KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

DECODING STUDENT RETENTION AND CHURN OF VODAFONE (TELECEL) IN KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY (KNUST): A SURVIVAL ANALYSIS APPROACH

MUSAH FARIDU OUBDA

KASSIM ASANA

SARPONG LINDA

TORSI EDMOND COLLINS

ASIAMAH EZEKIEL

A THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF BACHELOR OF SCIENCE IN ACTUARIAL SCIENCE

AUGUST, 2024

Declaration

We, therefore, affirm that this paper is entirely our own original work and has not been submitted in whole or in part for another degree award in any other university. Due credit has also been given to previously published works and online publications through referencing.

Musah Faridu Oubda . . . .

Index Number (4325620) Signature Date

Kassim Asana . . . .

Index Number (4323020) Signature Date

Sarpong Linda . . . .

Index Number (4330720) Signature Date

Torsi Edmond Collins . . . .

Index Number (4331420) Signature Date

Asiamah Ezekiel . . . .

Index Number (4316720) Signature Date

Certified by:

Sandra Addai-Henne . . . .

Supervisor Signature Date

Certified by:

Prof. Gabriel Asare Okyere . . . .

Head of Department Signature Date

Abstract

This study examines the factors influencing student churn at Kwame Nkrumah University of Science and Technology (KNUST), with a particular focus on the telecommunications services provided by Vodafone (now Telecel). The study was centered on students in their 4th year in the College of Science. Utilizing survival analysis methods, the Cox Proportional Hazards (PH) model and the Lognormal Accelerated Failure Time (AFT) model were used to identify key predictors of churn and analyze the survival rates of the 4th year students across different academic levels. The findings indicate that poor network quality significantly increases both the churn rate and the time to churn. While the Cox PH model demonstrates superior predictive value (0.96 concordance value), the Lognormal AFT model offers a better overall fit (290.678 loglikehood ratio). It was also noticed that, students churned more at the end of their 3rd year. Poor Network Quality was the covariate that affected the increase in churn rate and accelerated the time to churn the most. The study concludes with recommendations for reducing churn and outlines directions for future research.

Acknowledgment

Dedication

We dedicate this project to the Almighty God and our parents, who have seen us through all of our challenges during this thesis period and has brought us this far. We also dedicate this effort to our families and friends, who have been there for us every step of the way with love and prayers. Finally, we dedicate this presentation to our supervisor, who stood by us through it all.

Table of Content

List of Tables

List of Figures

List of Abbreviations

ACE   Africa Coast to Europe

AFT   Accelerated Failure Time

AIC   Akaike Information Criterion

AT   AirtelTigo

BIC   Bayesian Information Criterion

CDF   Cumulative Distribution Function

CPHM   Cox Proportional Hazard Model

CPH   Cox Proportional Hazard

C-index   Concordance Index

GSES   Ghana Satellite Earth Station

HQ   Hannan-Quinn Criterion

KM   Kaplan-Meier

KNUST   Kwame Nkrumah University of Science and Technology

PDF   Probability Density Function

PH   Proportional Hazards

RSF   Random Survival Forest

SA   Survival Analysis

SAT3   South Atlantic 3

S(t)   Survival Function

SMS   Short Message Service

UITS   University Information Technology Services

WACS   West Africa Cable System

h(t)   Hazard Function

Chapter 1

Introduction

1.1 Background of Study

Ghana's telecommunications industry has experienced significant growth in recent years, with companies such as Vodafone playing a crucial role in providing mobile and internet services to people across the country [Bandim, 2022]. The industry is highly competitive, making customer retention vital for sustaining market share and profitability. Obtaining new customers is more expensive than retaining existing ones due to marketing activities, incentives and campaigns involved. Therefore, retaining a customer is preferable to acquiring a new one. Customer churn, also known as customer attrition, refers to the loss of subscribers or customers who cease using a company's service or product within a given period. Understanding the reasons behind customer churn helps to develop strategies to improve customer retention and help reduce churn rates in the long run.

Vodafone is one of the leading national telecommunications providers in Ghana. As of January 2020, it had over 9.3 million mobile voice subscribers, representing 13.81% of Ghana's market share. Since becoming the majority shareholder, Vodafone Ghana has been operating the Ghana Satellite Earth Station (GSES) since 2008 [Wikipedia, 2020]. GSES allows Ghana to access and utilize communications satellites orbiting the Earth for various applications, such as telephone services, internet connectivity, television broadcasting, data transmission, disaster management emergency communications. The operation of the earth station by Vodafone Ghana proves the company's commitment to investing in and upgrading Ghana's satellite communications capabilities.

In 2016, Vodafone partnered with Kwame Nkrumah University of Science and Technology (KNUST) to provide enhanced packages of services to the various faculties across the university's campuses to improve education services [Wikipedia, 2020]. This collab-

oration included telecommunications services such as SIM cards and data plans for the student and employee communities.

In February 2023, Telecel acquired 70% shares in Vodafone, rebranding the company name to Telecel in 2024 [Wikipedia, 2024]. This rebranding aimed to improve service offerings across voice and data services, money transfers business solutions. Telecel, founded in 1986, is an Africa-focused telecommunication company that operates primarily in Africa and converges telecommunications with fintech, e-commerce tech startups.

Student churn is a major issue every telecom company encounters. It leads to a loss of revenue and increases the cost of acquiring new customers. In the highly competitive telecom market, where customers have multiple service provider options, retaining customers becomes even more challenging. Companies use modern technology, computer software survival analysis approaches to identify at-risk students and devise strategies to enhance retention rates. This method of recognizing unsatisfied customers is known as churn prediction.

On March 14, 2024, Vodafone, along with several other telecommunications companies, was hit by outages on several underwater fiber optic cables, leading to disruptions in services, particularly internet services [Ghana Web, 2024]. It affected about 10 countries in West Africa, including Ghana. Initially, it was estimated that the problem would be fixed within 3 days; however, this was not the case. The damage was massive, affecting the West Coast route to Europe, the West Africa Cable System (WACS) the Africa Coast to Europe (ACE), resulting in MainOne and South Atlantic 3 (SAT3) going offline [AP News, 2024]. The only network that was properly functioning at the moment in Ghana was AirtelTigo (AT). To quickly prevent the loss of valuable customers to AirtelTigo, Vodafone (Telecel) took the initiative to update its customers daily on their progress and offer various bonuses to prevent customers from defecting to their competitors. This continued until the problem was fixed on April 29, 2024, when the WACS cable was repaired. This study aims to focus on the KNUST student population, uncover and analyze data gathered from students, apply survival analysis models to identify patterns related to churn (such as demographics, usage patterns, etc.), develop retention strategies, evaluate

the success of these efforts and make informed decisions.

1.2   Problem Statement

Despite KNUST's efforts to partner with Telecel for affordable and accessible mobile communication services, student churn remains a persistent issue. The core problem lies in the insufficient understanding of the factors influencing student churn and retention, which hampers the development of effective strategies to tackle the issue. This research aims to address this gap by investigating the factors contributing to student churn and developing a survival analysis model to identify students at risk. The goal is to create targeted strategies that improve retention rates based on the model's insights.

The existing gap in knowledge regarding the precise reasons behind student churn undermines the ability to devise targeted interventions that could effectively address the issue. As a result, the strategies that have been implemented so far have not achieved the desired outcomes in reducing churn rates. This research seeks to bridge this gap by thoroughly investigating the various factors contributing to student churn and by developing a robust survival analysis model to accurately identify students who are at risk of churning. The objective is to leverage the insights gained from this model to design and implement targeted strategies aimed at improving retention rates. By addressing these issues, the research aims to facilitate an understanding of churn dynamics and ultimately enhance the effectiveness of retention efforts.

1.3   Research Objectives

The objectives of the research are;

1. To use survival analysis to estimate student churn rates at various academic years.

2. To determine which covariates affect the churn rate.

3. To provide strategies to combat churn.

3

1.4 Significance of the Study

The study offers a comprehensive understanding of the factors impacting student churn from Vodafone. This will enable the development of precise strategies to enhance retention rates, thereby minimizing churn. The study will empower KNUST to improve the telecom services offered to its students. By identifying the factors influencing student retention and churn, KNUST can work closely with Telecel to guarantee the delivery of top-notch, dependable services to its students, thereby solidifying the partnership between KNUST and Telecel. The study is all about understanding what the populace expects from Vodafone (Telecel). By knowing their specific needs and preferences, the services can be improved upon. This means improved connectivity, service plans less hassle from switching providers. Ultimately, it ensures a more stable and reliable service for students.

