Customer Churn Prediction and Management

Project report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology

In

Electronics and Communication Engineering

Submitted by

Debanjali Saha (1810110059)

Under supervision of

Premchand Jain
Department of Electrical Engineering, Shiv Nadar University

Amit Kumar (Mentor)
Ericsson India Global Services Private Limited, Gurgaon

Anand Joshi (Line Manager)
Ericsson India Global Services Private Limited, Gurgaon

Department of Electrical Engineering
School of Engineering
Shiv Nadar University, Delhi NCR

(Spring 2022)

Candidate Declaration

I hereby declare that the thesis entitled "Customer Churn Prediction and Management" was submitted for the B.Tech. Degree program. This thesis has been written in my own words. I have adequately cited and referenced the original sources.

Debanjali

Debanjali Saha

(1810110059)

Date: 28/04/2022

CERTIFICATE

It is certified that the work contained in the project report titled “Customer Churn Prediction and Management” by “Debanjali Saha” has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

(Signature)
Dr. Premchand Jain
Dept. of Electrical Engineering
School of Engineering
Shiv Nadar University
Date: 29/04/2022

(Signature)
Anand Joshi
Product Development Leader
BSS PDG E-billing
Date: 28/04/2022

Abstract

A **BSS (Business Support System)** is a component used by Telecommunication Service Providers to run their business operations towards customers. The term is used to broadly describe customer-facing functionalities for a business. The tools which are a part of BSS allow organizations to connect with their customers (like CRM or Customer Relationship Management), create offers for customers (like products or services), take orders from customers, tackle payment issues (billing and rating), revenues, etc.

Ericsson develops such consumer telecom BSS stacks for **CSPs (Communication Service Providers)** enabling them to monetize the network as a platform; serve enterprise customers; facilitate ecosystem partners; support dedicated networks, IoT platforms, connectivity, communication, and collaboration.

The aim of the project that I am a part of at the R&D unit at Ericsson is to apply **AI-enabled personalization** to the application of **Retention Policies** to reduce customer dropout. This is the next step in automating the process of identification and solution to **Customer churn**. Customer churn occurs when subscribers stop their subscription with a particular service provider

The goal of my project is to use a combination of **Survival Analysis** and **Machine Learning** techniques to automate this complex and highly subjective task of detecting **"churners"**. Further, the next step after optimally solving the problem is to go ahead and explain the results using **Explainable AI** to our customers (TSPs) and justify our results.

We first implement the **Kaplan Meier method** to get an idea about the behavior of the entire population as a whole as well as of various cohorts within the population with respect to the different classes within every feature. Next, we find the effect of every covariate or feature and the magnitude of the effect it has on churn by implementing the **Cox Proportional hazard model**. Thereafter, we combine the **Accelerated Failure Time (AFT)** model fitted in a **Weibull distribution** along with **Extreme Gradient Boosting (XGBoost)** machine learning model.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Customer churn, also known as customer attrition is the name given to the phenomenon wherein customers of a business no longer purchase or interact with the business. In other words, it occurs when customers or subscribers decide to discontinue using a company's product or service during a certain time frame leading to the end of their association. Customer churn is one of the most important metrics for a growing business to evaluate because though it is not the happiest measure, it is a number that can help a company identify its customer retention.

Every business deals with churn, and it is practically much easier and less costly to keep an existing customer than it is to gain a new customer. For instance, an increase in customer retention of just 5% can create at least a 25% increase in profit. This is because returning customers will likely spend 67% more on a company's products and services. As a result, the company can spend less on the operating costs of having to acquire new customers. It is also much easier and more logical to save a customer before they leave than it is to convince the customer to come back. Customer loyalty is something all businesses strive for and so, understanding and preventing churn is critical to achieving this.

In this project, a statistical method called Survival Analysis has been used to build the churn prediction model. **Survival analysis** is a collection of statistical techniques used for analyzing the expected duration of time until one event occurs, such as a death in biological organisms and failure in mechanical systems. It is also referred to as 'Time-to-Event' analysis as the goal is to estimate the time for an event of interest. This estimated time is the duration between birth and death events.

The term was coined in medicine to estimate time until death and was originally developed and used by Medical Researchers and Data Analysts to measure the lifetimes of a certain population. But, it also can be used in several other useful applications such as predicting churning customers/employees, estimation of the lifetime of a Machine, etc. The birth event in our case is the time a customer starts their membership with a company, and the death event can be considered as the customer leaving the company.

Survival analysis constitutes a culmination of several statistical analysis methods that together evaluate the effect of predictors on time until an event, rather than the probability of an event, occurs. It is also known as reliability analysis in the engineering discipline, duration analysis in the economics discipline, and event history analysis in the discipline of sociology.

More generally, survival analysis involves the modeling of time to event data; in the context of this project, “churn” will be considered an "event" in the survival analysis literature – traditionally only a single event occurs for each subject, after it is dead or broken. In survival analysis we use the term ‘failure’ to define the occurrence of the event of interest. The term ‘survival time’ specifies the length of time taken for failure to occur.

By utilizing the science of Survival Analysis, not only can companies predict if customers are likely to stop doing business but also after how much time the event might actually happen.

1.1 Background

The global telecommunications industry has become a highly mature and competitive industry over the last couple decades. Maintaining a sound customer base is a primary concern for all telecommunications industries, specifically the mobile telecommunications sector. Marketing efforts are often designed to sustain market share and increase profit levels have switched from focusing on acquiring new customers to retaining existing customers or reducing customer churn altogether.

Customer churn in the telecommunication industry is the process by which existing customers end their service with a provider. The telecom industry experiences an average 30-35% annual churn rate and according to research, it costs 5-10 times more to recruit a new customer than to retain an existing one (Lu, J., 2002). Although churn is an issue in any industry that offers a service or subscription, it is most prevalent in the mobile telecommunications industry as it impacts business directly.

Churn prediction is the ability to detect customers who are likely to end their service or subscription prior to contract termination and so being able to determine which customers are most likely to churn allows a company to target their marketing efforts (incentives and discounts) toward a specific segment of the existing customer base in an attempt to improve customer retention and increase revenue (Owczarczuk, 2010).

Churn prediction would also help to allocate current customers into different groups based on their probability of churning. Shapoval and Setzer (2015) discuss the terms of contract in predicting churn. Hence, it is critical for businesses to have a solid performing churn prediction model that accurately predicts each customer's probability of churning because it can be extremely costly and inefficient to market to all existing customers, especially loyal and long tenure customers who have very low probabilities of churning.

Thus, a critical component to an efficient and successful marketing strategy that improves customer retention is accurately predicting the existing customers who are most likely to churn. To do this, companies must understand their customers' churning process and the factors that increase the probability of customer churn.

Moreover, it is not only necessary to predict which customers will churn, it is also necessary to calculate when they will do so. Various statistical techniques can be used to develop churn prediction models- Logistic regression and data mining methods, such as neural networks, have often been used to score customers based on their predicted probability of churning.

However, both techniques have certain shortcomings. Logistic regression accurately predicts probabilities of churning and explains why customers will churn based on usage data and customer personal information, but it does not provide any information on when a customer will churn. Likewise, neural networks also perform well in predicting probabilities of churning, but the method is more of a "black box" which refers to the challenge of interpreting the results. Neural networks, unlike logistic regression, do not possess the ability to explain why a customer will churn, but rather only provide the predicted probability of churn. Both these methods do not offer any information about the timing of when a customer will churn.

An alternative statistical technique is survival analysis which is commonly used when the objective of the model is to predict the timing of an event from its origin to its occurrence. This method lends well to churn prediction modeling, where the objective is to predict when a customer will churn from a time of origin. Survival analysis fills the void that logistic regression and data mining techniques leave by predicting the timing of when a customer will churn, in addition to the probability of churning. Like logistic regression, survival analysis also identifies the key drivers that cause customers to churn and is now becoming the modeling technique of choice since traditional regression techniques identify churners but not predict when they will churn.

1.2 Purpose, Scope, Baseline and Compatibility

Churn is the proportion of subscribers who leave the network. Customer churn causes a direct loss of revenue to a business. Operators must be able to retain their customer base and minimize churn. To enable this, one of the important tasks is to identify the customers who are likely to churn and organize them in groups with priority so that preventive measures can be taken. There are multiple activities involved in churn identification – data preparation, ML model training and organizing the churn scores and exposing downstream applications as insights.

1.3 BUC Requirements

Following are the requirements defined for our Business Use-Case:

1.3.1 Churn prediction training template configuration

As a system user, I want to train a churn prediction model utilizing a training strategy template. This template would enable the following:

1. Churn Model Training
2. Model persistence
3. Training Metric calculation
4. Updating model metric
5. Update model training information
6. Update training run metadata

1.3.2 Churn Prediction Model Training

As a system user,

1. I want to predict the lifetime value of a subscriber at different time points in the future based on observed past churn history and customers who did not churn at the point of observation.
2. I want to be able to see how different input features affect the likelihood of churn
3. I want to be able to select a subset of features out of all the features offered. So that only these features are used as part of model inputs and influence recommendation as part of the strategy execution framework

4. Persist serialized model in a shared storage location. So that the output model can be utilized later to derive the churn probability of all subscribers at a given point in time.

1.3.3 Churn Prediction Model Metrics Calculation

As a system user, I would like to:

- Calculate model metrics that will enable me to compare and select from different models generated by churn prediction strategy and save the same in persistent storage.
- Save training information of the model in persistent storage

1.3.4 Churn Training job containerization

As a developer, I should be able to package the churn training job as a docker image, create Spark application specification and execute the same using the Spark operator.
So that churn training job can be run as a Kubernetes workload.

1.3.5 Workflow tasks

Initializing strategy run status
As a system user, I want to initialize the strategy execution run status and a new version of the model meta entity as the first step in the training workflow.

Model training and metric generation
As a system user, I want to trigger the churn model training job towards and create a new version of the model.

Metric update
As a system user, I want to associate generated metrics with model metadata. So that the training effectiveness of the model can be inspected by the operations team.

Update strategy run status
As a system user, I want to update the execution status of a churn training strategy and update model status to indicate the availability of the model for subsequent scoring workflow.

1.4 Solutions Overview

The following diagram describes the Solutions pipeline for customer churn prediction and management to retain customer:

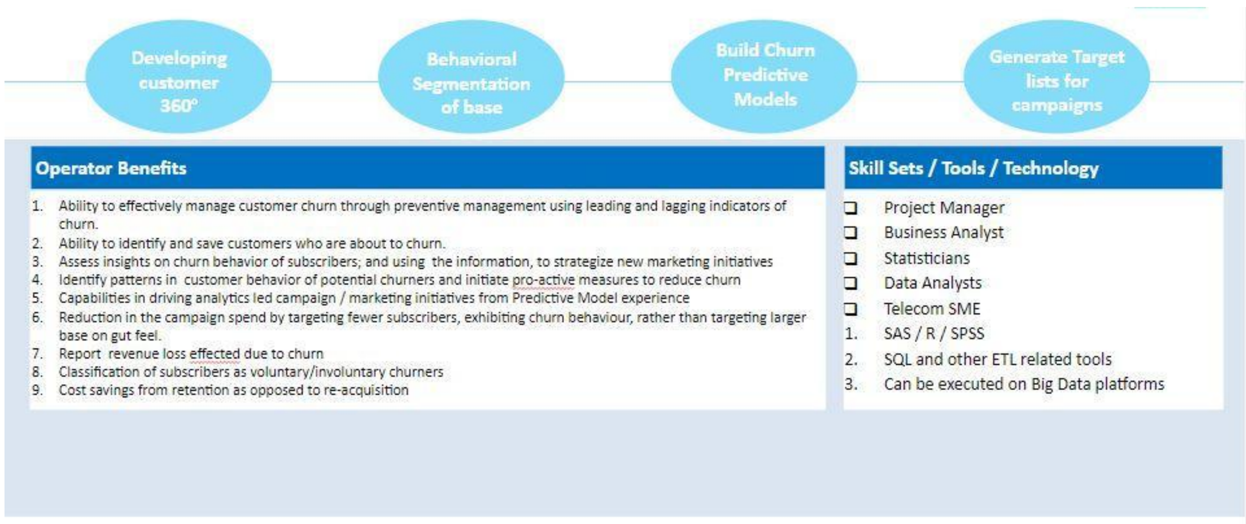

Fig 1.1 Solutions Overview Pipeline

Customer churn causes a direct loss of revenue to a business. Considering the highly competitive landscape and the ever-increasing cost of acquisition, businesses emphasize on retention rather than new customer acquisition. To enable this, it is important to know who the customers are at risk of churning out of the business, estimate and analyze the amount of revenue the customer is likely to generate in his lifetime (CLTV) to be able to build a strategy to retain them. The entire process of managing churn involved numerous activities – data preparation, model training, model scoring, and visualization.

Chapter 2
Literature Survey

2.1 Customer Segmentation Based on Survival Character

"Customer Segmentation" is the process of dividing a company's customers into groups that reflect similarities among customers in each group. It is usually done by single data mining technology from a special point, rather than from a systematic framework. Furthermore, one of the key purposes of customer segmentation is customer retention.

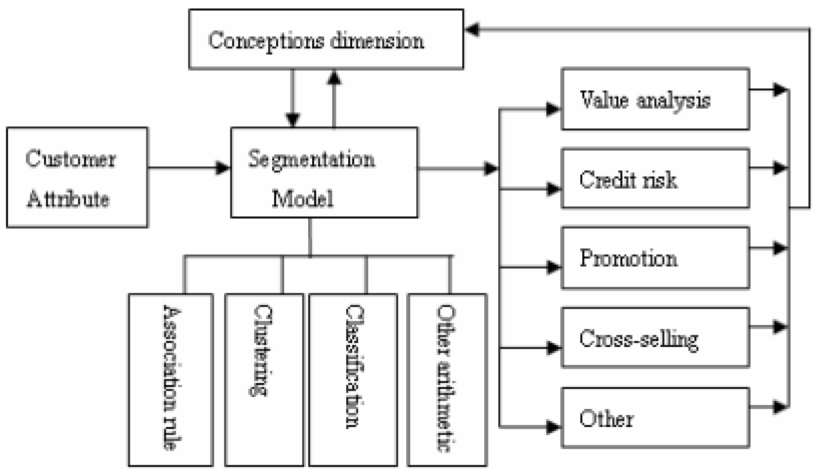

Fig 2.1 Customer Segmentation Framework Based on Data Mining

Although previous segment methods may identify which group needs more care, it is unable to identify customer churn trends for adopting different actions. The paper focused on proposing a customer segmentation framework based on data mining and constructs a new customer segmentation method based on customer survival character

A new customer segmentation method is proposed and constitutes two steps. Firstly, using K-means clustering arithmetic clusters customers into different segments which have a similar survival function (churn trend) inside. Secondly, the method uses survival analysis to predict each cluster's survival/hazard function to test the validity of clustering and identify customer churn trends.

