximately one year of history (since our time step is equal to one week) is sufficient to achieve reasonable accuracy. Having in mind that customers have varying observed period, we perform zero padding to make the length of all sequences equals to fifty. Therefore, sequences that are shorter than 50 time-steps are padded with zero from the beginning until they reach the length equals to 50. Sequences longer than 50 are truncated from the beginning by keeping the most recent values and removing the oldest ones.

Recurrent neural networks (RNN[27]) add the explicit handling of order between observations when learning a mapping function from inputs to outputs. They are a type of neural networks that natively supports sequential input data and can exhibit temporal dynamic behavior. Instead of mapping inputs to outputs alone, the network is capable of

[27] https://en.wikipedia.org/wiki/Recurrent_neural_network (accessed on December 21, 2022)

learning a mapping function for the inputs over time to an output. This capability of RNN has been used to in complex natural language processing models such as translation where the complex inter-relationships between words should be learned within a given language and across languages in translating form. Such a capability of RNN can be used for other tasks where sequences of data are involved.

LSTM[28] is a type of RNN that has feedback connections and can process entire sequences of data. GRU[29] is a gating mechanism in RNN similar to LSTM, but it lucks an output gate and has fewer parameters than LSTM. Its performance on certain tasks is found to be similar to that of LSTM and even better on some small datasets. **Figure 58** shows the structure of the neural network we use for our regression model. Our ANN has two GRU layers followed by three dense layers and linear output. Few dropout layers[30] are included to prevent overfitting.

Procedure to obtain estimated values for remaining life is similar to survival regression models, except that we train ANN model on the subset that contains all dead 'customers' and make predictions for those who remains alive by the end of study period. Subset containing 'dead' customers is divided into to two groups: training and validation. Training data is used to fit ANN model, validation data has two purposes: it is used for early stopping (where ANN model should stop training) and to compare predicted and true values of the remaining life. MAE obtained on validation data is equal to 8.1 which is very low value comparing to results from survival regression. However, MAE calculated on predictions, made on test dataset that contains only censored clients, is 35.06 which is significantly larger. Median absolute error on test data is equal to 16 weeks which is comparable with best results obtained by survival models.

[28] https://en.wikipedia.org/wiki/Long_short-term_memory (accessed on December 21, 2022)
[29] https://en.wikipedia.org/wiki/Gated_recurrent_unit (accessed on December 21, 2022)
[30] https://keras.io/api/layers/regularization_layers/dropout/#:~:text=The%20Dropout%20layer%20randomly%20sets,over%20all%20inputs%20is%20unchanged. (accessed on December 21, 2022)