1.5 Organization of the Study

The study on student retention and churn for Telecel services at KNUST aims to investigate the factors that drive students' decisions on Telecel services usage. The study follows a structured approach. Chapter one discusses the background information and the foundation for the research. The problem statement and goals are stated here. Chapter two of the study contains a literature review, including several studies related to the subject at hand. The third chapter focuses on the methodology and data collection techniques. There are several ways of approaching this problem, but the main focus of the study is using a survival analysis modeling tool. The fourth chapter presents the models and practical application to the dataset. Detailed analysis is presented here with the aid of figures and diagrams to determine the optimal model for the research. Chapter five delves into the conclusion and summarizes the results obtained. Based on the findings, recommendations are formulated.

1.6 Limitation of the Study

The research specifically concentrates on students at the College of Science at KNUST. As a result, it does not encompass students from other colleges within the university. This limitation means that the findings and insights derived from this study are relevant primarily to the College of Science and may not be fully applicable to students in other colleges.

Chapter 2

Literature Review

2.1 Introduction

Mobile technology has become widely used, changing how students interact, communicate obtain information. Student retention and churn are critical issues for universities and telecom companies alike. In the context of Vodafone's services offered to students, understanding student behavior and preferences is crucial for effective retention strategies. This project addresses this gap using survival analysis, a statistical method for examining data over time [Box-Steffensmeier and Jone, 2020]. In this context, the event of interest is student churn, defined as the discontinuation or cessation of Vodafone SIM cards. By applying survival analysis, this project will provide valuable insights into student retention and churn factors.

2.1.1 Telecommunication Industry and Customer Behavior

The telecommunication industry is highly competitive, with customer acquisition and retention being pivotal to a company's success. Research indicates that factors such as service, pricing, customer support competitive offerings significantly influence customer churn [Lai, 2021].

2.1.2 Student Retention and Telecom Services

Students, as a demographic, exhibit unique usage patterns and service expectations. Studies show that reliable connectivity, affordable pricing additional benefits like educational resources influence their choice of telecom services [Smith and Brown, 2020]. For telecom companies, understanding these factors is crucial to tailoring services that meet

students' needs and enhance retention.

2.1.3 History of Survival Analysis (SA)

The origins of SA and its history spread far back to the early work on mortality by John Graunt, who published the book "Natural and Political Observations on the Bills of Mortality" in 1977. The concept of "Life Tables" was introduced in this book [Graunt, 1977]. A new, modern era in SA started during World War II in the USA, where the reliability of military equipment was tested using SA. After World War II, SA became popular and spread to various other disciplines [Jerenz, 2008]. The most influential papers on SA were published by Kaplan and Meier [Kaplan and Meier, 1958] and Cox [Cox et al., 1972], in which the Kaplan-Meier product limit estimator and Cox proportional hazard model were introduced, respectively.

2.1.4 Application of Survival Analysis

Survival analysis (SA) has been successfully applied in diverse fields, including medical research, engineering (reliability analysis), finance and economics. In finance, SA has been used to analyze a firm's vulnerability to global financial crises and identify firms at risk of failure, enabling timely interventions [Lee, 2014, Pereira, 2014, Kumar and Gepp, 2015, Iwasaki, 2014]. SA has also been employed to predict the optimal time to buy or sell stocks within the stock market [Kumar and Gepp, 2015]. In the banking sector, SA is utilized to assess banks' survival during financial crises, measure their strength when entering new markets [Leung et al., 2003, Evrensel, 2008] and conduct credit risk analysis [Tsujitani and Baesens, 2012, Baesens et al., 2005]. The widespread adoption of SA across these domains underscores its versatility and effectiveness in modeling time-to-event data and identifying the factors that influence outcomes of interest.

2.2   Conceptual Review

Customer retention refers to a company's capacity or ability to maintain or retain their customer base over a specific period. High retention rates are crucial for maintaining revenue streams, reducing marketing costs associated with acquiring new customers and fostering a positive brand reputation. Customer churn, also known as customer attrition, is the rate at which customers stop using a company's product or service [Masarifoglu and Buyuklu, 2019]. It can be caused by various factors such as dissatisfaction with the service, better offers from competitors, or changes in personal circumstances. In the telecommunications industry, market competitiveness is measured by churn rate. Telephone, internet and mobile services are all part of telecommunications.

2.3   Empirical Studies

A paper by [Imani, 2020] tackles customer churn in telecoms, emphasizing that retaining customers is cheaper than acquiring new ones. It evaluates various machine learning methods for predicting churn by comparing classifiers, target detectors, feature extraction and feature selection techniques. The study finds that Random Forest and Feed-Forward Neural Networks with Genetic Algorithm showed high accuracy. Effective target detectors included a Spectral wrangle Mapper and an Adaptive Subspace Detector. Clustering-Based Feature Extraction and Advanced Binary Ant Colony Optimization excelled in improving classification and feature selection. The study recommends implementing Random Forest and Neural Networks for effective churn prediction.

Another paper by [Masarifoglu and Buyuklu, 2019] emphasizes the importance of customer retention in telecoms, noting that retaining customers is more cost-effective than acquiring new ones. It employs survival analysis to estimate customer survival and hazard functions, focusing on factors like campaigns, tariffs, tenure, age and auto-payment. The paper introduces survival analysis (SA) as a statistical method to analyze time-to-event data. It describes the Cox Proportional Hazard Model (CPHM), a widely used technique

for quantifying the risk of an event occurring during the observation period. The paper uses Kaplan-Meier estimates and the Cox model to analyze the impact of various predictors on customer churn. The results show how different factors, such as campaigns and tariffs, affect customer survival probabilities. The study finds that survival analysis is effective for predicting churn and recommends its use for targeting at-risk customers and improving retention strategies.

The CPH model, a well-established method, is known for its interpretability, allowing researchers to understand the relationship between covariates and survival time [Nurhaliza et al., 2022]. However, it assumes proportional hazards, which may not hold in all cases. In contrast, Random Survival Forest (RSF) is a non-parametric ensemble method that can handle complex, non-linear relationships and does not require the proportional hazards assumption. The study uses data from various domains, such as medicine and engineering, to assess the performance of both models. The findings suggest that while RSF generally offers better predictive accuracy, especially in cases with non-linear relationships, CPH remains valuable for its simplicity and interpretability. The paper concludes that the choice between CPH and RSF should be based on the specific requirements of the analysis, balancing the need for interpretability against predictive accuracy.

Chapter 3

Methodology

3.1 Introduction

This chapter structures the research methods and procedures followed by the analysis to predict student churn of Vodafone (Telecel) in KNUST. A comprehensive explanation of the research models, mathematical formulations and interpretations are presented here. The paper compares their performance using the Concordance Index thus shedding light on their predictive capabilities.

3.2 Research Design

The study employs quantitative research. This type of research aims to establish the cause-and-effect relationships between variables. Specifically, it focuses on quantifying various aspects of students' usage and satisfaction, rather than exploring underlying meanings or personal experiences in an open-ended manner.

3.3 Pilot Survey

Before the actual study was carried out, a pilot survey was held to grasp the scope of students' understanding of the topic. It was carried out using final year students in Actuarial Science. Out of 50 questionnaires sent out, a total of 46 answered the questions.

3.4 Data Collection

The primary data source for this research was a survey conducted among students at Kwame Nkrumah University of Science and Technology (KNUST) in the College of Sci-

ence. The survey targeted students in their final years (Level 400). The dataset includes a variety of questions, each capturing specific aspects relevant to the study with the aid of Google Forms.

3.5   Sample Size

The sample size of the research was determined through simple random sampling. Simple random sampling is a sampling technique in which each member of the population has an equal chance of being selected. This method ensures that the sample is representative of the population, as every individual has an equal probability of being included. This sampling technique is employed to determine the optimal sample size as it is a straightforward method that allows for unbiased estimation of the population parameters.

The sample size for simple random sampling can be determined using the formula below:

$$n_0 = \frac{z^2 * p * (1 - p)}{E^2} \tag{3.1}$$

$$n = \frac{n_0}{1 + \left(\frac{n_0 - 1}{N}\right)} \tag{3.2}$$

Where:

(N) is the population size.