This method was applied in a dataset from China Telecommunications Company, whose result proposes some useful management measures and suggestions.

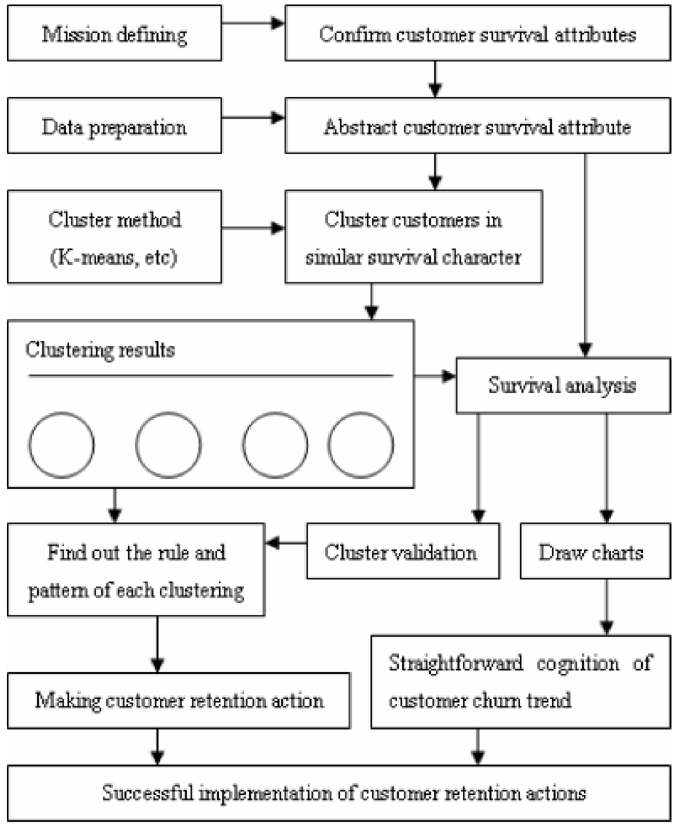

Fig 2.2 Customer Segmentation based on survival character

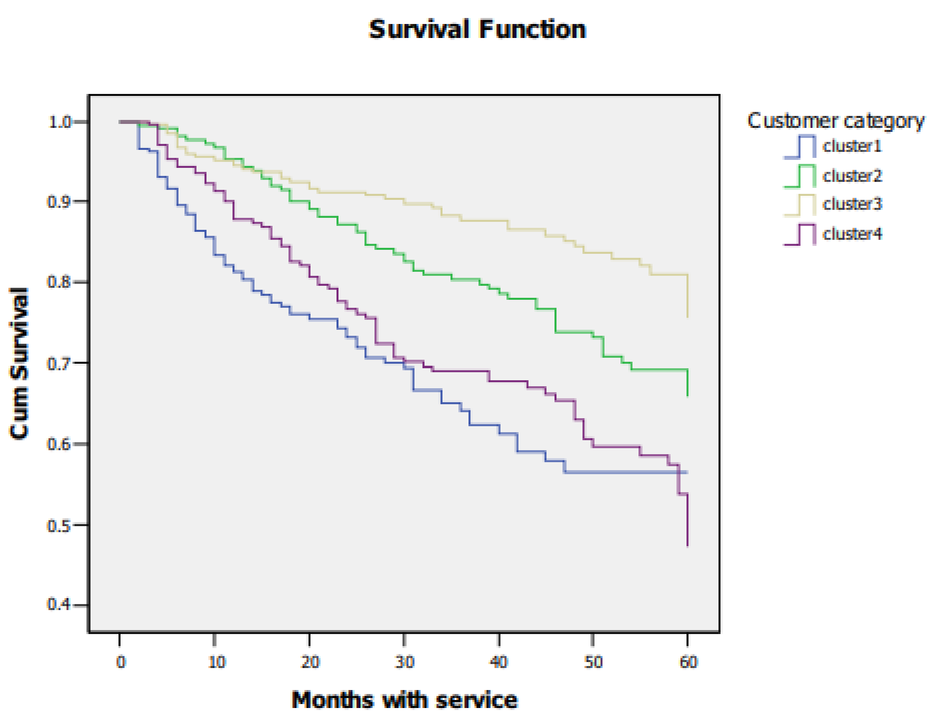

Fig 2.3 Survival Function

2.2 Modeling Customer Lifetime Value Using Survival Analysis — An Application in the Telecommunications Industry

Increasingly, companies are viewing customers in terms of their lifetime value – the net present value of customers' calculated profit over a certain number of months. Customer lifetime value is a powerful and straightforward measure that synthesizes customer profitability and churn (attrition) risk at the individual customers' level.

For existing customers, customer lifetime value can help companies develop customer loyalty and treatment strategies to maximize customer value. For newly acquired customers, customer lifetime value can help companies develop strategies to grow the right customers. The calculation of customer lifetime value varies across industries.

In the telecommunications industry, customer monthly margin and customer survival curve are the two major components to calculate the customer lifetime value. Since customer monthly margin is from accounting models, the key to estimating customer lifetime value is the customer survival curve. In this study, survival analysis is applied to estimate the customer survival curve, therefore customer lifetime value is calculated.

Survival/Churn – In the telecommunications industry, the broad definition of churn is the action that a SUGI 28 Data Mining Techniques customer's telecommunications service is canceled. In this study, both service-provider initiated churn and customer initiated churn are included. An example of service-provider initiated churn is a customer's account being closed because of payment default. Customer initiated churn is more complicated and reasons behind vary from customer to customer. Customer survival is the opposite of customer churn, and both terms are used in the study.

Active – Active is a customer status. Customers whose service is being involuntarily terminated and are in collection stage are not in "active" status.

Granularity – This study examines customer survival/churn at the account level. Customer

Contract – This study does not distinguish customers with or without contracts, although separate models may be desirable for each contract status.

Exclusions – This study does not include employee accounts.

CUSTOMER LIFETIME VALUE

The calculation of customer lifetime value (LTV) varies across industries. In the telecommunications industry, customer monthly margin and customer survival curve are the two major components of customer lifetime value. The customer lifetime value is the net present value of customers' calculated profit over a certain number of months. Here is the formula to calculate customer lifetime value:

$$LTV = MM \times \sum_{i=1}^{T} \left(p_i / (1 + r/12)^{i-1} \right)$$

where MM is the monthly margin for the last three months for existing customers, or the last month's monthly margin for newly acquired customers. MM is either calculated from accounting models or estimated through a set of regression models. The calculation of monthly margin is not the focus of this study and therefore not covered. T is the number of months in consideration to calculate customer lifetime value. r is the discount rate. p_i is the series of customer survival probabilities (customer survival curve) from month 1 through Month T , where $p_1 = 1$. p_i is estimated through a customer survival model.

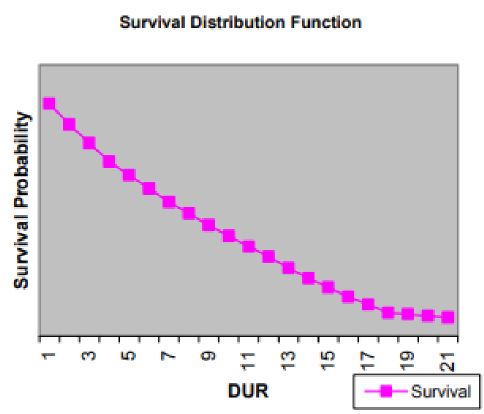

Fig 2.4 Customer Survival Function

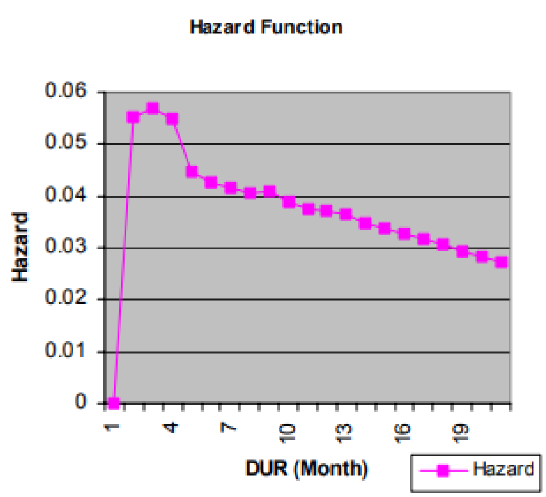

Fig 2.5 Customer Hazard Function

Chapter 3
Theory

3.1 Customer Churn

Fig 3.1 What is customer churn?

Customer churn rate or customer attrition rate is the mathematical figure denoting the percentage of customers who are not likely to make another purchase from a business. The churn rate can be calculated by dividing the number of customers lost during that time period by the number of customers present at the beginning of that time period. For a company to expand its clientele, its growth rate, which is measured by the number of new customers, must exceed its churn rate.

The churn rate not only includes when customers terminate service without but also includes when customers switch services within the same business to a lower profitable service. This measurement is most valuable in subscriber-based businesses in which subscription fees comprise most of the revenues such as in the telecommunications industry. the churn rate is an important factor in the telecommunications industry. In most areas, many of these telecom companies compete, making it easy for people to transfer from one provider to another. Hence, the churn rate is an important factor in the telecommunications industry.

For the first half of the term, my work was dedicated to the prediction of Churn among customers using traditional machine learning models and analyzing which model would give the best performance on the real customer dataset.

There are broadly 4 different types of churn scenarios for telecommunication service providers.

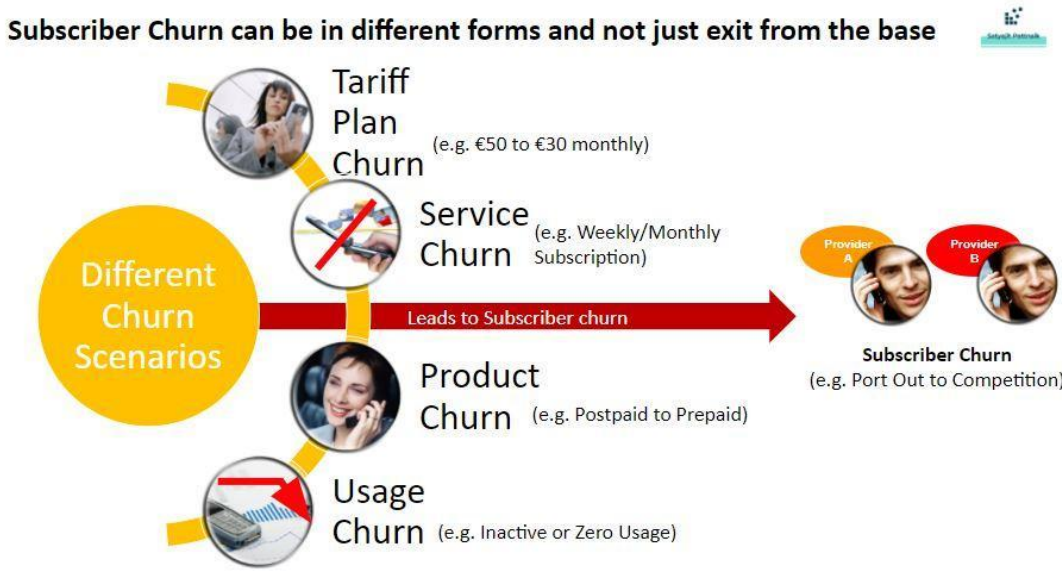

Fig 3.2 Different churn scenarios

In order to understand customer behavior, it is extremely important to estimate the decision cycle of a subscriber. These decisions are often related to the churn segment that a subscriber belongs to.

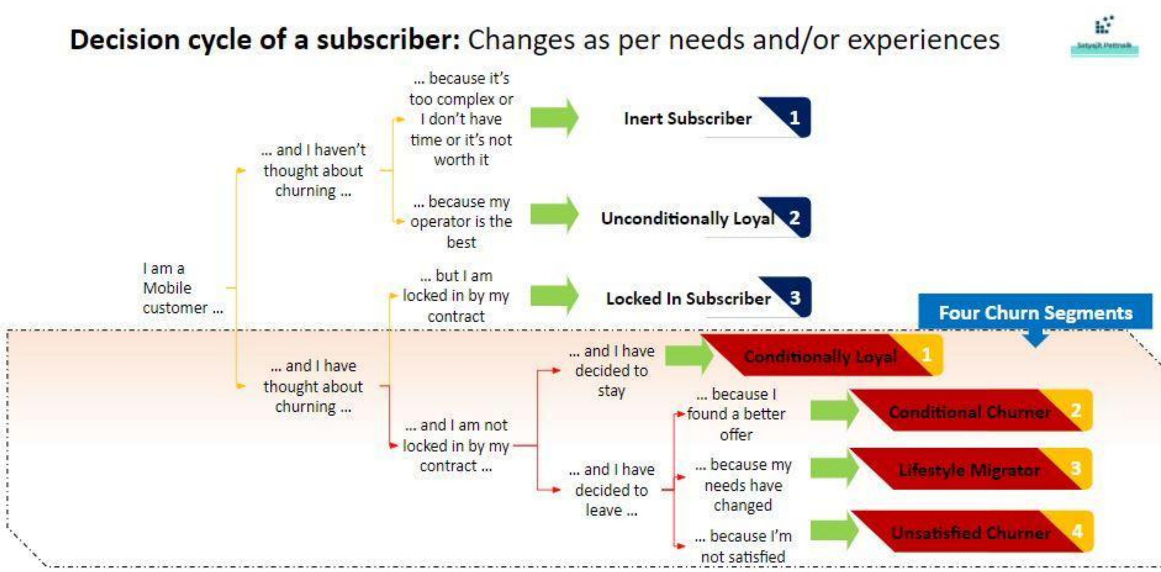

Fig 3.3 Decision Cycle of a Subscriber

In the population of subscribers, the churn segments are differentiators that define the reason for churn for a particular segment of subscribers who decide to discontinue the service. There are broadly 4 churn segments.