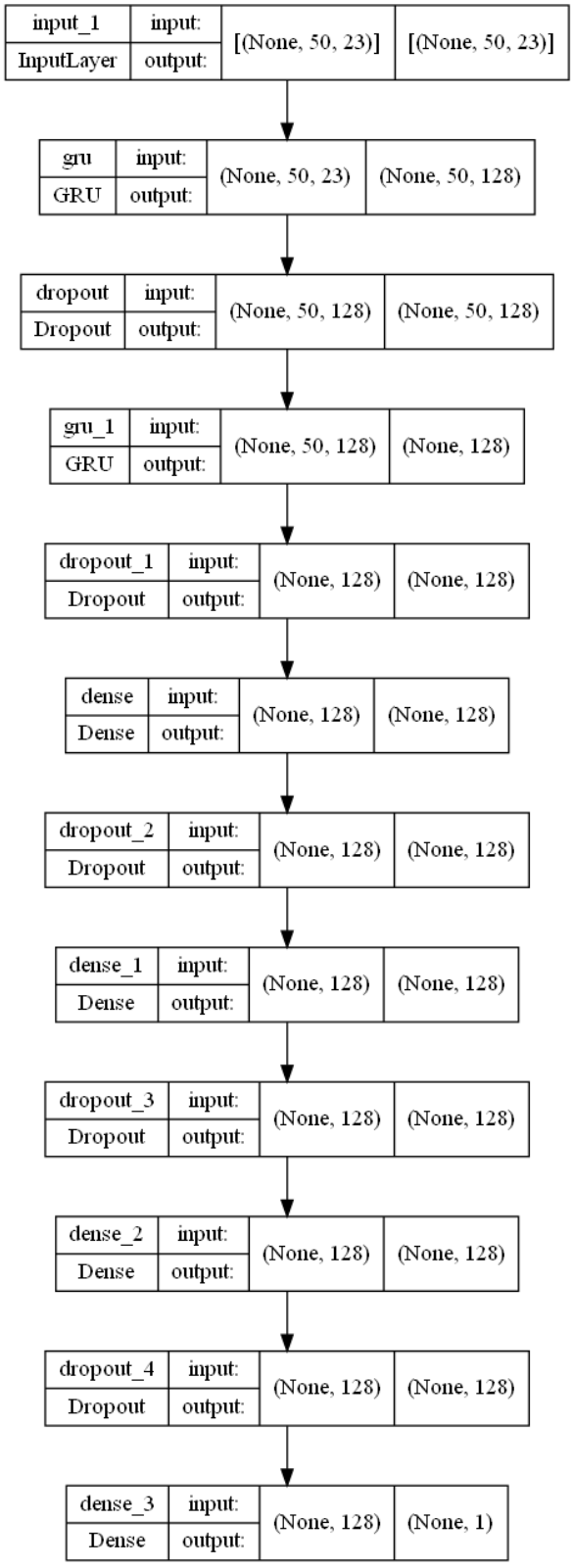

Figure 58. Artificial neural network architecture.

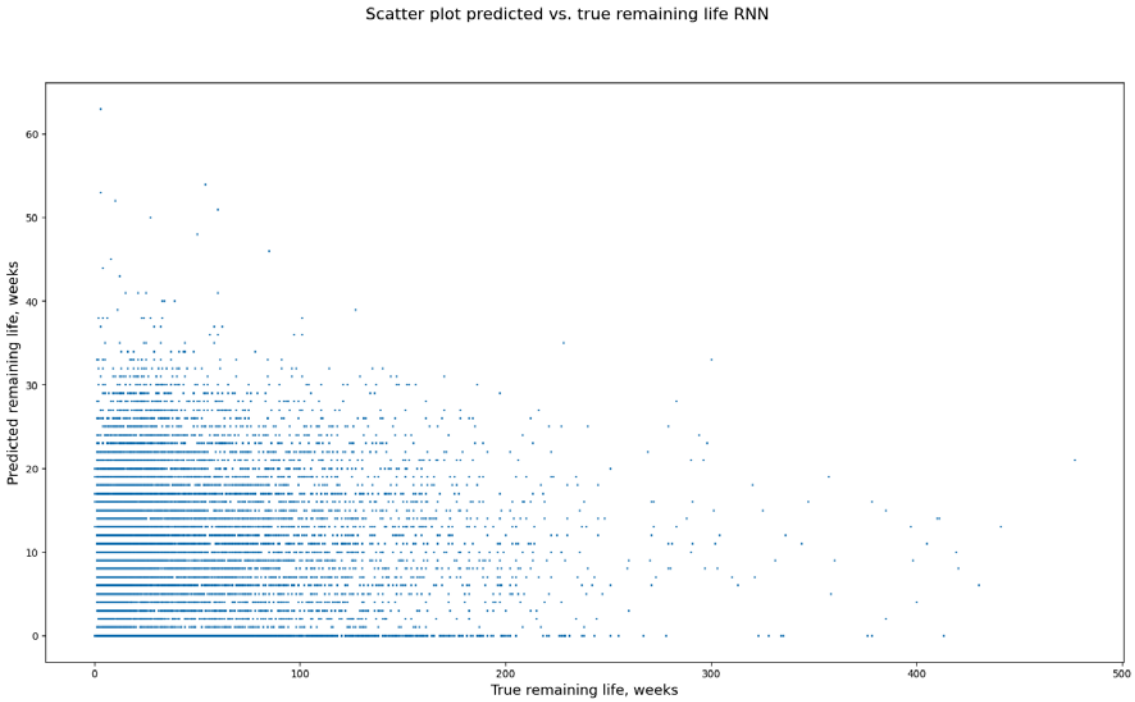

Figure 59. Scatter plot predicted vs. true remaining life estimated by RNN model on censored observations over 20 simulated synthetic data.