(n_0) is the initial sample size for simple random sampling.

(n) is the adjusted sample size for the population.

(z) represents the critical value.

(p) is the estimated proportion of the population

(E) is the margin of error.

3.6 Data Pre-Processing

In the realm of data analysis, ensuring the quality and suitability of data is paramount for deriving meaningful insights and making informed decisions. The initial phase of the study involved a thorough examination of the dataset to identify and handle missing data appropriately to ensure that subsequent analyses were conducted on a complete and representative dataset. One of the critical pre-processing tasks involved the transformation of categorical variables into numeric format. This was achieved using label encoding, a technique that assigns unique integer labels to each category with the aid of Python. The transformation restructured the dataset to facilitate survival analysis. The data was then organized to facilitate essential components such as time duration, event indicators and relevant covariates to the variables. The lifelines package played a pivotal role in this section by providing essential survival analysis models in Python. The models are crucial for analyzing data where the time to student churn is important.

3.7 Concept of Survival Analysis

Survival analysis is a branch of statistics used to analyze time-to-event data. The primary interest lies in the time until the occurrence of the event of interest. The time variable is usually referred to as survival time since it gives the time that an individual has survived over some follow-up period. The event is also referred to as a failure because the event of interest is usually death, disease incidence, or some other negative experience.

3.8 Censoring

In survival analysis, not all subjects may experience the event of interest within the study period. Censoring occurs when the survival time of a subject is not fully observed. It happens when a subject leaves the study before an event occurs or when the study ends before the event has occurred for all subjects.

3.8.1 Right-Censored

Right-censoring occurs when a subject experiences some loss to follow-up, or the study concludes before the event of interest has taken place. In such cases, we only know that the lifetime exceeds a certain value, meaning the true survival time is equal to or greater than the observed survival time. Right-censoring is common in survival analysis, especially in studies where follow-up periods vary among subjects or the study period is fixed.

3.8.2 Left-Censored

Left censoring happens when the event of interest has already occurred before the subject enters the study. Thus, the exact survival time is only known to be less than a certain value, indicating that the true survival time is less than or equal to the observed survival time. Left-censoring can be more challenging to handle because it typically requires additional assumptions or data about the time before the study begins.

3.8.3 Interval-Censored

Interval censoring arises when the exact time of the event of interest is unknown, but it is known to have occurred within a specific time interval. This type of censoring incorporates aspects of both right-censoring and left-censoring. For example, if a subject is only observed periodically and the event occurs between two observation points, the exact timing is unknown but constrained within the interval. Interval censoring requires specialized methods to analyze because it involves handling data within a range rather than a specific time point.

3.8.4 Type I Censoring

Type I censoring occurs when the study ends at a predetermined time and subjects who have not experienced the event by that time are censored. This is a form of right-censoring. For example, in a clinical trial with a fixed duration, all patients who do not

experience the event by the end of the study are censored at the study's conclusion.

3.8.5 Type II Censoring

Type II censoring happens when the study is designed to end after a certain number of events have occurred. All remaining subjects who have not experienced the event by the time this number is reached are censored. This type of censoring is often used in reliability testing, where a test might be stopped once a predetermined number of failures have occurred.

3.8.6 Random Censoring

Random censoring occurs when the censoring times are random and independent of the event times. This type of censoring can occur due to various reasons, such as a subject withdrawing from the study for personal reasons or being lost to follow-up. The analysis methods for random censoring need to account for the randomness of the censoring times to avoid bias in the estimation of survival functions.
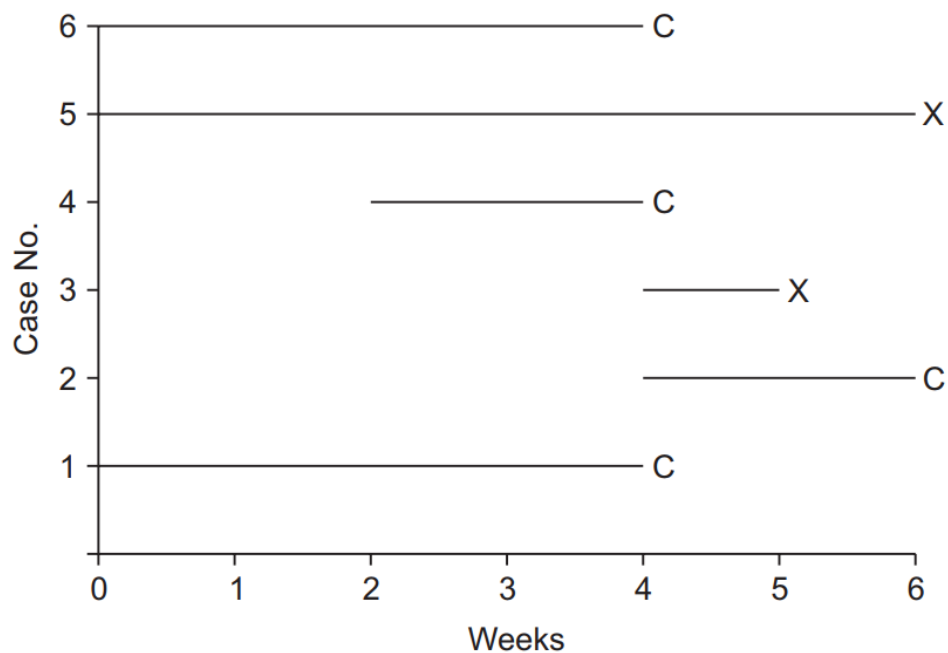
Figure 3.1: Censoring Graph

C indicates censored data

X indicates observed events

3.9 Fundamental Concepts of Survival Analysis

3.9.1 Survival Function $S(t)$

The survival function $S(t)$ also known as the survival probability function, gives the probability that a person survives longer than some specified time. Let T be a non-negative continuous random variable representing the elapsed time until an event occurs. The probability density function (pdf) of T is denoted by $f(t)$ and the cumulative distribution function (cdf) is denoted by $F(t)$. The cdf $F(t)$ is defined as:

$$F(t) = \Pr\{T < t\} \tag{3.3}$$

The survival function $S(t)$ is the complement of the cumulative distribution function and is given by:

$$S(t) = 1 - F(t) \tag{3.4}$$

$$S(t) = \Pr\{T \geq t\} \tag{3.5}$$

$$S(t) = \int_t^\infty f(u)\, du \tag{3.6}$$

The derivative of $S(t)$ with respect to t is:

$$S'(t) = -F'(t) \tag{3.7}$$

$$S'(t) = -f(t) \tag{3.8}$$

In Figure 3.2 below, the curve on the left is a theoretical curve of the survival function $S(t)$ which ranges from 0 up to infinity. It is non-increasing and therefore slopes downward as t increases. At $t = 0$, $S(0) = 1$ and when $t = \infty$, $S(\infty) = 0$. When actual data is used, the survivor curve does not result in a smooth curve but rather obtains a step function

graph. The step function is illustrated on the right in Figure 3.2. The study period is never infinite in duration as there may be competing risks for failure, not everyone may obtain the event of interest. The estimated survivor function $\hat{S}(t)$, thus may not go all the way down to 0 at the end of the study.
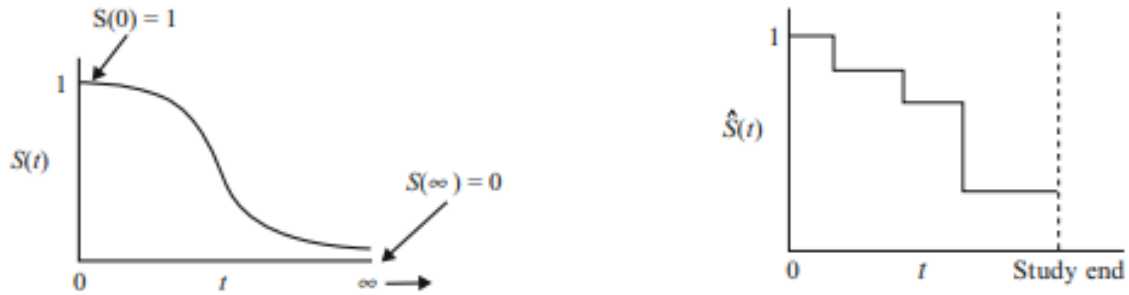
Figure 3.2: Survivor Curves

3.9.2 Hazard Function $h(t)$

The hazard function $h(t)$ denotes the instantaneous rate of failure at time t, given that the subject has survived up to time(t).