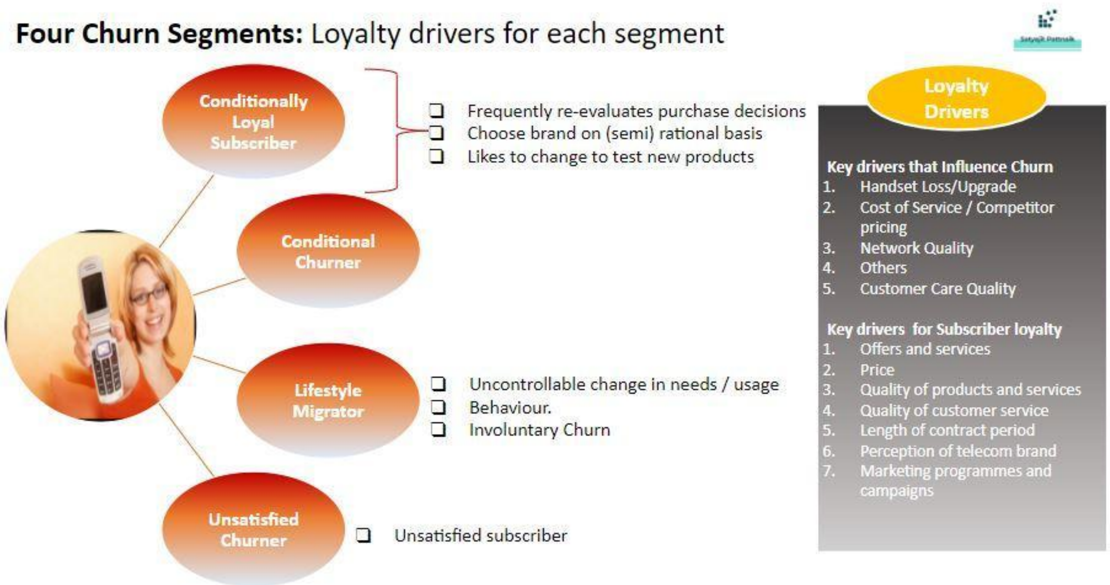

Fig 3.4 Churn Segments

3.2 Survival Analysis

Survival analysis can be used in several ways:

- To describe the survival times of the members of a group:
    - ➢ Life tables
    - ➢ Kaplan–Meier curves
    - ➢ Survival function
    - ➢ Hazard function

- To compare the survival times of two or more groups:
    - ➢ Log-rank test

- To describe the effect of categorical or quantitative variables on survival;
    - ➢ Cox proportional hazards regression
    - ➢ Parametric survival models
    - ➢ Survival trees
    - ➢ Survival random forests

Definitions of common terms in survival analysis:

The following terms are commonly used in survival analyses:

- ❖ Event: Death, disease occurrence, disease recurrence, recovery, or other experience of interest
- ❖ Time: The time from the beginning of an observation period (such as surgery or beginning treatment) to (i) an event, or (ii) end of the study, or (iii) loss of contact or withdrawal from the study.
- ❖ Censoring / Censored observation: Censoring occurs when we have some information about individual survival time, but we do not know the survival time exactly. The subject is censored in the sense that nothing is observed or known about that subject after the time of censoring. A censored subject may or may not have an event after the end of observation time.
- ❖ Survival function S(t): The probability that a subject survives longer than time t.

3.2.1 Data for Survival Analysis

In survival analysis, we do not need the exact starting points and ending points of every entity. All the observations do not always start at zero. A subject can enter at any time in the study and may not experience the event of interest. Now, for the analysis, all the subjects are brought to a common starting point where the time t is zero ($t = 0$) and all subjects have the survival probabilities equal to one, i.e their chances of not experiencing the event of interest (death, churn, etc) is 100%.

There may arise situations where the volume of the data prevents it from being used completely in Survival Analysis. For such situations, Stratified Sampling helps in which the goal is to have an equal or nearly equal number of subjects from each group of subjects in the whole population. Each group is called a Strata and the whole population is stratified (divided) into groups based on some characteristic.

Now, in order to pick a certain number of subjects from each group, you can use Simple Random Sampling. The total number of subjects is specified at the start and you split the total number required among each group and you pick that number of subjects randomly from each group.

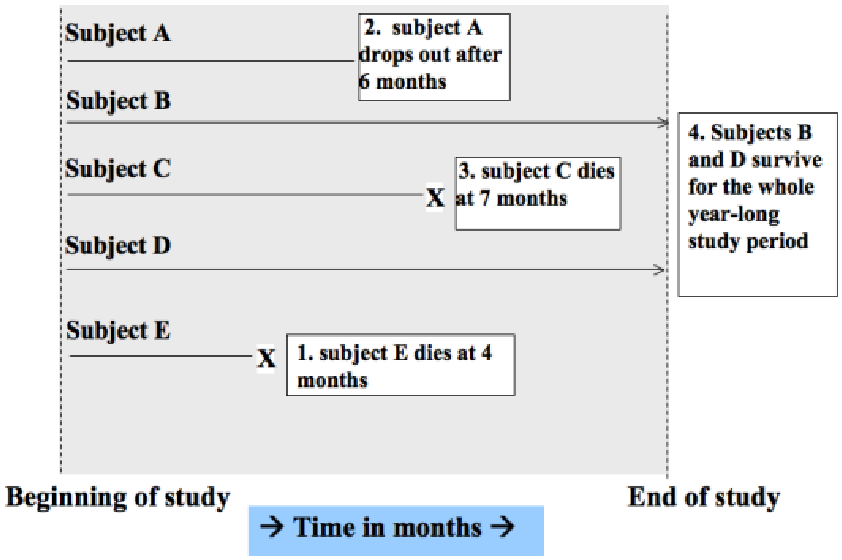

Fig 3.5 Survival Data Example

3.2.2 General Principles

3.2.2.1 Describing time-to-event

Death Density

When the variable under consideration is the length of time taken for an event to occur (e.g. death) a frequency histogram can be constructed to show the count of events as a function of time. A curve fitted to this histogram produces a death density function f(t), as shown in Figure 1. If we set the area under the curve of the death density function to equal 1 then for any given time t the area under the curve to the left of t represents the proportion of individuals in the population who have experienced the event of interest. The proportion of individuals who have died as a function of t is known as the cumulative death distribution function and is called F(t).

Survival

Considering again the death density function shown in Figure 1. The area under the curve to the right of time t is the proportion of individuals in the population who have survived to time t, S(t). S(t) can be plotted as a function of time to produce a survival curve, as shown in Figure 2. At t = 0 there have been no failures so S(t) = 1. By day 15 all members of the population have failed and S(t) = 0. Because we use counts of individuals present at discrete time points, survival curves are usually presented in step format. The survival function S(t), is the probability that a subject survives longer than time t. S(t) is theoretically a smooth curve.

$$\text{Survival Function: } S(t) = P(X > t) = 1 - F(t)$$

$$\text{Cumulative failure function: } F(t) = P(X \leq t)$$

<u>Hazard</u>

The instantaneous rate at which a randomly-selected individual known to be alive at time (t - 1) will die at time t is called the conditional failure rate or instantaneous hazard, h(t). Mathematically, instantaneous hazard equals the number that fail between time t and time t $+ \Delta(t)$ divided by the size of the population at risk at time t, divided by $\Delta(t)$. This gives the proportion of the population present at time t that fail per unit time.
Instantaneous hazard is also known as the force of mortality, the instantaneous death rate, or the failure rate.

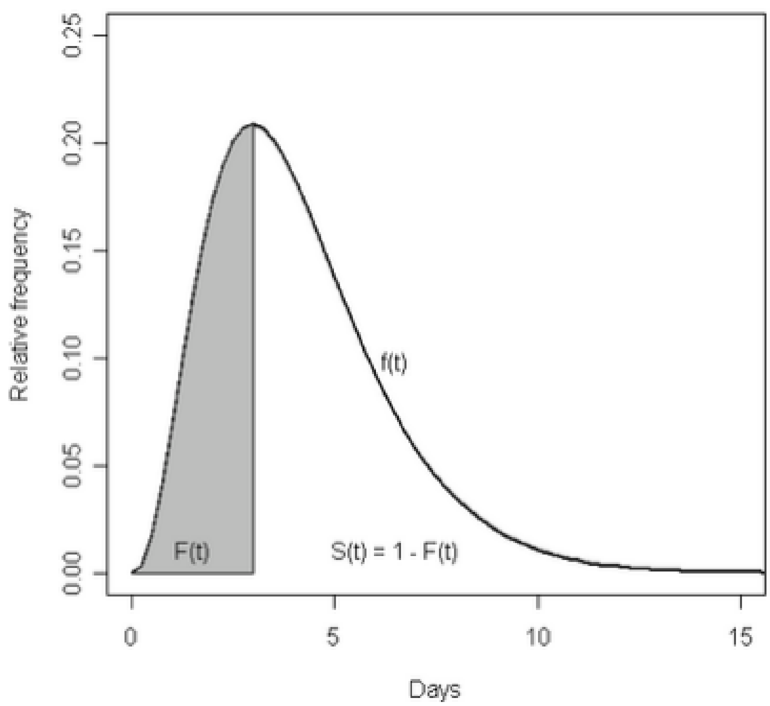

Fig 3.6 Survival Data ExampleLine plot f(t) (death density) as a function of time. The cumulative proportion of the population that has died up to time t equals F(t). The proportion of the population that has survived to time t is $S(t) = F(t) - 1$.

The cumulative hazard (also known as the integrated hazard) at time t, H(t) equals the area under the hazard curve up until time t. A cumulative hazard curve shows the (cumulative) probability that the event of interest has occurred up to any point in time.

3.2.2.2 Censoring and Truncation

It is important to understand that not every member of the population will experience the Event of Interest (death, churn, etc) during the study period. For example, there will be customers who are still a member of the company, or employees still working for the company, or machines that are still functioning during the observation/study period. We do not know when they will experience the event of interest as of the time of the study. All we know is that they haven't experienced it yet. Their survival times are longer than their time in the study. Their survival times are thus, labeled as 'Censored'. This indicates that their survival times were cut-off. Therefore, Censoring allows you to measure lifetimes for the population who haven't experienced the event of interest yet.
It is worth mentioning that the people/subjects who didn't experience the event of interest need to be a part of the study as removing them completely would bias the results towards everyone in the study experiencing the event of interest. So, we cannot ignore those members and the only way to distinguish them from the ones who experienced the event of interest is to have a variable that indicates censoring or death (the event of interest).

Censoring: Sources/events can be detected, but the values (measurements) are not known completely. We only know that the value is less than some number.

Truncation: An object can be detected only if its value is greater than some number; and the value is completely known in the case of detection. For example, objects of certain type in a specific region of the sky will not be detected by the instrument if the apparent luminosity of objects is less than a certain lower limit. This often happens due to instrumental limitations or due to our position in the universe.

The main difference between censoring and truncation is that a censored object is detectable while the object is not even detectable in the case of truncation.

There are different types of Censoring done in Survival Analysis as explained below. Censoring must be independent of the future value of the hazard for that particular subject.

- Right Censoring: This happens when the subject enters at t=0 i.e at the start of the study and terminates before the event of interest occurs. This can be either not experiencing the event of interest during the study, i.e they lived longer than the duration of the study, or could not be a part of the study completely and left early without experiencing the event of interest, i.e they left and we could not study them any longer. The exact value X is not measurable, but only $T = \min(X, C)$ and $\delta = I(X \leq C)$ are observed.

- Left Censoring: This happens when the birth event wasn't observed. Another concept known as Length-Biased Sampling should also be mentioned here. This type of sampling occurs when the goal of the study is to perform analysis on the people/subjects who already experienced the event and we wish to see whether they will experience it again. The lifelines package has support for left-censored datasets by adding the keyword left_censoring=True. Note that by default, it is set to False. Only $T = \max(X, C)$ and $\delta = I(X \geq C)$ are observed.

- Interval/Double Censoring: This happens when the follow-up period, i.e time between observation, is not continuous. This can be weekly, monthly, quarterly, etc. This occurs when we do not observe the exact time of failure, but rather two time points between which the event occurred:
  $(T, \delta) = (X, 1) : L < X < R$ or $(R, 0) : X > R$ or $(L, -1) : X < L$
  where L and R are left and right censoring variables.

★ Left Truncation: It is referred to as late entry. The subjects may have experienced the event of interest before entering the study. There is an argument named 'entry' that specifies the duration between birth and entering the study. If we fill in the truncated region then it will make us overconfident about what occurs in the early period after diagnosis. That's why we truncate them.

★ Right Truncation: a subject is right truncated if it leaves the population at risk some stage after the study starts (and we know that there is no way the event of interest could have occurred after this date). It happens when we cannot identify the population of interest since their initiating events are latent (like the times of infections), and the population becomes "visible" only when another event, that can be more easily identified, happens.

3.2.3 Non-parametric survival

Once we have collected time to event data, our first task is to describe it — usually this is done graphically using a survival curve. Visualization allows us to appreciate temporal patterns in the data. It also helps us to identify an appropriate distributional form for the data. If the data are consistent with a parametric distribution, then parameters can be derived to efficiently describe the survival pattern and statistical inference can be based on the chosen distribution. Non-parametric methods are used when no theoretical

distribution adequately fits the data. In epidemiology non-parametric (or semi-parametric) methods are used more frequently than parametric methods. There are three non-parametric methods for describing time to event data:
(1) the Kaplan-Meier method
(2) the life table method
(3) the Nelson-Aalen method.

3.2.3.1 Kaplan-Meier method

The Kaplan-Meier method is based on individual survival times and assumes that censoring is independent of survival time (that is, the reason an observation is censored is unrelated to the cause of failure). The Kaplan-Meier estimator of survival at time t is shown in Equation 1. Here t j , j = 1, 2, ..., n is the total set of failure times recorded (with t+ the maximum failure time), dj is the number of failures at time tj, and rj is the number of individuals at risk at time tj. A worked example is provided in Table 1.
(1) for each time period the number of individuals present at the start of the period is adjusted according to the number of individuals censored and the number of individuals who experienced the event of interest in the previous time period
(2) for ties between failures and censored observations, the failures are assumed to occur first.

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{(r_j - d_j)}{r_j}, \text{ for } 0 \leq t \leq t^+$$

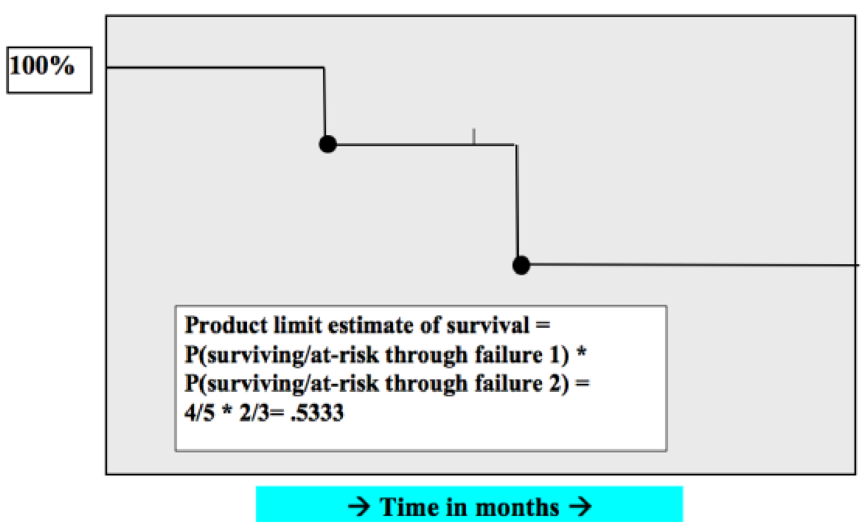

Fig 3.7 Corresponding Kaplan-meier Curve for data in Fig 2.5

The KM curve can be interpreted as follows:

- The x axis is time, from zero (when observation began) to the last observed time point.
- The y axis is the proportion of subjects surviving. At time zero, 100% of the subjects are alive without an event.
- The solid line (similar to a staircase) shows the progression of event occurrences.
- A vertical drop indicates an event.
- These events are indicated by the vertical drops in the KM plot at those time points.