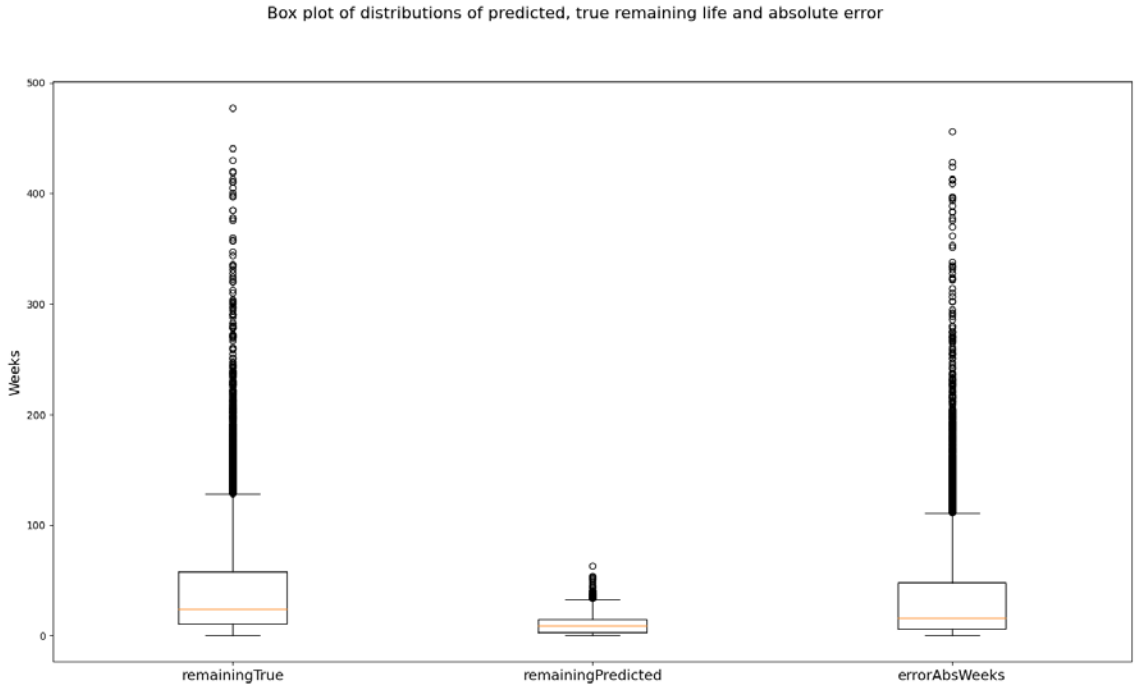

Figure 60. Box plots of distributions of true remaining life (left), predicted remaining life (middle) and MAE (right). Predictions obtained by RNN on subset of censored observations over 20 simulated synthetic datasets.

Results obtained by RNN regression model by predicting censored observations from 20 simulations are presented on **Figure 59** and **Figure 60**. It can be clearly seen that our RNN model significantly underestimates the outcome for censored customers, which is not surprising since the model is trained only on the subset of ‘dead’ clients. Even is the model has large capacity and the ability to learn complex dependencies from temporal sequential data, it does not take into account the concept of censorship and therefore is biased towards shorter values of the outcome.

Conclusion

Studying churn is important for businesses, and non contractual settings are especially challenging. We described and analyzed a reasonable solution based on experience and intuition. While many different strategies exist, we focused on using survival tools to predict customers' lifetime. The experiments described herein show that survival analysis can be applied to predict customers' lifetime for non-contractual setting. Discrete-time model and features created within its scope can serve as predictors for survival regression for lifetime estimation. Univariate models can be used as quick way to estimate populations' mean and median survival time, which in its turn might be useful to predict future purchases and therefore, companies' profit. Best model could be selected via either AIC/BIC criteria or by comparison with Kaplan-Meier estimator (IAE / ISE), which can be treated as 'true' survival function. In our experiment, generalized gamma model was the best among the other univariate models in terms of available metrics we used, except the last one in simulated data (the error between 'true' and predicted profit made by censored customers). However, univariate models do not take into consideration any customers specific characteristics that could be learned from the past buying behavior.