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{3.9}$$

$$h(t) \geq 0 \tag{3.10}$$

The hazard function $h(t)$, is given by the formula: $h(t)$ equals the limit, as Δt approaches zero, of a probability statement about survival, divided by Δt, where Δt denotes a small interval of time.

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T > t)}{\Delta t} \qquad (3.11)$$

$$= \lim_{\Delta t \to 0} \left[\frac{\Pr(t \leq T < t + \Delta t)}{\Delta t} \cdot \frac{\Delta t}{\Pr(T > t)} \right] \qquad (3.12)$$

$$= \lim_{\Delta t \to 0} \left[\frac{\Pr(t \leq T < t + \Delta t)}{\Delta t} \right] \cdot \lim_{\Delta t \to 0} \left[\frac{1}{\Pr(T > t)} \right] \qquad (3.13)$$

$$= \frac{1}{S(t)} \cdot \lim_{\Delta t \to 0} \left[\frac{\Pr(t \leq T < t + \Delta t)}{\Delta t} \right] \qquad (3.14)$$

$$= \frac{1}{S(t)} \cdot \lim_{\Delta t \to 0} \left[\frac{F(t + \Delta t) - F(t)}{\Delta t} \right] \qquad (3.15)$$

$$= \frac{f(t)}{S(t)} \qquad (3.16)$$

The hazard function is also known as the conditional failure rate. It is a rate rather than a probability. In the hazard function formula, the expression to the right of the limit sign gives the ratio of two quantities. The numerator is a conditional probability while the denominator, Δt denotes a small-time interval. By the division, a probability per unit of time is obtained, which is no longer a probability but a rate. In particular, the scale for this ratio is not 0 to 1 like a probability but rather ranges between 0 and infinity depending on whether time is measured in days, weeks, months, or years.

$$H(t) = \int_0^t h(x) dx \qquad (3.17)$$

The cumulative hazard function $H(t)$ can be derived from the hazard function. It is the integral of the hazard function up to time t. It represents the total hazard experienced up to time t. $H(t)$ and provides a straightforward cumulative measure of risk or failure over time.

$$S(t) = \exp \left[-\int_0^t h(x) \, dx \right] \qquad (3.18)$$

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right] \tag{3.19}$$

The equations above clearly define the relation between the hazard function and survival function. This can be seen in equation (3.18) and equation (3.19) in which the $h(t)$ gives the $S(t)$ and vice versa.

3.10 Approaches in Survival Analysis

The approaches to use in survival analysis are underlined by their strengths and weaknesses. The choice is based on statistical assumptions, data characteristics and the complexity of the survival patterns to model.

3.10.1 Parametric Methods

Parametric survival analysis is based on the assumption that survival times conform to a particular statistical distribution. The exponential, Weibull, log-normal and gamma distributions are common models that can be applied to various kinds of survival data [Aalen, 2004]. Maximum likelihood estimation, which finds the values that best fit the observed data, is commonly used for parameter estimation. As an alternative, Bayesian techniques can be applied, which refine estimates based on data by combining previous knowledge.

3.10.2 Non-Parametric Methods

In survival analysis, non-parametric techniques are useful since they assume very little about the survival distribution. A popular tool for estimating survival probabilities and visualizing survival curves is the Kaplan-Meier Estimator. In contrast, the Nelson-Aalen Estimator calculates the cumulative hazard function and offers insights into the event's risk over time. The survival distributions of the groups are compared using the Log-Rank

Test to see if there are any notable variations. When examining survival data and the precise distribution is unclear, these techniques can be helpful.

3.10.3   Semi-Parametric Methods

Semi-parametric models are used to assess survival data by combining parametric and non-parametric components. To accommodate different data patterns, they employ a non-parametric method to estimate the baseline hazard function and a parametric technique to represent the effect of variables. The Cox Proportional Hazards model is one illustration.

3.11   Kaplan Meier

The Kaplan-Meier estimator is employed in survival analysis to analyze the time until an event occurs. The Kaplan-Meier estimator calculates the survival probability at a specific time step by multiplying the probability of surviving each previous time step. The estimator is computed as

$$S\left(t\right) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.20}$$

Where:

t is a time,

d_i the number of events (churn) at time, t_i

$S\left(t\right)$ is the survival probability at time t and

n_i is the number of individuals at risk just before time. t_i

The estimator essentially calculates the probability of surviving from one time step to the next and the product of these probabilities gives the overall survival probability up to time t.

3.12   Cox Proportional (Cox PH) Hazard Model

The Cox Proportional Hazards model is a popular semi-parametric model to check the relationship between the survival time and a set of predictors.

$$h\left(t \mid x\right) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \tag{3.21}$$

Where:

$h(t \mid x)$ is the hazard function, i.e., the instantaneous rate of the event occurring at time t given the predictor variables x.

$h_0\left(t\right)$ is the baseline hazard function, representing the hazard for individuals with all predictor variables equal to zero,

$\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients for the predictor variables.

The coefficients β are estimated using maximum likelihood estimation and the model assumes a proportional hazard ratio, meaning the effect of the predictors on the hazard is constant over time. An important assumption on the Cox PH is that it has a constant hazard function proportion for each time. The Hypothesis of the assumption is as follows:

H_0 : The Assumption of Proportional Hazard is fulfilled

H_1 : The Assumption of Proportional Hazard is not fulfilled

with the rejection test, if the $p-value < 0.05$, then the hazard ratios are constant over time.

3.13   Accelerated Failure Time (AFT)

When the Cox proportional hazards assumption is not satisfied, the parametric model approach can be used. Accelerated Failure Time (AFT) is one of the popular parametric models used in survival analysis. The model assumes that the survival function $S(t)$ follows a parametric continuous distribution. This implies that the distribution follows a Weibull, lognormal, or log-logistic distribution. An AFT model aims to account for the influence of multiple covariates on the survival time by either accelerating or decelerating it.

$$\lambda(x) = \exp(b_0 + \sum_{i=1}^{n} b_i x_i) \tag{3.22}$$

Where:

$\lambda(x)$ is the accelerating factor

b_0 is the baseline accelerating factor when all covariates are 0

b_i is the regression coefficient for the i-th covariate

x_i are the covariates

$$H_0 : \text{The covariates do not impact the survival time}$$

$$H_1 : \text{The covariates have an impact on the survival time}$$

with the rejection test, if the $p - value < 0.05$, the covariates significantly affect the survival time in the AFT model.

3.13.1 Weibull AFT

Depending on its shape parameter, the Weibull distribution can model increasing, decreasing, or stable hazard rates, making it a very versatile tool in survival analysis. Because of its adaptability, it can be used for a variety of survival data, including medical outcomes and mechanical failures.

$$S(t) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\gamma}\right) \tag{3.23}$$

where λ is the scale parameter and γ is the shape parameter.

3.13.2 Lognormal AFT

In the case of right-skewed data, the model assumes that the survival time has a normal distribution. It converts the survival time into a log-normal distribution, which fits skewed survival data better and can effectively depict data with a long right tail.

$$S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right) \tag{3.24}$$

where $\mu = b_0 + \sum_{i=1}^{n} b_i x_i$ is the location parameter, σ is the scale parameter and Φ is the cumulative distribution function of the standard normal distribution.

3.13.3 Log-logistic AFT

The model can handle both growing and decreasing hazard rates and it assumes that the survival time follows a logistic distribution. Like the Weibull model, it is adaptable and diverse, which makes it helpful for simulating various risk patterns across time.

$$S(t) = \left(1 + \left(\frac{t}{\lambda}\right)^{\gamma}\right)^{-1} \tag{3.25}$$

where λ is the scale parameter and γ is the shape parameter.

3.14 Information Criteria and Loglikehood

Information criteria are used to evaluate and compare the relative quality of statistical models. They help in model selection by balancing model fit with complexity and penalizing models that use more parameters to avoid overfitting.

3.14.1 Log-Likelihood

The log-likelihood function is used to measure how well a model explains the observed data. It is the logarithm of the likelihood function, which represents the probability of observing the given data under a specific model.

$$\ln(L) = \sum_{i=1}^{n} \ln(f(y_i \mid \theta)), \tag{3.26}$$

Where:

n is the number of observations.

y_i represents each data point.