3.2.3.2 Life Table method

The life table method (also known as the actuarial or Cutler Ederer method) is an approximation of the Kaplan-Meier method. It is based on grouped survival times and is suitable for large data sets. Calculation details are shown in Table 2. The life table method assumes that subjects are withdrawn randomly throughout each interval — therefore, on average they are withdrawn half way through the interval. This is not an important issue when the time intervals are short, but bias may be introduced when time intervals are long. This method also assumes that the rate of failure within an interval is the same for all subjects and is independent of the probability of survival at other time periods. Life tables are produced from large scale population surveys (e.g. death registers) and are less-frequently used these days (the Kaplan-Meier method being preferred because it is less prone to bias).

3.2.3.3 Flemington-Harrington estimator

Instantaneous hazard is defined as the proportion of the population present at time t that fail per unit time. The cumulative hazard at time t, H(t) is the summed hazard for all time up to time t. The relationship between cumulative hazard and survival is as follows:

$$H(t) = -\ln[S(t)], \text{ or } S(t) = e-H(t)$$

The Nelson-Aalen estimator of cumulative hazard at time t is defined as:

$$\hat{H}(t) = Xj:tj \leq tdj\ rj, \text{ for } 0 \leq t \leq t+$$

The Flemington-Harrington estimate of survival can be calculated using the Nelson-Aalen estimate of cumulative hazard using the relationship between survival and cumulative hazard described in the second equation.

3.2.4 Parametric survival

On some occasions the pattern of survivorship for our study subjects follows a predictable pattern. In this situation, parametric distributions can be used to describe time to event. An advantage of using a parametric distribution is that we can reliably predict time to event well after the period during which events occurred for our observed data. Several parametric distributions are used to describe time to event data. Each parametric distribution is defined by a different hazard function, as shown in Table.

Table 3.1 Parametric Survival Distributions and their equations

Distribution	$f(t)^a$	$h(t)^b$	$H(t)^c$	$S(t)^d$
Exponential	λ exp[-λt]	λ	λt	exp[-λt]
Weibull	λpt^{p-1} exp[-λt^p]	λpt^{p-1}	λt^p	exp[-(λt)^p]
Gompertz	a exp[bt] exp[-a/b (exp[bt] - 1)]	a exp[bt]	a/b (exp[bt] - 1)	exp[-a/b (exp[bt] - 1)]
Log-logistic	abt^{b-1} / (1 + at^b)^2	(abt^{b-1}) / (1 + at^b)	log(1 + at^b)	(1 + at^b)$^{-1}$

[a] $f(t)$ probability density.
[b] $h(t)$ instantaneous hazard.
[c] $H(t)$ cumulative hazard.
[d] $S(t)$ survival.

As a general approach to the analysis of time to event data we should plot the hazard function for the observed data and determine whether or not it is consistent with a parametric distribution.

If the data follows a parametric distribution, parametric methods are preferred to non-parametric methods for describing and quantifying factors that influence time to event. In veterinary epidemiology, the most important parametric forms are the exponential and Weibull distributions.

3.2.4.1 The exponential distribution

The exponential distribution is described by the mean, λ. A feature of the exponential distribution is that the instantaneous hazard does not vary over time. Observed survival distributions can be checked for consistency with the exponential distribution by plotting instantaneous hazard as a function of time: exponential distributions in this case will yield a straight line. Alternatively, the log of cumulative hazard can be plotted as a function of the log of time: exponential distributions will yield a 45◦line.

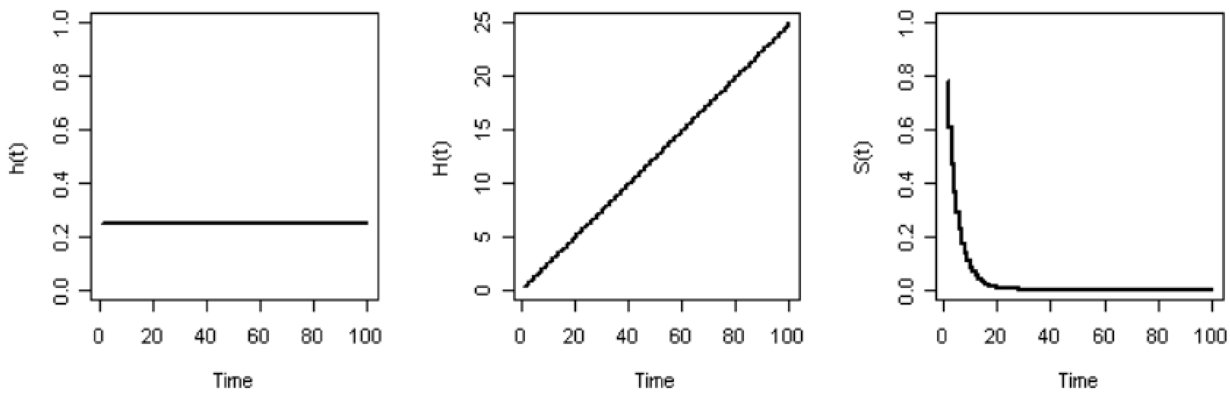

Fig 3.8 Instantaneous hazard, cumulative hazard and survival as a function of time for the exponential distribution. In this example λ = 0.25

3.2.4.2 The Weibull distribution

The Weibull distribution is described by a scale parameter λ and shape parameter p. If p is less than 1 instantaneous hazard monotonically decreases with time, if p equals 1 instantaneous hazard is constant over time (equivalent to the exponential distribution) and if p is greater than 1 instantaneous hazard increases with time. Figure 7 is an example of a Weibull distributed survival pattern with $p < 1$. Time to event data can be checked for consistency with the Weibull distribution by plotting the log cumulative hazard as a function of log time: Weibull distributions in this case will yield a straight line.

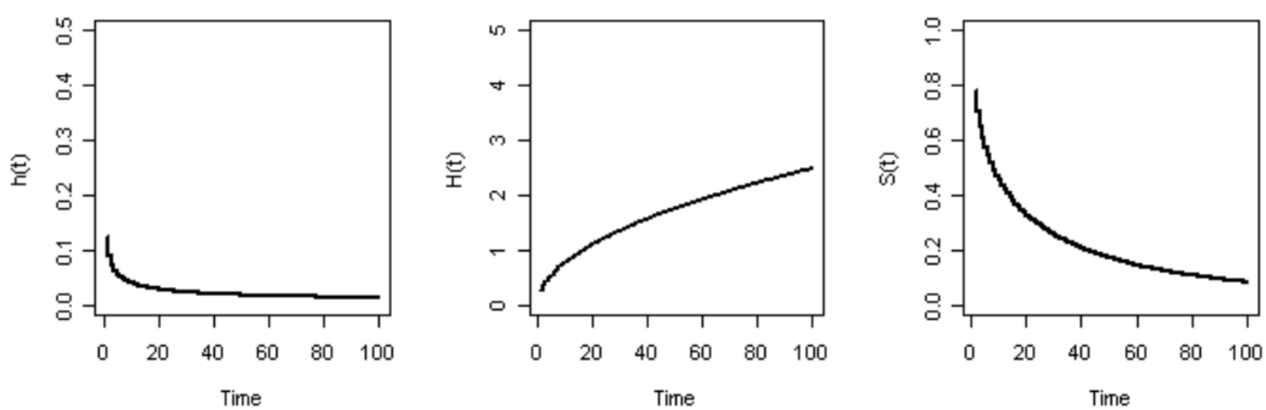

Fig 3.9 Instantaneous hazard, cumulative hazard and survival as a function of time for the Weibull distribution. In this example λ = 0.25 and p = 0.5.

3.2.5 Comparing survival distributions

In addition to providing useful information about how time to event distributions differ among groups, separate survival curves for different levels of covariates provide an effective screening process that helps one to identify factors that are influential in determining survival. Once influential factors are screened using these methods their influence can then be tested using multivariate analyses. When there are no censored observations, standard non-parametric tests can be used to compare two survival distributions. If the groups are independent, a Wilcoxon or Mann-Whitney U test may be used. If the groups are dependent the Sign Test may be used.

3.2.5.1 The log-rank test

The log-rank test (also known as the Mantel log-rank test, the Cox Mantel log-rank test, and the MantelHaenszel test) is the most commonly used test for comparing survival distributions. It is applicable to data where there is progressive censoring and gives equal weight to early and late failures. It assumes that hazard functions for the two groups are parallel. The test takes each time point when a failure event occurs and a 2×2 table showing the number of deaths and the total number of subjects under follow up is created. For each table the observed deaths in each group, the expected deaths and the Variance of the expected number is calculated. These quantities are summed over all tables to yield a χ 2 statistic with 1 degree of freedom (known as the Mantel-Haenszel or log-rank test statistic). The log-rank test calculations also produce for each group the observed to expected ratio which relates the number of deaths observed during the follow up with the expected number under the null hypothesis that the survival curve for that group would be the same as that for the combined data.

3.2.5.2 Other tests

Breslow's test (also known as Gehan's generalized Wilcoxon test) is applicable to data where there is progressive censoring. It is more powerful than the log-rank test when the hazard functions are not parallel and where there is little censoring. It has low power when censoring is high. It gives more weight to early failures. The Cox Mantel test is similar to the log-rank test. It is applicable to data where there is progressive censoring. More powerful than Gehan's generalized Wilcoxon test. The Peto and Peto modification of the Gehan-Wilcoxon test is similar to Breslow's test and is used where the hazard ratio between groups is not constant. Cox's F test is more powerful than Breslow's test if sample sizes are small.

3.2.6 Non-parametric and semi-parametric models

Survival models are used to quantify the effect of one or more explanatory variables on failure time. This involves specification of a linear-like model for the log hazard. A parametric model based on the exponential distribution may be parameterised as follows:

$$\log hi(t) = \alpha + \beta 1xi1 + \beta 2xi2 + ... + \beta kxik \ (4)$$

or, equivalently:

$$hi(t) = \exp(\alpha + \beta 1xi1 + \beta 2xi2 + ... + \beta kxik) \ (5)$$

In this case the constant α represents the log-baseline hazard since log hi(t) = α when all the x's are zero. The Cox proportional hazards model is a semi-parametric model where the baseline hazard α(t) is allowed to vary with time:

$$\log hi(t) = \alpha(t) + \beta 1xi1 + \beta 2xi2 + ... + \beta kxik \ (6)$$

$$hi(t) = h0(t) \exp(\beta 1xi1 + \beta 2xi2 + ... + \beta kxik) \ (7)$$

If all of the x's are zero the second part of the above equation equals 1 so hi(t) = h0(t). For this reason the term h0(t) is called the baseline hazard function. With the Cox proportional hazards model the outcome is described in terms of the hazard ratio. We talk about the hazard of the event of interest at one level of an explanatory variable being a number of times more (or less) than the hazard of the specified reference level of the explanatory variable.

Assumptions of the Cox proportional hazards model are as follows:

- The ratio of the hazard function for two individuals with different sets of covariates does not depend on time.
- Time is measured on a continuous scale.
- Censoring occurs randomly.

3.2.6.1 Model building

Selection of covariates

We now discuss how a set of variables are selected for inclusion in a regression model of survival. Begin with a thorough univariate analysis of the association between survival time and all important covariates. For categorical variables this should include Kaplan-Meier estimates of the group-specific survivorship functions. Tabulate point and interval estimates of the median and quartiles of survival time. Use one or more of the significance tests to compare survivorship among the groups defined by the variable under investigation. Continuous covariates should be broken into quartiles (or other biologically meaningful groups) and the same methods applied to these groups.

Tied events

A central assumption in survival analysis is that time is continuous. Sometimes (particularly in veterinary epidemiological research) the outcome of interest is not measured on a continuous scale and outcome events may occur simultaneously (e.g. service number when conception occurred). When the number of tied events is large, approximate methods yield regression coefficients that are biased towards zero. There are three common methods for dealing with ties:

1. Breslow approximation. There is a contribution to the partial likelihood from each of the tied failure times. For each failure time, the risk set comprises all subjects failing at or after the failure time. This includes all subjects whose failure times are tied with that of the subject contributing to the numerator.

2. Efron approximation. In the Breslow approximation, if m subjects share the same survival time, they all contribute to the risk set for each of the m failure times as if each one of the m subjects failed, all others were still alive. In the Efron approximation, the contribution to the denominator from the m subjects with tied survival times is weighted down by a factor of $(m - k)/m$ for the kth term.

3. Exact partial likelihood. Assuming that no two subjects ever failed simultaneously (this would be the case if we measured the time of failure down to milliseconds), there is a true (unknown) unique ordering of the tied survival times. The exact partial likelihood can be obtained by taking the sum (or average) of the partial likelihoods for all possible orderings of the tied survival times. Computationally intensive.

Fitting a multivariable model
A multivariable model should contain at the outset all covariates significant in the univariate analyses at the $P = 0.20$ to 0.25 level and any others that are thought to be of clinical importance. You should also include any covariate that has the potential to be an important confounder.

Following the fit of the multivariable model, use the P values from the Wald tests of the individual coefficients to identify covariates that might be deleted from the model. The partial likelihood ratio test should confirm that the deleted covariate is not significant. Also check if removal of a covariate produces a 'significant' change (say 20%) in the coefficients of the covariates remaining in the model. Continue until no covariates can be deleted from the model. At this point, work backwards and add each of the deleted covariates back into the model one at a time — checking that none of them are significant or show evidence of being a confounder.

Check the scale of continuous covariates
The next thing is to examine the scale of the continuous covariates in the preliminary model. Here we need to check that the covariate is linear in its log hazard. Replace the continuous covariate with three design variables using Q1, Q2, and Q3 as cutpoints. Plot the estimated coefficients for the design variables versus the midpoint of the group. A fourth point is included at zero using the midpoint of the first group. If the correct scale is linear, then the line connecting the four points should approximate a straight line. Consider transforming the continuous variable if this is not the case. Another method to check this property of continuous covariates using fractional polynomials.