CoxPH and gradient boosting models have significantly better performance than best univariate model (generalized gamma). However, if study interval is relatively short (censored individuals have larger lifetime than study period), those models cannot correctly predict the remaining life, since individuals' survival functions do not go below the value 0.5. Therefore, from models' perspective, those customer have infinitely long lifetime, which is not the case in real life. GB model has best performance characteristics most of the time (highest CI and AUC and lowest IBS), but on simulated data its predictions suffer of underestimation of clients' lifetime. Therefore, this model might be used as a predictor for recent customers churn (if customers remaining life predicted by GB model is very low, this might be an indicator that he is going to churn and requires an attention of retention team). If study period is sufficient, CoxPH model is a good choice for lifetime estimation. However, if the

number of predictions that model cannot handle (predicts infinity lifetime) is significant, AFT model is an excellent choice to do the job. The huge advantage of AFT model is its functional form: once parameters of the model is determined via learning on training data, the survival function for any subject we would like to make prediction can be extended as far as necessary (at least until it will cross the altitude of 0.5 or even conditional zero). As a result, either expected or median survival time can be calculated for any censored customer. It should be mentioned that both top performed AFT models log-normal and log-logistic showed results not very far from best performing models either on real data (3^{rd} and 2^{nd} rank on CI index) or on simulated data (MAE and AbsError).

We find that although many features are just derived from the transaction dates, their use appear to improve the performance of the regression methods. Some methods ignore the censoring, and surprisingly, they yielded reasonable results in our context. Globally, using methods with failed assumptions may still provide good predictions, or not. Making numerous approximations does not have an automatic negative cost on performance but may have a hidden cost of a reduce trustability of the model. Although we are aware of real-life instances where very reasonable assumptions in a rigorous model did not yield the expected performance – outside of Monte Carlo studies, we have no real control on the data generating process.

Bibliography

Anderson, E., & Weitz, B. (1989). Determination of continuity in convential industrial channel dyads. *Marketing Science 8(4)*, 310–323.

Ansell, J., Harrison, T., & Archibald, T. (2007). Identifying cross-selling opportunities, using lifestyle segmentation and survival analysis. *Marketing Intelligence & Planning, 25*(4), 394-410. doi:https://doi.org/10.1108/02634500710754619

Ascarza, E., Netzer, O., & Hardie, B. G. (2018b). Some customers would rather leave without saying goodbye. *Marketing Science, 37*(1), 54-77.

Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications, 39*, 13517-13522.

Bayrak, A. T., Guven, Y., Bahadır, M. B., & Yalcınkaya, S. M. (2022). 2022 IEEE International Conference on Big Data and Smart Computing (BigComp). *Comparative Methods for Personalized Customer Churn Prediction with Sequential Data* (pp. 222-225). IEEE.

Bayrak, A. T., Guven, Y., Bahadır, M. B., & Yalcınkaya, S. M. (2022). Comparative Methods for Personalized Customer Churn Prediction with Sequential Data. *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 222-225). IEEE.

Bhattacharya, C. B. (1998). When customers are members: Customer retention in paid membership contexts. *Journal of the Academy of Marketing Science*(26), 31-45.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*(164).

Chen, Y., Zhang, L., Zhao, Y., & Xu, B. (2022). Implementation of penalized survival models in churn prediction of vehicle insurance. *Journal of Business Research, 153*, 162-171.

Clemente-Císca, M., San Matías, S., & Giner-Bosch, V. (2014). A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings. *European Journal of Operational Research, 239*(1), 276-285. doi:https://doi.org/10.1016/j.ejor.2014.04.029

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry:. *Journal of Business Research*(66), 1629-1636.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187-220.

Djeundje, V. B., & Crook, J. (2019). Dynamic survival models with varying coefficients for credit risks. *European journal of operational research, 275*, 319-333.