$f(y_i \mid \theta)$ is the probability density function of the data given the parameters θ.

A higher log-likelihood value indicates that the model provides a better fit to the data. When comparing different models, the one with the highest log-likelihood is typically preferred, as it suggests that the model is more likely to have generated the observed data. However, it's important to balance this with model complexity, which is where information criteria like Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion.

3.14.2   Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a tool used to evaluate and compare statistical models. It balances the model's fit to the data with its complexity by penalizing models with more parameters. A lower AIC value indicates a better model, as it suggests a good fit with fewer parameters.

$$AIC = 2k - 2\ln(L), \tag{3.27}$$

Where:

k is the number of parameters in the model.

L is the likelihood of the model.

3.14.3   Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is used for model selection, similar to AIC but with a stronger penalty for the number of parameters. This makes BIC more conservative, favoring simpler models, especially with larger sample sizes.

$$BIC = k\ln(n) - 2\ln(L), \tag{3.28}$$

Where:

k is the number of parameters in the model.

n is the number of observations in the dataset.

L is the likelihood of the model.

3.14.4 Hannan-Quinn Criterion (HQ)

The Hannan-Quinn Criterion (HQ) is a model selection tool that lies between AIC and BIC in terms of penalizing model complexity. It provides an intermediate option, with a penalty for the number of parameters that is less severe than BIC but more stringent than AIC.

$$HQ = 2k \ln(\ln(n)) - 2\ln(L),$$ (3.29)

Where:

k is the number of parameters in the model.

n is the number of observations in the dataset.

L is the likelihood of the model.

3.15   Concordance Index in Survival Analysis

A vital statistical metric used in survival analysis to evaluate the discriminative power of predictive models is the Concordance Index, also referred to as the C-index or Harrell's C-index [Harrell et al., 1982]. It is a gauge of a model's ability to distinguish across subjects with varying event durations, concentrating on the model's precision in assigning a person a risk rating. In order to ascertain if people with higher expected risks actually experience events sooner than those with lower risks, the C-index assesses the concordance between the predicted risk scores and the actual event outcomes.

$$C = \frac{Number\ of\ Concordant\ Pairs}{Number\ of\ Concordant\ Pairs\ + Number\ of\ Discordant\ Pairs}$$ (3.30)

This metric is particularly valuable in survival analysis because it provides a single,

interpretable summary of the model's overall discriminatory power. A higher C-index indicates that the model is more accurate in predicting which subjects are at greater risk, making it an essential tool for evaluating and comparing the performance of different survival models. The C-index has become a standard measure in survival analysis and has been widely adopted in both clinical and research settings for validating the effectiveness of prognostic models.

Chapter 4

Results and Analysis

4.1 Introduction

This section presents the study's findings and discusses what each output means toward the research goals. It includes tables, graphs and computer results based on the methods used. The analysis closely examines details of the model building, diagnostics and evaluation. The information here was obtained from using Python for computer analysis during the research. This chapter seeks to explain the use of survival models in the methodology to determine the Vodafone (Telecel) churn rate among students.

4.2 Sample Size Calculation

The sample size was calculated using the following parameters:

Z-score (z): A confidence level of 95% is chosen, resulting in a Z-score of 1.96.

Population (N): The population size of the College of Science students is about 2,835.

Estimated Proportion (p): To maximize sample size, we use 0.5.

Margin of Error (E): 5% since the confidence level is 95%.

$$n_0 = \frac{(1.96^2 \times 0.5 \times 0.5)}{0.05^2} \tag{4.1}$$

$$n_0 \approx 383.82 \tag{4.2}$$

$$n_0 \approx 384 \tag{4.3}$$

$$n = \frac{384}{1 + \left(\frac{384-1}{2835}\right)} \tag{4.4}$$

$$n \approx 338.297 \tag{4.5}$$

$$n \approx 338 \tag{4.6}$$

From a population of 2,835 students in College of Science, a sample size of 338 students were selected.

4.3 Data Description

The dataset has a shape consisting of 338 rows and 14 columns from a sample of students in KNUST's College of Science. The Churn variable is assigned as the event while the Churn Level is assigned as the time to the event (duration). The remaining covariates are used to explain the event and duration.

Variable	Description	Data Type
Gender	The gender of the student	Categorical
Churn	Whether the student has churned	Categorical
Residence	Place of residence in school	Categorical
Usage Frequency	Frequency of Vodafone Sim usage	Categorical
Voice Calls	Usage of voice calls	Categorical
Mobile Data Internet	Usage of mobile data/internet	Categorical
SMS Text Messaging	Usage of SMS/text messaging	Categorical
Data Exhaustion	Whether monthly data allowance (5GB) is exhausted	Categorical
Multiple Networks	Use of other networks	Categorical
Poor Network Quality Coverage	Perception of poor network quality/coverage	Categorical
Unsatisfactory Customer Service	Perception of unsatisfactory customer service	Categorical
High Costs Pricing	Perception of high costs/pricing	Categorical
Monthly Data Usage	Monthly Data Usage	Categorical
Churn Level	Level at which the student churns	Numeric

Table 4.1: Description of Variables

Table 4.1 contains behavioral data like usage frequency, voice calls and mobile data internet usage, in addition to demographic data like gender and place of residence.

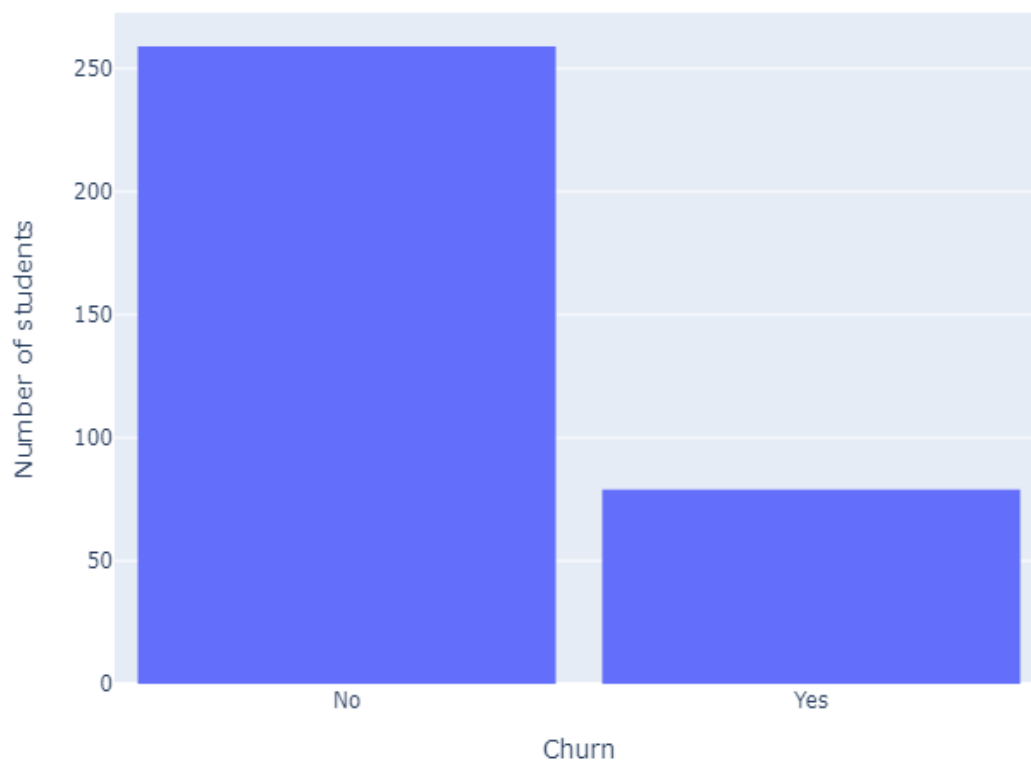## 4.4   Descriptive Analysis

Figure 4.1: Descriptive Analysis of Churn Data

In Figure 4.1 above, out of the 338 students, 79 were recorded to have churned. The survey also recorded 259 students who had not observed the event of interest (churn).