Another method is to use two residual-based plots:
(1) a plot of the covariate values versus the Martingale residuals (and their smooth) from a model that excludes the covariate of interest
(2) a plot of the covariate values versus the log of the ratio of smoothed censor to smoothed cumulative hazard.

To construct the second plot:
(1) fit the preliminary main effects model, including the covariate of interest (e.g. 'age')
(2) save the Martingale residuals (M_i) from this model
(3) calculate $H_i = c_i - M_i$, where c_i is the censoring variable
(4) plot the values of c_i versus the covariate of interest and calculate a lowess smooth (called cLSM)
(5) plot the values of H_i versus the covariate of interest and calculate a lowess

smooth (called HLSM)
(6) the smoothed values from these plots are used to calculate:

$$y_i = \ln \ cLSM\ HLSM\ + \beta_{age} \times age_i$$

and the pairs (y_i , age_i) are plotted and connected by straight lines. There should be a linear relationship between the covariate values and each of the described parameters.

Interactions

The final step is to determine whether interaction terms are required. An interaction term is a new variable that is the product of two other variables in the model. Note that there can be subject matter considerations that dictate that a particular interaction term (or terms) should be included in a given model, regardless of their statistical significance. In most settings there is no biological or clinical theory to justify automatic inclusion of interactions.

The effect of adding an interaction term should be assessed using the partial likelihood ratio test. All significant interactions should be included in the main-effects model. Wald statistic P-values can be used as a guide to selecting interactions that may be eliminated from the model, with significance checked by the partial likelihood ratio test.

At this point we have a 'preliminary model' and the next step is to assess its fit and adherence to key assumptions.

3.2.6.2 Testing the proportional hazards assumption

Once a suitable set of covariates has been identified, it is wise to check each covariate to ensure that the proportional hazards assumption is valid. To assess the proportional hazards assumption we examine the extent to which the estimated hazard curves for each level of strata of a covariate are equidistant over time.

A plot of the scaled Schoenfeld residuals (and a loess smoother) as a function of time and the partial likelihood ratio test may be used to test proportionality of hazards. In a 'well-behaved' model the Schoenfeld residuals are scattered around 0 and a regression line fitted to the residuals has a slope of approximately 0. The fitted curves can be interpreted as estimates of $\beta(t)$ if the coefficients are allowed to vary over time.

For categorical covariates the proportional hazards assumption can be visually tested by plotting -log[-logS(t)] vs time for strata of each covariate. If the proportionality assumption holds the two (or more) curves should be approximately parallel and should not cross. Alternatively, run a model with each covariate(individually). Introduce a time-dependent interaction term for that covariate. If the proportional hazards assumption is valid for the covariate, the introduction of the time-dependent interaction term won't be significant. This approach is regarded as the most sensitive (and objective) method for testing the proportional hazards assumption.

What do you do if a covariate violates the proportional hazards assumption? The first option is to stratify the model by the offending covariate. This means that a separate baseline hazard function is produced for each level of the covariate. Note you can't obtain a hazard ratio for the covariate you've stratified on because its influence on survival is 'absorbed' into the (two or more) baseline hazard functions in the stratified model. If you are interested in quantifying the effect of the covariate on survival then you should introduce a time-dependent interaction term for the covariate, as described above.

3.2.6.3 Residuals

Residuals analysis provides information for evaluating a fitted proportional hazards model. They identify leverage and influence measures and can be used to assess the proportional hazards assumption. By definition, residuals for censored observations are negative and residual plots are useful to get a feeling for the amount of censoring in the data set — large amounts of censoring will result in 'banding' of the residual points.

There are three types of residuals:

1. Martingale residuals. Martingale residuals are the difference between the observed number of events for an individual and the conditionally expected number given the fitted model, follow up time, and the observed course of any time-varying covariates. Martingale residuals may be plotted against covariates to detect non-linearity (that is, an incorrectly specified functional form in the parametric part of the model). Martingale residuals are sometimes referred to as Cox-Snell or modified CoxSnell residuals.

2. Score residuals. Score residuals should be thought of as a three-way array with dimensions of subject, covariate and time. Score residuals are useful for assessing individual influence and for robust variance estimation.

3. Schoenfeld residuals. Schoenfeld residuals are useful for assessing proportional hazards. Schoenfeld residuals provide greater diagnostic power than unscaled residuals. Sometimes referred to as score residuals.

3.2.6.4 Overall goodness-of-fit

To assess the overall goodness-of-fit of a Cox proportional hazards regression model Arjas (1988) suggests plotting the cumulative observed versus the cumulative expected number of events for subjects with observed (not censored) survival times. If the model fit is adequate, then the points should follow a 45◦line beginning at the origin. The methodology is as follows:
(1) create groups based on covariate values (e.g. treated yes, treated no) and sort on survival time within each group
(2) compute the cumulative sum of the zero-one censoring variable and the cumulative sum of the the cumulative hazard function within each group
(3) plot the pairs of cumulative sums within each group only for subjects with an observed survival time.

As in all regression analyses some sort of measure analogous to R2 may be of interest. Schemper and Stare (1996) show that there is not a single simple, easy to calculate, easy-to-interpret measure to assess the goodness-of-fit of a proportional hazards regression model. Often, a perfectly adequate model may have what, at face value, seems like a very low R2 due to a large amount of censoring. Hosmer and Lemeshow recommend the following as a summary statistic for goodness of fit:

$$R_M^2 = 1 - exp\left[\frac{2}{n}(L_0 - L_M)\right]$$

Where:
L0: the log partial likelihood for the intercept-only model,
LM: the log partial likelihood for the fitted model,
n: the number of cases included.

3.2.7 Parametric models

As discussed, semi-parametric models make no assumption about the distribution of failure times, but do make assumptions about how covariates change survival experience. Parametric models, on the other hand, make assumptions about the distribution of failure

times and the relationship between covariates and survival experience. Parametric models fully specify the distribution of the baseline hazard/survival function according to some (defined) probability distribution. Parametric models are useful when we want to predict survival rather than identify factors that influence survival. Parametric models can be expressed in:
(1) proportional hazard form, where a one unit change in an explanatory variable causes proportional changes in hazard
(2) accelerated failure time (AFT) form, where a one unit change in an explanatory variable causes a proportional change in survival time. The advantage of the accelerated the failure time approach is that the effect of covariates on survival can be described in absolute terms (e.g. numbers of years) rather than relative terms (a hazard ratio).

3.2.7.1 Exponential model

The exponential model is the simplest type of parametric model in that it assumes that the baseline hazard is constant over time:

$$h(t) = h0 \exp\beta X, \text{ where } h0 = \lambda \quad (10)$$

The assumption that the baseline hazard is constant over time can be evaluated in several ways. The first method is to generate an estimate of the baseline hazard from a Cox proportional hazards model and plot it to check if it follows a straight, horizontal line. A second approach is to fit a model with a piecewise-constant baseline hazard. Here, the baseline hazard is allowed to vary across time intervals (by including indicator variables for each of the time intervals). The baseline hazard is assumed to be constant within each time period, but can vary between time periods.

3.2.7.2 Weibull model

In a Weibull model it is assumed that the baseline hazard has a shape which gives rise to a Weibull distribution of survival times:

$$h(t) = h0 \exp\beta X, \text{ where } h0 = \lambda ptp-1 \quad (11)$$

Where βX includes an intercept term $\beta 0$. The suitability of the assumption that survival times follow a Weibull distribution can be assessed by generating a log-cumulative hazard plot. If the distribution is Weibull, this function will follow a straight line. The estimated shape parameter from the Weibull model gives an indication of whether hazard is falling ($p < 1$), constant ($p = 1$), or increasing ($p > 1$) over time.

3.2.7.3 Accelerated failure time models

The general form of an accelerated failure time model is:

$$\log(t) = \beta X + \log(\tau) \text{ or } t = \exp \beta X \tau \quad (12)$$

where log(t) is the natural log of the time to failure event, βX is a linear combination of explanatory variables and log (τ) is an error term. Using this approach τ is the distribution of survival times when $\beta X = 0$. If we assume that τ follows a log-normal distribution, then the log of survival times will have a normal distribution, which is equivalent to fitting a linear model to the natural log of survival time (assuming that you can ignore the problem of dealing with censored observations). Equation 12 can be re-expressed as follows:

$$\tau = \exp{-\beta X t} \text{ or } \ln(\tau) = -\beta X + \log(t) \quad (13)$$

The linear combination of predictors in the model (βX) can act additively or multiplicatively on the log of time: they speed up or slow down time to event by a multiplicative factor. In this case $\exp{-\beta X}$ is called the acceleration parameter such that if $\exp{-\beta X} > 1$ time passes more quickly, if $\exp{-\beta X} = 1$ time passes at a normal rate, and if $\exp{-\beta X} < 1$ time passes more slowly.

Exponential and Weibull models can be parameterised as either proportional hazards models or as accelerated failure time models. Other parametric models (e.g. the log-normal, the log-logistic, and gamma) can only be expressed as accelerated failure time models (the predictors in these models do not necessarily multiply the baseline hazard by a constant amount). Accelerated failure time coefficients represent the expected change in ln(t) for a one unit change in the predictor.

Chapter 4
Work Done

4.1 Dataset

The churn dataset preparation job would scan records of customers who already have churned, take a certain number of random snapshots of past behavior of the same population and combine it with a population of active customers of size certain times that of churned population. For a demonstration of the actual work done which is non-disclosable as per the Code-Of-Business-Ethics at Ericsson, this report explains all the R&D done with open-source Kaggle Dataset called Telco Customer Churn.

The first step is to import the data and perform some basic cleaning.

For each customer, we will need two important data points:
(1) 'Tenure': how long they have been a customer when the data is observed
(2) 'Churn': whether or not the customer left when the data was observed

We will first identify these features and ensure the data type is correct. It is extremely important to note here that many customers in our data have not churn yet which means that the dataset is Right-Censored. We drop the columns which are not of any use to our use-case. We change object data types to float/int/binary(boolean)/multi-class(one-hot encoded) as per the data present as each feature.

```
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
```

Fig 4.1 Dataset Information

For the actual dataset to be used for the service, the following are some details:

4.1.1 Configuration Parameters

The below table gives a short elaboration of configuration parameters:

Table 4.1 Configuration Parameters

Parameter Name	Description
X	Threshold of Tenure in days; customers with Tenure below X would be filtered out for both churned and retained customers.
M	Snapshot duration in history in days; the number of days in history from which snapshots of feature data are to be taken for both churned and retained customers.
N	The number of snapshots in integer; N indicates how many snapshots of feature data in history are to be taken for dataset preparation.
F	Sampling factor in decimal; the ratio of retained customers and churned customers within the scope of dataset preparation.
H	Churn history duration in days; the number of days over which records of churn need to be taken for calculating the number of retained customers.

4.1.2 Schema Configuration

The below table describes Schema details:

Table 4.2 Schema Configuration details

Field Name	is_feature	is_censored	is_target
lifetime_days	✗	✗	✓
censored	✗	✓	✗
data_last_30days	✓	✗	✗
data_last_60days	✓	✗	✗

4.1.3 Dataset Configuration

The Dataset will contain a select query to make a union of the following two tables:

- churned_training_censored_ds

- churned_training_uncensored_ds

Note as churned_training_uncensored_ds will be purged from time to time, the union query doesn't need to contain any filtering clause.

4.1.4 Containerization of Churn Dataset Preparation Job

Churn dataset preparation job would be built to create a docker image and would be executed as Pod on K8S using Spark Operator.

4.2 Kaplan-Meier Estimation

Next step is to estimate the survival rate for the average customer using a Kaplan-Meier survival curve. So, we fit a KM survival curve to the customer churn data, and plot our survival curve with a confidence interval.

The survival curve is cumulative. Meaning, in the graph below, after 20 months, the chance of a customer not canceling service is just above 80%.

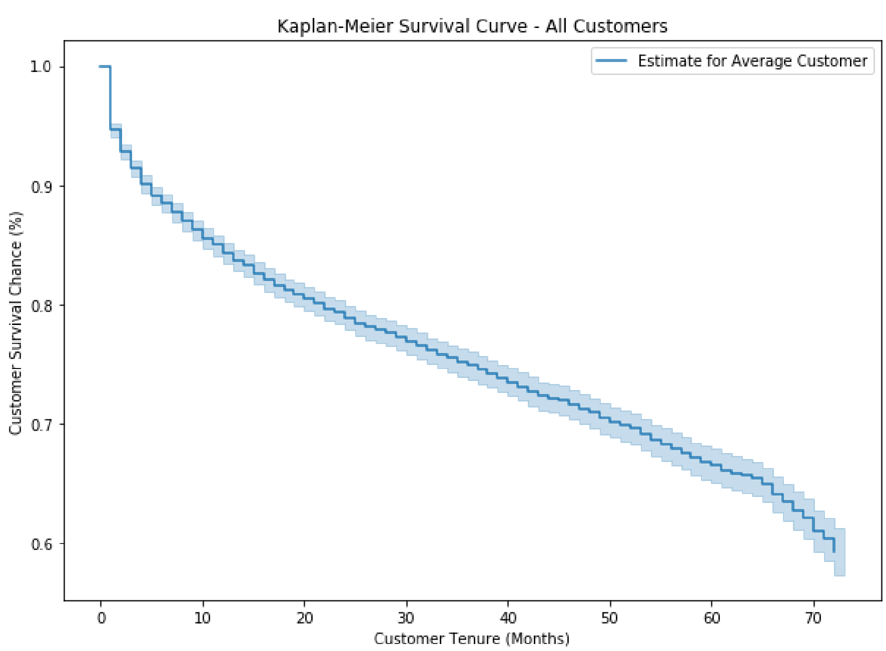

Fig 4.2 Kaplan Meier Plot for entire population

The above should give some basic intuition about the customers. As we would expect for telecom, churn is relatively low. Even after 72 months, the company is able to retain about 60% or more of their customers.

4.2.1 Customer Retention Recommendations based on Kaplan-Meier curve plots segmented by categorical variables

How can our telecom company reduce customer churn?

We can make recommendations along three dimensions: contract specification, customer selection, and payment systems.

To visualize some of our findings, we will fit categorically based Kaplan-Meier curves and plot them, allowing us to see differences in churn rate between customer categories.