Fader, P. S., Hardie, B. G., & Shang, J. (2010). Customer-Base Analysis in a Discrete-Time Noncontractual Setting. *Marketing Science*, 1086-1108.

Glady, N., Baesens, B., & Croux, C. (2009). A modified Pareto/NBD approach for predicting customer lifetime value. *Expert Systems with Applications*, 2062-2071.

Harrell, F., Califf, R., Pryor, D., Lee, K., & Rosati, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association, 247*, 2543–2546.

Hu, S., Chen, P., & Chen, X. (2021). Do personalized economic incentives work in promoting shared mobility? Examining customer churn using a time-varying Cox model. *Transportation Research Part C, 128*, 103224.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician, 50*(4), 361-365.

Jahromi, A. T., Sepehri, M. M., Teimourpour, B., & Choobdar, S. (2010). Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. *Journal of Strategic Marketing, 18*(7), 587–598.

Jahromi, A. T., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management 43*.

Jerath, K., Fader, P. S., & Hardie, B. G. (2011). New Perspectives on Customer "Death" Using a Generalization of the Pareto/NBD Model. *Marketing Science, 30*(5), 866-880. doi:http://dx.doi.org/10.1287/mksc.1110.0654

Karnstedt, M., Hennessy, T., Chan, J., Basuchowdhuri, P., Hayes, C., & Strufe, T. (2010). *Handbook of Social Network Technologies and Applications.* Boston: Springer US.

Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., & Pentland, A. (2018). Behavioral attributes and financial churn prediction. *EPJ Data Science, 7*(1). doi:doi.org/10.1140/epjds/s13688-018-0165-5

Lemon, K. N., White, T. B., & Wine, R. S. (2002). Dynamic customer relationship management: Incorporating future considerations into the service retention decision. *Journal of Marketing*(66), 1-14.

Li, Y., Li, Y., & Li, Y. (2019). What factors are influencing credit card customer's default behavior in China? A study based on survival analysis. *Physica A, 526*, 120861.

Liu, X. (2012). *Survival Analysis Models and Applications.* West Sussex: Higher Education Press.

Morrison, D., & Schmittlein, D. (1988). Generalizing the NBD Model for Customer Purchases: What Are the Implications and Is It Worth the Effort? *Journal of Business & Economic Statistics, 6*(2), 145-159.

Netzer, O., Lattin, J. M., & Srinivasan, V. (2008). A Hidden Markov Model of Customer Relationship Dynamics. *Marketing Science, 27*(2), 185-204. doi:doi 10.1287/mksc.1070.0294

Perisic, A., Jung, D. S., & Pahor, M. (2022). Churn in the mobile gaming field: Establishing churn definitions and measuring classification similarities. *Expert Systems with Applications, 191*, 116277.

Reinartz, W., & Kumar, V. (2002). The mismanagement of customer loyalty. *Harvard Business Review*, 86-94.

Rothenbuehler, P., Runge, J., Garcin, F., & Faltings, B. (2015). SAI Intelligent Systems Conference 2015. *Hidden Markov Models for churn prediction* (pp. 723-730). London: IntelliSys.

Schmittlein, D. C., & Peterson, R. A. (1994). Customer base analysis: An industrial purchase process application. 13 (1), 41–67. *Marketing Science*(13), 41-67.

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Sci.*, 1-24.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine, 30*(10), 1105-1117.

Uno, H., Cai, T., Pencinac, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *National institute of health, 30*(10), 1105-1117.

Wong, K. K.-K. (2011). Using Cox regression to model customer time to churn in the wireless telecommunications industry. *Journal of Targeting, Measurement and Analysis for Marketing, 19*, 37-43. doi:doi: 10.1057/jt.2011.1 ;

Wu, C., & Chen, H.-L. (2000). Counting your customers: Compounding customer's in-store decisions, interpurchase time and repurchasing behavior. *European Journal of Operational Research*(127), 109-119.

Zhang, Y., Bradlow, E., & Small, D. S. (2014). Predicting Customer Value Using Clumpiness: From RFM to RFMC. *Marketing Science, 34*(2), 195-208.