Variable	Response	Number of students
Gender	Female	233
	Male	105
Churn	No	259
	Yes	79
Residence	Off-campus	279
	On-campus	59
Usage Frequency	Daily	170
	Several times a week	67
	Occasionally	60
	Never	23
	Rarely	18
Voice Calls	Yes	285
	No	53
Mobile Data Internet	Yes	302
	No	36
SMS Text Messaging	Yes	180
	No	158
Data Exhaustion	Yes	283
	No	55
Multiple Networks	Yes	312
	No	26
Poor Network Quality Coverage	No	276
	Yes	62
Unsatisfactory Customer Service	Yes	288
	No	50
High Costs Pricing	Yes	291
	No	47
Monthly Data Usage	8 and more	231
	6-8	41
	2-4	33
	0-2	22
	4-6	11

Table 4.2: Numbers of Responses for each Variable

The response for the number of students in each variable is shown in Table 4.2. The number of students who choose each response category is indicated next to it, giving a clear picture of the distribution of data across several variables.

4.4.1 Kaplan Meier Analysis

The provided Kaplan-Meier estimator output in Table 4.2 below summarizes the survival curve in Table 4.1 over the levels in KNUST.

Event Time	Number of Students	Churned Students	Survival Probability	Lower 95% Confidence Interval	Upper 95% Confidence Interval
0	338	0	1.000000	1.000000	1.000000
1	338	25	0.926036	0.892497	0.949406
2	313	22	0.860947	0.819283	0.893630
3	291	32	0.766272	0.717405	0.807836

Table 4.3: Kaplan-Meier Survival Analysis Results

Initially, at time 0 in Table 4.3 (when students initially start the academic year), all 338 students are considered to be at risk. With no events (churns) recorded yet, the survival probability is 1.

As the students ascend the academic ladder, the number at risk begins to gradually decrease as some begin to experience the event. The higher the number of events, the more the number at risk decreases.

At the end of the first year, 25 students experienced the event of interest and therefore were omitted from the study at the beginning of the second year. Anytime the event of interest is experienced, they are omitted from the study until the study duration ends.

The confidence intervals (Lower and Upper 95% cond) provide ranges within which the true survival probabilities lie with a certain level of confidence.

4.4.2 Kaplan Meier (KM) Curve

The Kaplan-Meier survival curve below is a tool to estimate the probability that students will remain enrolled in the study over a given period.
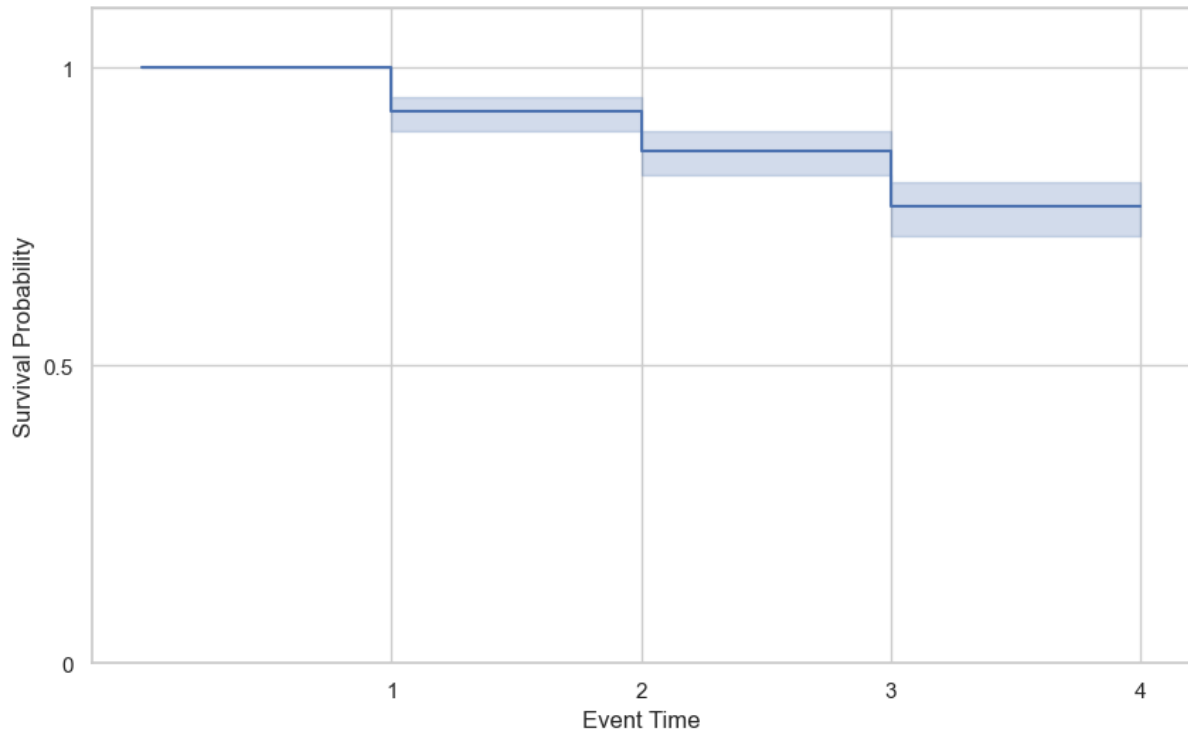
Figure 4.2: Kaplan Meier Curve

The x-axis represents the event time (Churn Level). The 0 indicates when the study began while the 4 represents when the study ended. The y-axis represents survival probability and ranges from 0 to 1. Each step down indicates an event, which decreases the overall survival probability. The shaded area around the line suggests the confidence interval, giving a range within which the true survival curve is expected to lie.

In churn prediction, this curve helps identify critical time points where student retention drops significantly and allows institutions to intervene proactively.

4.5 Cox Proportional Hazard (COX PH)

In the Cox Proportional Hazard modeling, two things are mostly done, namely, partial testing for each predictor variable and testing the Cox Proportional Hazard assumption. Furthermore, the results of parameter estimation and partial testing are presented in the Table below.

Variable	Coefficient (β)	Standard Error	P-value
Gender	0.33	0.28	0.24
Residence	0.32	0.27	0.25
Usage Frequency	-0.08	0.07	0.27
Voice Calls	-0.57	0.32	0.08
Mobile Data Internet	0.32	0.46	0.48
SMS Text Messaging	0.48	0.25	0.06
Data Exhaustion	-0.20	0.38	0.61
Multiple Networks	-0.13	0.44	0.78
Poor Network Quality Coverage	2.60	0.35	<0.005*
Unsatisfactory Customer Service	-1.40	0.30	<0.005*
High Costs Pricing	-1.13	0.29	<0.005*
Monthly Data Usage	0.21	0.09	0.02*

Table 4.4: Cox Proportional Hazards Model Results

* indicates significant covariate.

The coefficient columns in Table 4.4 provide information about the relationship between each independent variable and the dependent variable. A positive coefficient indicates an increase in the independent variable is associated with an increase in the hazard (risk) of churn (event). A higher value of these variables is associated with a higher likelihood of churn occurring.

Conversely, a negative coefficient suggests a decrease in the hazard (risk) of churn (event). A higher value of these variables is associated with a lower likelihood of churn occurring. If the model were to be fitted on a new random sample from the same population, the standard error would show how much the predicted coefficient is likely to change. A smaller standard error denotes greater precision in the coefficient estimate, whereas a bigger standard error denotes less precision.