4.2.1.1 Contract Specification

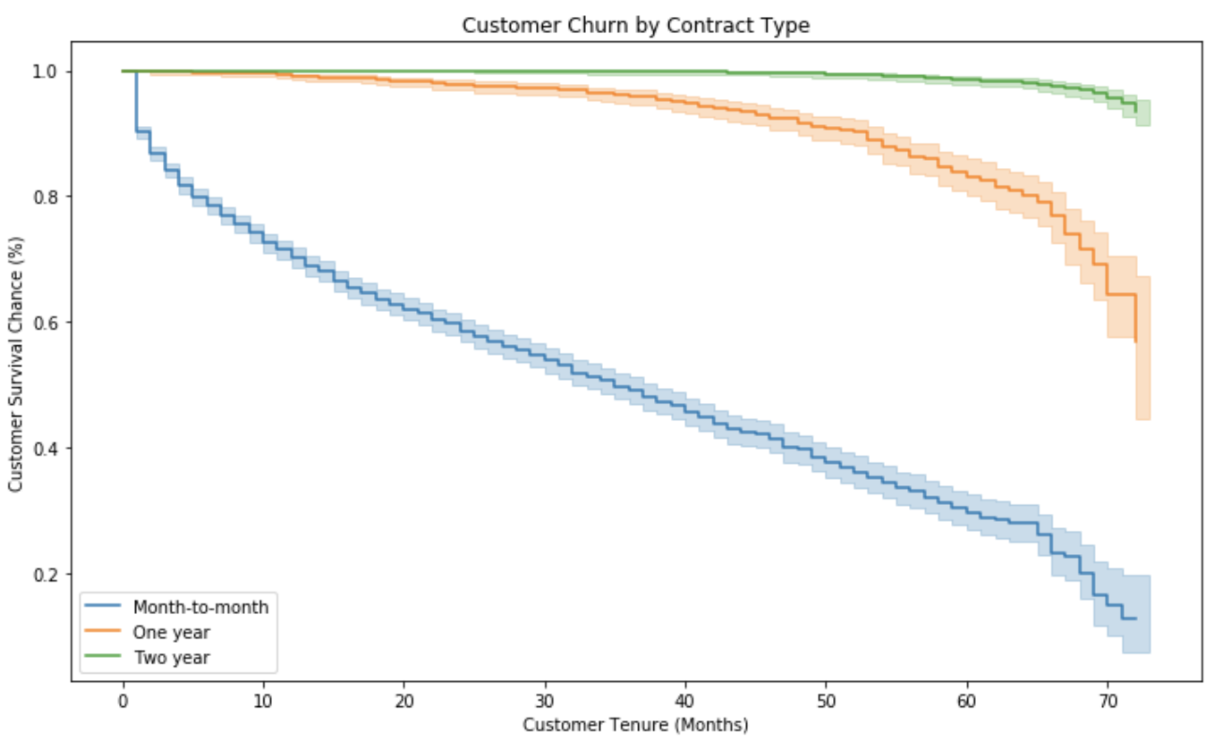

Fig 4.3 Kaplan Meier Plot for Customer Churn by Contract Type

The most important feature, by far, is the presence of a 1 or 2 year contract. Customers are .25 and .02, respectively, times as likely to cancel their service if they are under contract.

Cancellation fees are a possible underlying cause. As long as these fees do not prohibit new sales, we would recommend continuing to put them into as many contracts as possible.

4.2.1.2 Customer Selection

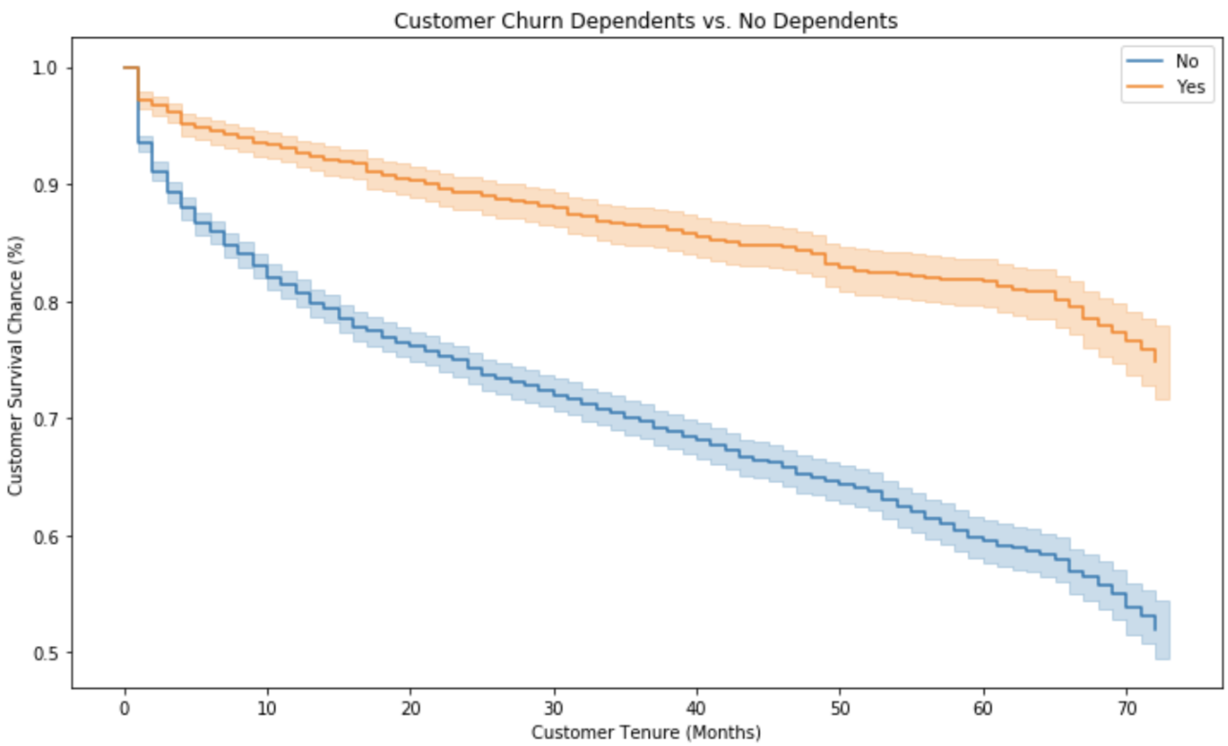

Fig 4.4 Kaplan Meier Plot for Customer Churn for customers with/without dependents

Customers with a partner or dependents are .82 and .91 times as likely to cancel as normal customers.

Families and other large households seem to be less likely to change providers. This could be due to higher incomes, less time to consider options, or another combination of factors.

4.2.1.3 Payment Systems

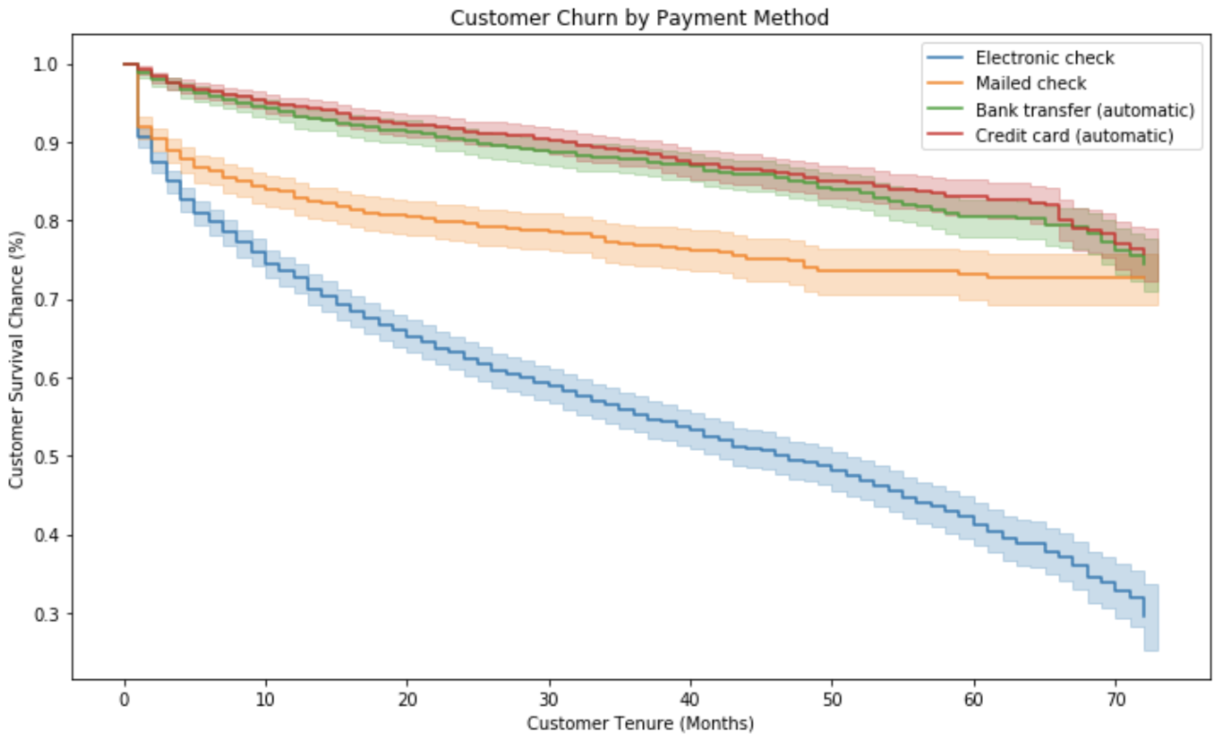

Fig 4.5 Kaplan Meier Plot for Customer Churn by Payment Method

There is a reason companies now default to opting employees into 401k plans. It takes effort for people to make a change, even if it is beneficial.

Make sure your customer's default is an automatic payment made monthly. This requires little effort from the customer to remain subscribed.

Conversely, sending a check, in the mail or electronically, is a pain. It requires effort to remain subscribed.

4.3 Survival Regression using Cox Proportional Hazard Model

To examine the effects of different features, we will use the Cox Proportional Hazards Model. We can think of this as a Survival Regression model.

'Hazards' can be thought of as something that would increase/decrease chances of survival. In our business problem, for example, a hazard may be the type of contract a customer has. Customers with multi-year contracts probably cancel less frequently than those with month-to-month contracts. One restriction is that the model assumes a constant ratio of hazards over time across groups. The library called "lifelines" offers a built-in check_assumptions method for the CoxPHFitter object.

After some data cleaning, including encoding categorical variables (k-1 dummies), we can fit a survival regression model to the data.

```
<lifelines.CoxPHFitter: fitted with 7043 observations, 5174 censored>
      duration col = 'tenure'
         event col = 'Churn'
number of subjects = 7043
  number of events = 1869
    log-likelihood = -12688.70
  time fit was run = 2019-01-23 02:40:53 UTC

---
                                      coef  exp(coef)  se(coef)      z      p  log(p)  lower 0.95  upper 0.95
SeniorCitizen                         0.03       1.03      0.06   0.60   0.55   -0.60       -0.08        0.14
Partner                              -0.19       0.82      0.06  -3.52 <0.005   -7.77       -0.30       -0.09   **
Dependents                           -0.10       0.91      0.07  -1.39   0.17   -1.80       -0.23        0.04
OnlineSecurity                       -0.38       0.68      0.07  -5.65 <0.005  -17.94       -0.51       -0.25  ***
OnlineBackup                         -0.29       0.75      0.06  -5.22 <0.005  -15.54       -0.40       -0.18  ***
DeviceProtection                     -0.16       0.85      0.06  -2.85 <0.005   -5.44       -0.27       -0.05    *
TechSupport                          -0.28       0.76      0.07  -4.19 <0.005  -10.50       -0.41       -0.15  ***
StreamingTV                          -0.27       0.77      0.06  -4.46 <0.005  -11.69       -0.38       -0.15  ***
StreamingMovies                      -0.26       0.77      0.06  -4.36 <0.005  -11.27       -0.38       -0.14  ***
PaperlessBilling                      0.16       1.17      0.06   2.79   0.01   -5.25        0.05        0.27    *
MonthlyCharges                        0.07       1.07      0.00  26.59 <0.005 -356.98        0.06        0.07  ***
TotalCharges                         -0.00       1.00      0.00 -40.10 <0.005    -inf       -0.00       -0.00  ***
MultipleLines_No phone service        0.64       1.89      0.12   5.50 <0.005  -17.07        0.41        0.87  ***
MultipleLines_Yes                    -0.20       0.82      0.05  -3.68 <0.005   -8.38       -0.30       -0.09   **
Contract_One year                    -1.40       0.25      0.10 -13.78 <0.005  -97.83       -1.60       -1.20  ***
Contract_Two year                    -4.05       0.02      0.20 -20.74 <0.005 -218.31       -4.43       -3.66  ***
PaymentMethod_Credit card (automatic) -0.01      0.99      0.09  -0.06   0.95   -0.05       -0.18        0.17
PaymentMethod_Electronic check        0.38       1.46      0.07   5.20 <0.005  -15.45        0.24        0.52  ***
PaymentMethod_Mailed check            0.52       1.68      0.09   5.96 <0.005  -19.81        0.35        0.69  ***
---
Signif. codes: 0 '***' 0.0001 '**' 0.001 '*' 0.01 '.' 0.05 ' ' 1

Concordance = 0.93
Likelihood ratio test = 5928.67 on 19 df, log(p)=-inf
```

Fig 4.6 Cox Proportional Hazard Model Summary

In the above regression, the key output is exp(coef). This is interpreted as the scaling of hazard risk for each additional unit of the variable, 1.00 being neutral.

For example, the last exp(coefficient), corresponding to PaymentMethod_Mailed check, means a customer that pays by mailing a check is 1.68 times as likely to cancel their service. For the company, exp(coef) below 1.0 is good, meaning a customer is less likely to cancel.