The p-value column helps assess the statistical significance of each independent variable. A low p-value (< 0.05) indicates that the covariate is likely to have a meaningful impact on the event. This can be seen in Monthly Data Usage, Poor Network Quality Coverage and High Costs Pricing.
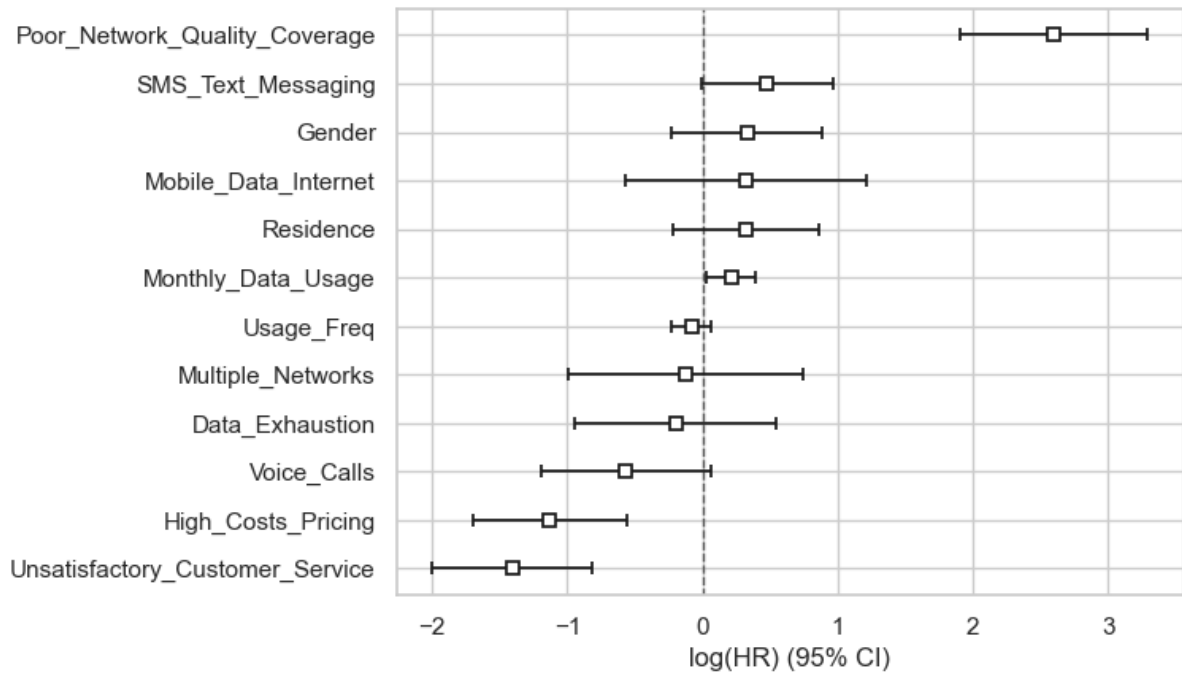
Figure 4.3: Cox PH Forest Plot

Figure 4.3 above is a forest plot of the Cox PH. A coefficient to the right of zero (positive log hazard ratio) indicates that an increase in that variable is associated with a higher risk of a student churning. A coefficient to the left of zero (negative log hazard ratio) indicates that an increase in that covariate is associated with a lower risk of a student churning. The further a coefficient is from zero, the stronger the effect of that covariate on the hazard of churn. The horizontal lines extending from the boxes are the 95% confidence intervals, showing the uncertainty around these estimates. If the confidence interval crosses zero, the covariate is not statistically significant.

4.5.1 Cox PH Assumption Test

This test checks if the impact of the predictor variables on the hazard rate is constant over time. Covariates violating this assumption might need further investigation or transformation.

The null hypothesis states that there is a significant relationship between the predictor variables (such as College and Voice Calls) and the likelihood of a student churning.

The alternative hypothesis suggests that there is no significant association between the predictor variables and the likelihood of churn.

Covariate	Test Statistic	P-value
Data Exhaustion	0.46	0.50
Gender	0.26	0.61
High Costs Pricing	1.74	0.19
Mobile Data Internet	0.63	0.43
Monthly Data Usage	1.49	0.22
Multiple Networks	0.09	0.77
Poor Network Quality Coverage	2.36	0.12
Residence	0.01	0.93
SMS Text Messaging	0.04	0.84
Unsatisfactory Customer Service	0.10	0.76
Usage Frequency	0.04	0.84
Voice Calls	2.70	0.10

Table 4.5: Test statistics of Cox-PH Assumption Test

The Cox Proportional Hazard method is an assumption that must be met. In Table 4.5, it can be seen that the covariates meet the assumptions as their p-values are above 0.05 meaning that there is no strong evidence against the proportional hazard's assumption for these covariates.

This means that the assumption that the hazard ratios are constant over time holds for all the covariates.

4.6   Accelerated Failure Time (AFT)

The Accelerated Failure Time (AFT) model is a possible alternative that can be used when the Cox Proportional Hazard (PH) model does not hold. The study uses the AFT to understand the direct effect of covariates on survival time.

Model	AIC	BIC	Hanna-Quinn
Weibull	316.433303	300.079395	306.67379
LogNormal	**305.805320**	**289.451411**	**306.67379**
LogLogistic	307.510343	291.156435	306.67379

Table 4.6: Comparison of AIC, BIC and Hanna-Quinn values for different AFT models

Results:

The AFT model with the lowest AIC is: *LogNormal*

The AFT model with the lowest BIC is: *LogNormal*

The AFT model with the lowest Hanna-Quinn is: *Weibull*

The study compares three Accelerated Failure Time (AFT) models—Weibull, Log-Normal and Log-Logistic—using AIC, BIC and Hanna-Quinn criteria in Table 4.6. The results reveal that the Log-Normal consistently has the lowest AIC and BIC, indicating a better overall fit despite the Weibull achieving the lowest Hanna-Quinn value.

4.6.1   Log-Normal Model

Variable	Coefficient (β)	Standard Error (SE)	P-value
Data Exhaustion	0.102	0.148	0.493
Gender	0.032	0.118	0.785
High Costs Pricing	0.515	0.122	<0.0005*
Mobile Data Internet	-0.287	0.207	0.165
Monthly Data Usage	-0.069	0.044	0.113
Multiple Networks	-0.104	0.202	0.607
Poor Network Quality Coverage	-1.041	0.123	<0.0005*
Residence	-0.146	0.124	0.239
SMS Text Messaging	-0.228	0.108	0.036*
Unsatisfactory Customer Service	0.809	0.122	<0.0005*
Usage Frequency	-0.002	0.033	0.940
Voice Calls	0.132	0.135	0.329
Intercept (for μ)	1.626	0.393	<0.0005
Intercept (for σ)	-0.616	0.084	<0.0005

Table 4.7: Lognormal Model Coefficients

* indicates significant covariate.

Unlike the Cox PH analysis, the coefficient columns in Table 4.7 provide information about the relationship between the covariates to event time (Churn_Level). A positive coefficient indicates a decrease in the time it takes to the event. A negative coefficient suggests an increase in the time to the event.

A higher value of these covariates is associated with a higher likelihood of the time to the event occurring if positive and a lower chance of occurring if negative. The standard error

quantifies the variation in the calculated coefficients among several population samples. A lower standard error (SE) suggests a more accurate approximation, implying that the coefficient is probably nearer the actual value in the population.

The p-value column helps assess the statistical significance of each covariate. A low p-value (< 0.05) indicates that the covariate is likely to have a meaningful impact on the Churn Level.

The mu_intercept is the mean baseline when the covariates are 0 and the sigma_intercept is the standard deviation used to estimate the variance to determine the spread when the covariates are 0.
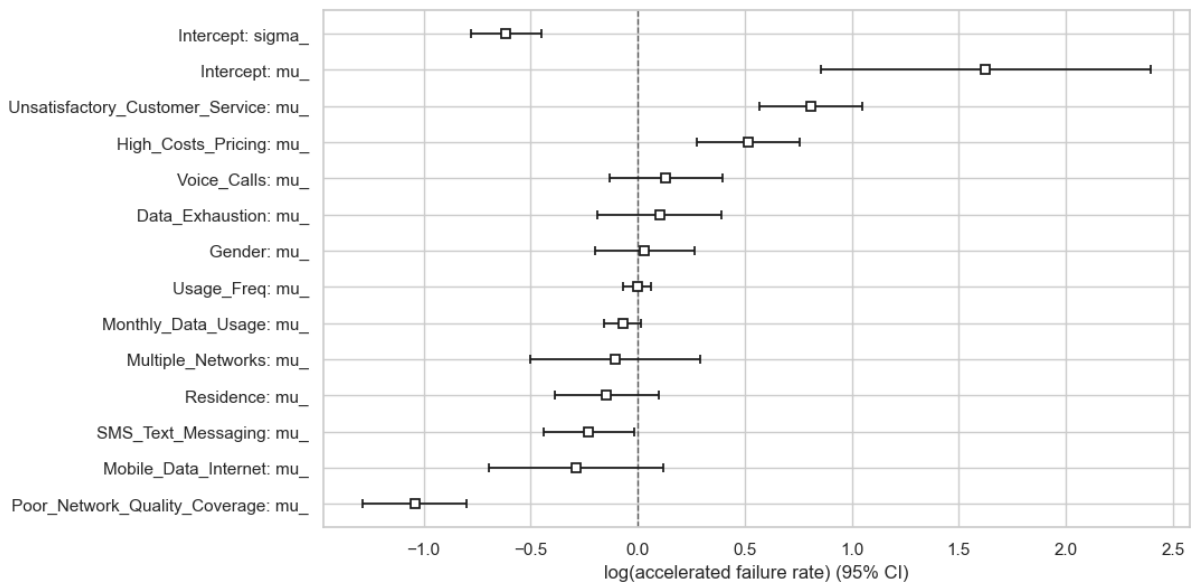

Figure 4.4: LogNormal Forest Plot

The boxes (squares) in the plot represent the estimated coefficients for each factor in the Lognormal AFT model in Figure 4.4. The positive values suggest that the covariate decreases the time to the event (churn), while negative values suggest it increases the time to the event (Churn Level). The horizontal lines extending from the boxes are the 95% confidence intervals, showing the uncertainty around these estimates. If the confidence interval crosses zero, the covariate is not statistically significant.

4.7   Model Comparison

In this section, the models used are compared based on the Loglikelihood and the concordance values. The higher the concordance, the better the model predictive value and the higher the loglikelihood, the better the model fit.

Model	Concordance	Loglike hood
LogNormal	0.958	290.678
Cox PH	0.96	266.68

Table 4.8: Model Concordance and AIC values

Based on the comparison of the concordance values and log-likelihood in Table 4.8, it is evident that the Cox PH model shows a higher concordance compared to the LogNormal model, indicating a better predictive value. However, the log-likelihood values suggest that the LogNormal model has a better fit.

Chapter 5

Summary of findings, conclusion and recommendations

5.1 Introduction

This section provides a comprehensive overview of the findings, conclusions and recommendations derived from the study. The primary objective of the research was to investigate student churn of Telecel in KNUST, aiming to provide actionable insights and guidance based on the data analyzed.

5.2 Summary of finding

Poor network quality, poor customer service, high costs pricing, monthly data usage and SMS text messaging were found to be important indicators of student churn. While the Lognormal Accelerated Failure Time (AFT) model provided a higher and better log-likelihood value, the Cox Proportional Hazards (PH) model showed a higher and better concordance value. Furthermore, students churned more at the end of their 3rd year.

5.3 Conclusion

The analysis identified that out of the significant covariates—Poor network quality, poor customer service, high costs pricing, monthly data usage and SMS text messaging —Poor Network Quality Coverage was the most significant covariate to increase both the churn rate and time to churn. The Cox Proportional Hazards model showed a higher predictive value, while the Lognormal Accelerated Failure Time (AFT) model provided a better fit

for the data. Additionally, students in their 3rd year had the highest churn rate, with 35 out of 79 churners coming from this group.

5.4 Recommendations

This section presents recommendations based on the study's findings. From the analysis of the level 400 students in the College of Science over the 4 years, recommendations to the University Information Technology Services (UITS) are as follows:

1. Improving network coverage at KNUST can reduce churn rates by ensuring consistent and reliable connectivity. This enhances students' access to online resources, supports smooth communication and minimizes technical issues, leading to a better overall academic experience and higher student satisfaction. Investing in better infrastructure and regular maintenance will help achieve these improvements.

2. Establishing diverse communication channels and feedback systems through digital platforms ensures that students can get timely and efficient help. This approach addresses concerns effectively and improves overall student satisfaction.

3. Future research could expand on this study by including other Colleges in KNUST to obtain a more diverse student population. Using data from the UITS will capture student sentiments more deeply in understanding the churn of Telecel.

References

[1] Aalen, O. O. (2004). 1. The statistical analysis of failure time data (2nd edn). JD Kalbfleisch and RL Prentice, Wiley-Interscience, Hoboken, New Jersey, 2002. No. of pages: 439. Price:£ 62.95. ISBN: 0-471-36357-X.

[2] AP News. (2024). Underwater cable disruptions affect West Africa's internet services. *Retrieved from:*
`https://apnews.com/article/africa-internet-outage-undersea-cables-failure-ac67fd`
Accessed: 2024-06-05.

[3] Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., & Vanthienen, J. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society, 56*(9), 1089-1098.

[4] Bandim, A. N. A. (2022). *Managing stakeholder communication in the Ghanaian telecommunication industry* (Doctoral dissertation, University of Pretoria).

[5] Box-Steffensmeier, J. M., & Zorn, C. J. (2001). Duration models and proportional hazards in political science. *American Journal of Political Science*, 972-988.

[6] Cox, D. R., & Lewis, P. A. W. (1972, January). Multivariate point processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 3, pp. 401-448).

[7] Evrensel, A. Y. (2008). Banking crisis and financial structure: A survival-time analysis. *International Review of Economics & Finance, 17*(4), 589-602.

[8] Ghana Web. (2024). Underwater fiber optic cable outages impact telecommunications in West Africa. *Retrieved from:* `https://www.ghanaweb.com/GhanaHomePage/`
`business/Details-of-the-outages-on-multiple-submarine-optic-fibre-cables-that-hi`
Accessed: 2024-06-05.

[9] Graunt, J. (1977). Natural and political observations mentioned in a following index and made upon the bills of mortality. *Mathematical demography*, *6*, 11-20.

[10] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543-2546.

[11] Imani, M. (2020). Customer Churn Prediction in Telecommunication Using Machine Learning: A Comparison Study. *AUT Journal of Modeling and Simulation*, *52*(2), 229-250.

[12] Iwasaki, I. (2014). Global financial crisis, corporate governance and firm survival:: The Russian experience. *Journal of comparative economics*, *42*(1), 178-211.

[13] Jerenz, A. (2008). *Revenue management and survival analysis in the automobile industry*. Betriebswirtschaftlicher Verlag Gabler.

[14] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457-481.

[15] Gepp, A., & Kumar, K. (2015). Predicting financial distress: A comparison of survival analysis and decision tree techniques. *Procedia Computer Science*, *54*, 396-404.

[16] Lee, M. C. (2014). Business bankruptcy prediction based on survival analysis approach. *International Journal of Computer Science & Information Technology*, *6*(2), 103.

[17] Leung, M. K., Rigby, D., & Young, T. (2003). Entry of foreign banks in the People's Republic of China: a survival analysis. *Applied Economics*, *35*(1), 21-31.

[18] Masarifoglu, M., & Buyuklu, A. H. (2019). Applying survival analysis to telecom churn data. *American Journal of Theoretical and Applied Statistics*, *8*(6), 261-275.

[19] Nurhaliza, S., Sadik, K., & Saefuddin, A. (2022). A COMPARISON OF COX PROPORTIONAL HAZARD AND RANDOM SURVIVAL FOREST MODELS IN

PREDICTING CHURN OF THE TELECOMMUNICATION INDUSTRY CUS-TOMER. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *16*(4), 1433-1440.

[**20**] Pereira, J. (2014). Survival analysis employed in predicting corporate failure: A forecasting model proposal. *International Business Research*, *7*(5), 9.

[**21**] Smith, R. A., & Brown, M. G. (2020). Far beyond postsecondary: Longitudinal analyses of topical and citation networks in the field of higher education studies. *The Review of Higher Education*, *44*(2), 237-264.

[**22**] Tsujitani, M., & Baesens, B. (2012). Survival analysis for personal loan data using generalized additive models. *Behaviormetrika*, *39*(1), 9-23.

[**23**] Wikipedia. (2020). Vodafone Ghana. *Retrieved from:* `https://en.wikipedia.org/wiki/Vodafone_Ghana#:~:text=As%20of%20January%202020%2C%20it,of%20the%20Ghanaian%20market%20shares`. Accessed: 2024-05-24.

[**24**] Wikipedia. (2020). Vodafone KNUST. *Retrieved from:* `https://en.m.wikipedia.org/wiki/Vodafone_Ghana#:~:text=In%202016%2C%20Vodafone%20made%20a,the%20university's%20campuses%20in%20Ghana`. Accessed: 2024-07-02.

[**25**] Wikipedia. (2024). Telecel Ghana. *Retrieved from:* `https://en.m.wikipedia.org/wiki/Telecel_Ghana#:~:text=They%20are%20the%20leading%20total,Vodafone%20Ghana%20into%20Telecel%20Ghana`. Accessed: 2024-05-19.