To better visualize the above, we can plot the coefficient outputs and their confidence intervals:

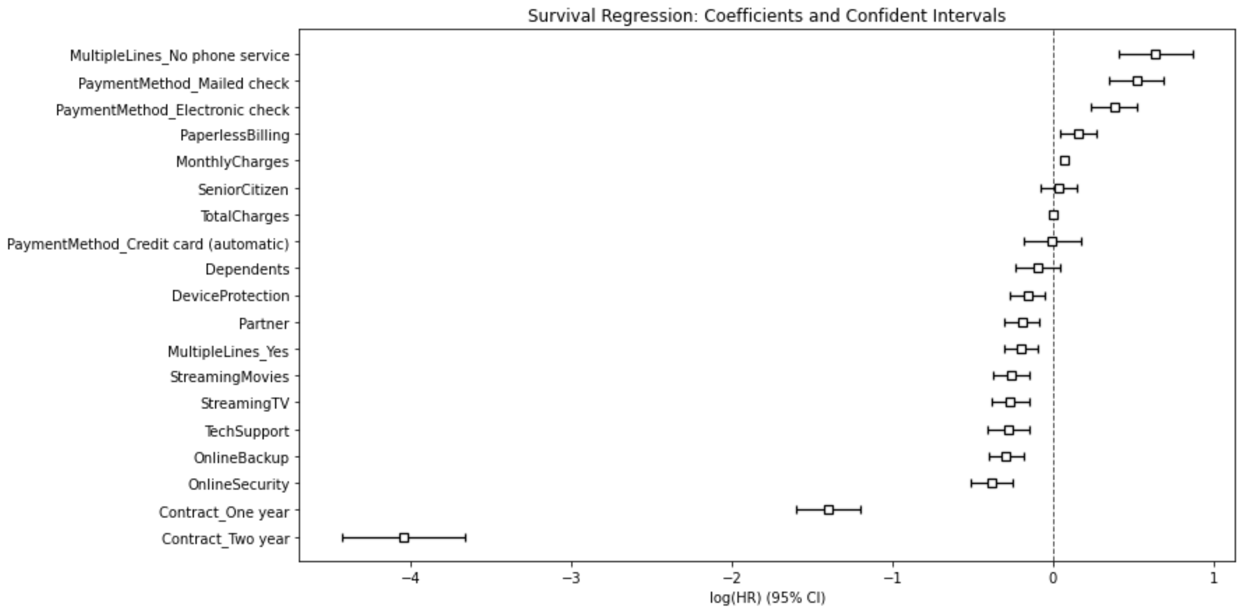

Fig 4.7 Boxplot of Cox Confidence intervals for each feature

4.4 Survival Regression using XGBoost+WeibullAFT

Churn management can be done with a combination of Customer Lifetime Value and a churn risk score.

CLTV: *The present value of the **future** cash flows attributed to the customer during his/her entire relationship with the company.* Computing CLTV has 4 major steps:

1. Forecast Customer lifetime
2. Forecast future revenue
3. Forecast future cost
4. Compute the discounted value of future cash flows

The following equation gives the formula for computing the CLTV of a subscriber.

$$CLV = \sum_{t=1}^{T} (ARPU_t - Cost_t) \frac{P_i(t)}{\left(1 + \frac{r}{12}\right)^{t-1}}$$

In this equation, $P_i(t)$ is the series of survival probabilities or 1 – the risk of churn for a subscriber forecasted at different time points in the future.

We describe the methodology to estimate these sequential churn risks based on the Weibull accelerated failure model.

4.4.1 Input Data

Each row in the training data is a subscriber observed at a specific historical date during his tenure. This is the column ‘car snapshot date’ in the table.

We also keep track of the churn date or the run date of the training data preparation job. The difference between either of these and the car snapshot date gives the remaining lifetime in days. A data point is censored if the subscriber has survived by the job run date, else it is uncensored. The column censored is 1 or 0 depending on whether the data row corresponds to a censored observation or not. The pair of columns “Remaining lifetime” and “censored” form the response for the model.

The covariates are specified by the FEATURES group and fetched from CAR history. Only those subscribers who have spent at least m months in the network will be considered for training the model.

Table 4.3 Training data example

		FEATURES					RESPONSE	
car snapshot date	msisdn	AVG_ZERO_BAL_MONETORY_DAY_PER_MTH	ZERO_BAL_MONETORY_DAY_MTD	AVG_ZERO_BAL_VOICE_DAY_PER_MTH	ZERO_BAL_VOICE_DAY_MTD	Churn Date / Job run date	Remaining Lifetime (days)	censored
2021-03-05	1	3	2	10	3	2021-03-11	6	1
2021-02-03	1	7	6	2	4	2021-03-11	36	1
2021-01-20	1	8	0	1	3	2021-03-11	50	1
2020-12-19	1	7	0	9	3	2021-03-11	82	1
2020-11-28	1	1	4	10	10	2021-03-11	103	1
2020-11-07	1	9	2	1	8	2021-03-11	124	1
2021-03-30	4	3	2	10	3	2021-03-31	1	0
2021-02-28	4	7	6	2	4	2021-03-31	31	0
2021-01-29	4	8	0	1	3	2021-03-31	61	0

2020-12-30	4	7	0	9	3	2021-03-31	91	0
2020-11-30	4	1	4	10	10	2021-03-31	121	0
2020-10-31	4	9	2	1	8	2021-03-31	151	0

4.4.2 Training XGBoost and Feature Importances

Some of the features in the training data are pre-processed before the training phase. One-hot encoding is performed for categorical features having cardinality lesser than the threshold mentioned by the input parameter. Label encoding is performed for features having cardinality greater than the threshold.

Here is an example of One-hot encoding.

ID	Gender
1	Male
2	Female
3	Not Specified
4	Not Specified
5	Female

ID	Male	Female	Not Specified
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1
5	0	1	0

Fig 4.8 Illustration of one-hot encoding

An example of Label encoding.

	occupation
0	programmer
1	data scientist
2	engineer
3	manager
4	ceo

Label Encoding

	occupation
0	4
1	1
2	2
3	3
4	0

Fig 4.9 Illustration of Label Encoding

The Average survival time predicted by the XGBoost model comes out as 594 days.

On training the XGBoost model, the following are the feature importances:

{'SeniorCitizen': 48.0,
'Partner': 30.0,
'Dependents': 15.0,
'OnlineSecurity': 48.0,
'OnlineBackup': 27.0,
'DeviceProtection': 18.0,

'TechSupport': 49.0,
'StreamingTV': 17.0,
'StreamingMovies': 43.0,
'PaperlessBilling': 49.0,
'MonthlyCharges': 527.0,
'TotalCharges': 494.0,
'MultipleLines_No phone service': 27.0,
'MultipleLines_Yes': 28.0,
'Contract_One year': 37.0,
'Contract_Two year': 35.0,
'PaymentMethod_Credit card (automatic)': 20.0,
'PaymentMethod_Electronic check': 42.0,
'PaymentMethod_Mailed check': 19.0}

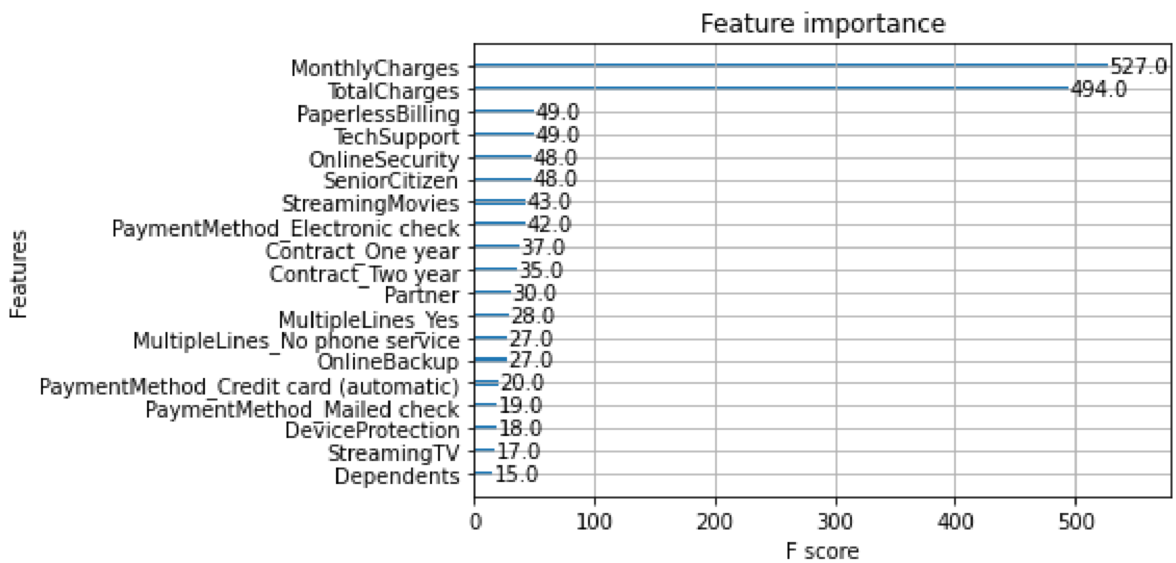

Fig. 4.10 XGBoost Feature Importance plot

4.4.3 Survival Regression using AFT

A survival regression model will be built to predict the remaining customer lifetime. A specific family of survival regression called the Accelerated Failure time model has been chosen.

4.4.3.1 Weibull Accelerated Failure Time model

Let t_i be the censored/uncensored survival time of subscriber i. Denote $x_i \in R^p$ to be the p-dimensional feature vector for subscriber i.

The accelerated failure time model is specified as:

$$\log log\ t_i = \beta' x_i + \sigma\epsilon_i$$

where $\beta \in R^p$ is the vector of regression coefficients, σ is a scale parameter and ϵ_i is i.i.d distributed according to some probability distribution.

The specific distribution chosen for this is called a Gumbel distribution, which has the p.d.f.

$$f(x) = e^x e^{-e^x}$$

This model is called a Weibull Accelerated Failure time model.

4.4.3.2 Training AFT model

The training data will be split into a train and validation split. The model will be trained on the training split and the model metrics will be computed on the validation split. The ratio of train test validation will be controlled by a user-specified configuration parameter.

model	lifelines.WeibullAFTFitter
duration col	'duration'
event col	'event'
number of observations	5634
number of events observed	1496
log-likelihood	-6291.78
time fit was run	2022-04-28 15:08:36 UTC

		coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
lambda_	risk	0.02	1.02	0.00	0.02	0.02	1.02	1.02	38.47	<0.005	inf
	Intercept	2.03	7.62	0.03	1.97	2.09	7.17	8.09	65.93	<0.005	inf
rho_	risk	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-30.90	<0.005	694.15
	Intercept	0.37	1.45	0.02	0.33	0.41	1.39	1.51	17.92	<0.005	236.12

Concordance	0.98
AIC	12591.57
log-likelihood ratio test	4382.62 on 2 df
-log2(p) of ll-ratio test	inf

Fig. 4.11 Weibull AFT fitted model summary

4.4.3.3 Loss Function

The following loss function is minimized for a Weibull AFT model.

$$L(\beta, \sigma; y_i) = \prod_{i=1}^{n} [f_Y(y_i)]^{\delta_i} [S_Y(y_i)]^{1-\delta_i}$$
$$= \prod_{i=1}^{n} \left\{ \frac{1}{\sigma} \exp(\frac{y_i - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma}) \exp\left[- \exp(\frac{y_i - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma}) \right] \right\} \left\{ \exp\left[- \exp(\frac{y_i - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma}) \right] \right\}$$

4.4.3.4 Model Performance

Model performance will be evaluated based on the following 2 metrics. Corresponding to every run of the model the following 2 performance metrics will be reported.

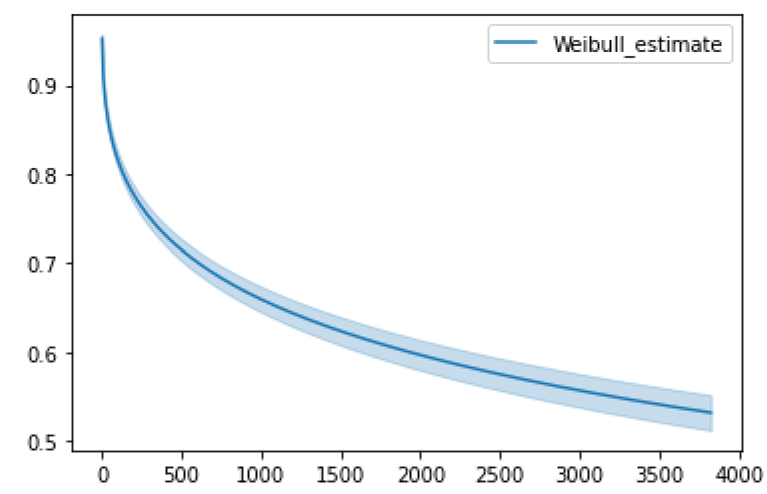

Fig. 4.12 Weibull Estimate

1) **Concordance Index (c-index)**

Assume T_i and y_i are the observed and predicted lifetimes of subscriber i. Assume δ_i is the indicator of churn. The concordance index is then given by the following formula:

$$c - index = \frac{\sum_{i,j} 1_{T_j < T_i} 1_{y_j < y_i} \delta_j}{\sum_{i,j} 1_{T_j < T_i} \delta_j}$$

The value of the c-index ranges from 0 to 1. A model with random guesses will have a c-index of about 0.5. A model with c-index greater than 0.5 performs better than a random. A c-index of 1 indicates a perfect model.
The concordance index for the Weibull AFT fitted model comes out to be 0.98 whereas the concordance index for trained XGBoost model was 0.967

2) **Brier score**

Another metric of importance is the Brier score. The formula is given below. A useful model will have a Brier score below 0.25.

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left(0-\hat{S}(t,\vec{x}_i)\right)^2 \cdot 1_{T_i \leq t, \delta_i = 1}}{\hat{G}(T_i^-)} + \frac{\left(1-\hat{S}(t,\vec{x}_i)\right)^2 \cdot 1_{T_i > t}}{\hat{G}(t)}\right)$$

4.4.3.5 Validation

Building good ML models requires these models to generalize well to unseen data. In other words, the model will be trained on an offline dataset and then deployed into production. Some guarantees of the performance in production need to be provided a-priori, otherwise, there might be catastrophic results because of deploying a sub-optimal model.

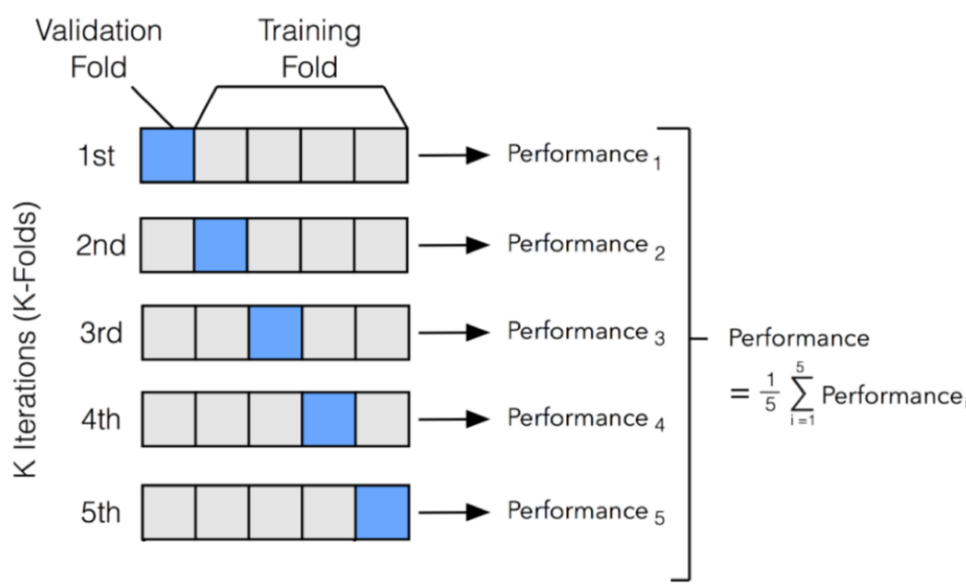

Fig 4.13 Illustration of a 5-fold cross-validation scheme

Validation of the ML model is required to provide some guarantee of performance on unseen data during the training process itself. Every model, with its own set of hyper-parameters, is required to report its generalization performance. To report the generalization performance of a model, the training process separates out a small random portion of the dataset and trains using the rest of the data. Once trained, the performance is reported on the small held-out set. To ensure that one specific random draw does not produce skewed results, this process can be repeated multiple times in a process called K-fold cross-validation. Both these validation methods are supported by the framework. Figure 3 and Figure 4 present a visual illustration of these 2 schemes of validation.

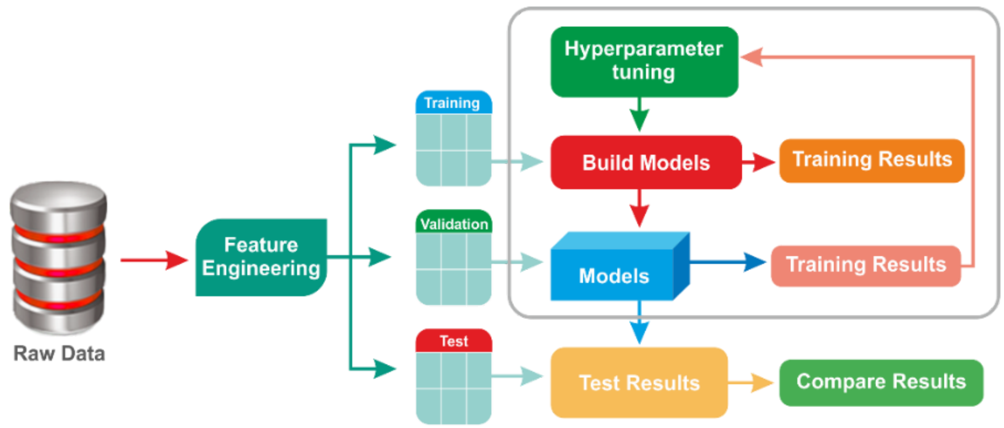

Fig 4.14 Illustration of a train-test validation scheme

4.4.3.6 Explainable AI using SHAP

Model explainability becomes a basic part of the machine learning pipeline. Keeping a machine learning model as a "black box" is not an option anymore. Luckily there are tools that are evolving rapidly and becoming more popular. This guide is a practical guide for XAI analysis of SHAP open-source Python package for a regression problem.

SHAP (Shapley Additive Explanations) by Lundberg and Lee (2016) is a method to explain individual predictions, based on the game theoretically optimal Shapley values. Shapley values are a widely used approach from cooperative game theory that come with desirable properties. The feature values of a data instance act as players in a coalition. The Shapley value is the average marginal contribution of a feature value across all possible coalitions

Force Plot:

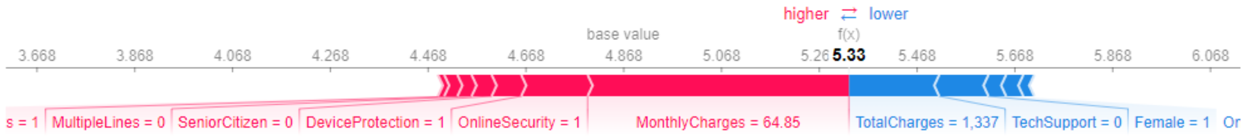

Fig 4.15 SHAP Force Plot

The output value is the prediction for that observation. Features that push the prediction higher (to the right) are shown in pink, and those pushing the prediction lower are in blue.

Feature Importance Summary plots:

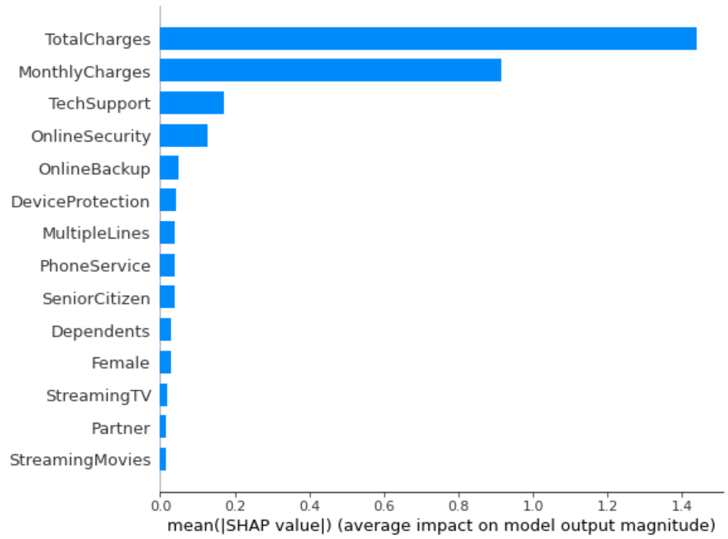

Fig 4.16 Feature importance bar plot

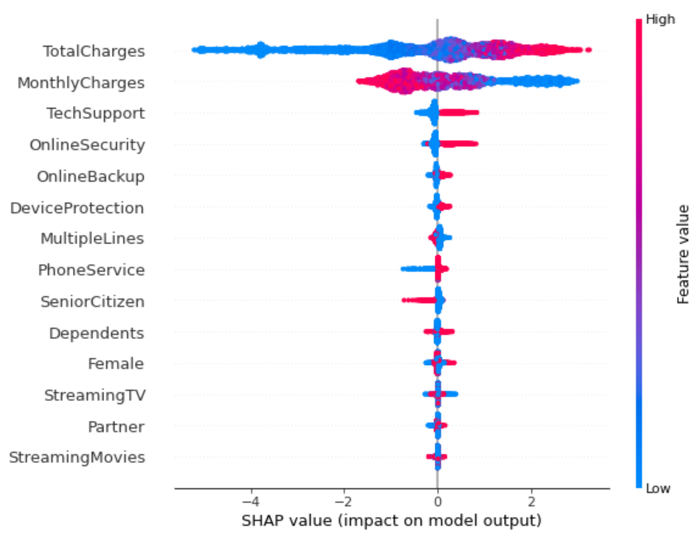

Fig 4.17 Feature importance beeswarm plot

The beeswarm plot is designed to display an information-dense summary of how the top features in a dataset impact the model's output. Each instance of the given explanation is represented by a single dot on each feature fow. The x position of the dot is determined by the SHAP value of that feature, and dots "pile up" along each feature row to show density. Color is used to display the original value of a feature.

Feature Dependence plots:

This shows how the model depends on the given feature, and is like a richer extension of the classical partial dependence plots. Vertical dispersion of the data points represents interaction effects.

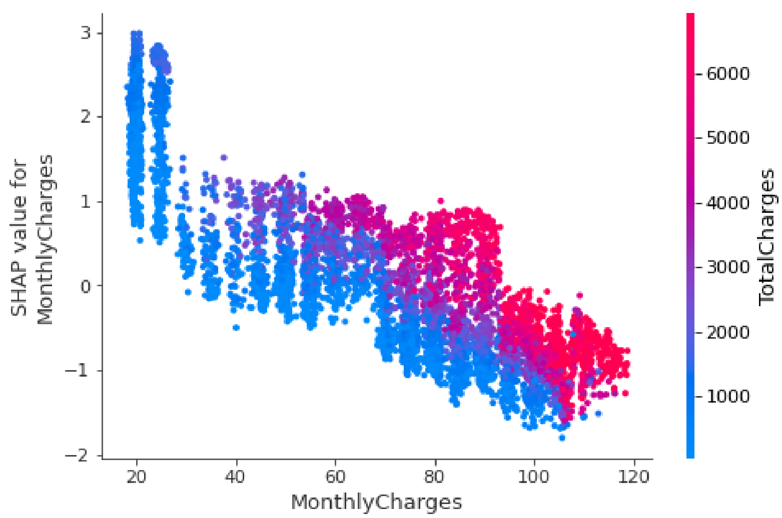

Fig 4.18 SHAP Dependence plot for Monthly Charges w.r.t Total Charges

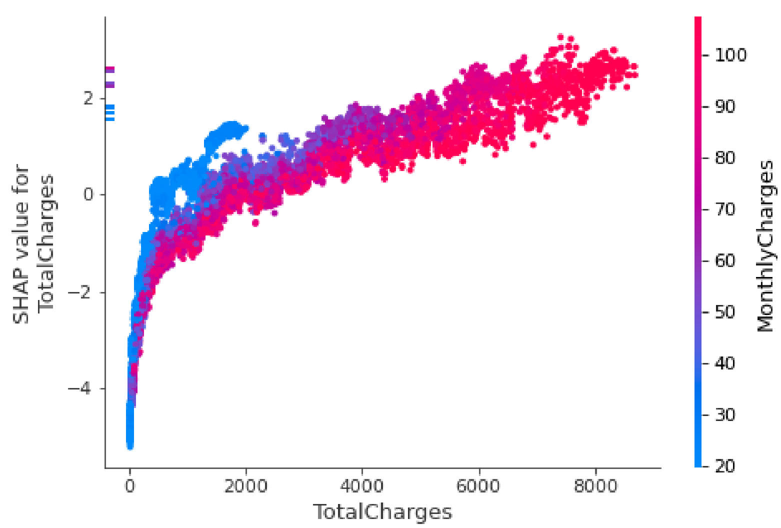

Fig 4.19 SHAP Dependence plot for TotalCharges w.r.t Monthly Charges

Chapter 5
Conclusion

With the existing technology and statistical methodology available today, Survival Analysis is the best way to go in order to perform churn prediction, survival time estimation and related analyses because there are some or the other associated limitations with other methods and models. It provides a comprehensive explanation at every step and has good library support in several programming languages.

Survival analysis, when combined with machine learning methods gives outstanding results as done in this project. After preparing our data, we went ahead with estimating the Kaplan-Meier plot as well as plotting the cumulative hazard graph using the Nelson-Aalen method. We also made recommendations to prevent telecom customer churn on the basis of the Kaplan-Meier curves plotted for particular covariables on which the churn depends. Kaplan-Meier estimation is a univariate analysis and tells us about the entire population as a whole or particular groups of population and their behavior.

Next, we went ahead to find out which features impact the desired event, i.e. the churn by performing Survival regression using the Cox Proportional Hazard model. We found out the performance on the basis of the concordance intex and observed which features are the ones impacting the churn most.

Then,, we performed a combination of the Extreme Gradient Boosting (or XGBoost) machine learning regression model and fitted the output of this model in a Weibull Accelerated Failure time model, giving us an optimum result and high performance. We also plotted the feature importances after training the XGBoost and found them to be in agreement with those from the Cox analysis.

Finally, we went ahead with explaining our analysis using “Explainable AI ', specifically SHAP (SHapley Additive exPlanations)) and plotting all our findings. which is a set of tools and frameworks to help you understand and interpret predictions made by your machine learning models. It can help debug and improve model performance, and help others understand a model’s behavior.

A lot of the work was also writing functions to perform the tasks that open-source python libraries do because during production, we cannot rely on constantly evolving and maintained open-source libraries as they may be discontinued at any time by the developers.

Chapter 6
Future Prospects

This project can be extended to use Deep Learning methods and use Artificial Neural Networks for churn prediction and survival time estimation. One can dig deeper into Neural Networks and experiment with them to find what works best for churn prediction.

In addition to the Explainable AI implemented in this project, what can be taken into account to authenticate the predictions to decide how good a model can be the Global and Local explainibilities of the model- the former revealing features important for the overall predictions and the latter telling us about which feature was more important for every individual prediction for a particular subscriber.

Furthermore Machine Reasoning may be used to automate this complex and highly subjective task of detecting “churners”. Machine Reasoning is a new category in AI/ML technologies that can enable a computer to work through complex processes that would otherwise require a human. It overcomes certain shortcomings of Machine Learning technologies to provide better and more efficient solutions.

Finally, The Cloud Computing Platform can be used to store and execute tasks will be Amazon Web Services- Amazon SageMaker Studio (an IDE for Machine Learning), Amazon Simple Queue Service or SQS (a fully managed message queuing service for future scalability), Amazon SageMaker Feature Store (a purpose-built repository to store, update, retrieve, share machine learning features) will be used.

References

[1]Gunjan Bansal, Adarsh Anand & V. S. S. Yadavalli (2019): Predicting effective customer lifetime: an application of survival analysis for telecommunication industry, communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2019.1570264

[2]Junxiang Lu, Ph.D. Overland Park, Kansas; Modeling Customer Lifetime Value Using Survival Analysis — An Application in the Telecommunications Industry, Paper 120-28, SUGI 28 Data Mining Techniques

[3]Junxiang Lu, Ph.D. Sprint Communications Company Overland Park, Kansas; Predicting Customer Churn in the Telecommunications Industry –– An Application of Survival Analysis Modeling Using SAS; Predicting Customer Churn in the Telecommunications Industry –– An Application of Survival Analysis Modeling Using SAS; SUGI 27 Data Mining Techniques

[4]Barry E. King et Jennifer Rice; Analysis of Churn in Mobile Telecommunications: Predicting the Timing of Customer Churn - AIMS International Journal of Management · August 2019

[5]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/

[6]https://xgboost.readthedocs.io/en/stable/tutorials/aft_survival_analysis.html

[7]https://arxiv.org/abs/2006.04920

[8]https://www.ucl.ac.uk/short-courses/search-courses/survival-analysis-time-event-data-introduction

[9]https://en.wikipedia.org/wiki/Survival_analysis

[10]https://www.questionpro.com/blog/customer-churn/

[11]https://www.ngdata.com/what-is-customer-churn/

[12]https://www.qualtrics.com/au/experience-management/customer/customer-churn/

[13]https://www.salesforce.com/resources/articles/how-calculate-customer-churn-and-revenue-churn/

[14]https://www.productplan.com/glossary/churn/

[15]https://www.profitwell.com/customer-churn/guide

[16]https://www.youtube.com/watch?v=v1QqpG0rR1k&t=164s

[17]https://www.ericsson.com/en/blog/2021/4/the-characteristics-of-consumer-telecom-bss-stacks-will-prevail-for-enterprises

[18]https://www.ericsson.com/en/telecom-bss

[19]http://www.biecek.pl/statystykaMedyczna/Stevenson_survival_analysis_195.721.pdf

[20]https://medium.com/@zachary.james.angell/applying-survival-analysis-to-customer-churn-40b5a809b05a

[21]https://www.iiap.res.in/astrostat/LecPresentations/JogeshBabu_TruncationCensoring.pdf

---