
NYANSAPƆ
WƆSANE NO
BADWENMA

Decoding Student Retention and Churn of Vodafone (Telecel) in Kwame Nkrumah University of Science and Technology (KNUST) - A Survival Analysis Approach

Abstract

This study investigates student retention and churn of Vodafone (Telecel) services at Kwame Nkrumah University of Science and Technology (KNUST) using survival analysis approaches. The research aims to understand the factors influencing student churn, analyze patterns related to demographics and usage, and develop strategies to enhance retention rates. By focusing on KNUST students, the study seeks to improve Telecel's service offerings, enhance student experience, and strengthen the partnership between Telecel and KNUST.

Dedication

This work is dedicated to my family and friends for their unwavering support and encouragement throughout my studies. Special thanks to my advisors and colleagues who provided invaluable insights and guidance during this research.

Contents

List of Figures

List of Tables

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 [1]. 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 [4]. 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. 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, and 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. This collaboration 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. This rebranding aimed to improve service offerings across voice and data services, money transfers, and business solutions. Telecel, founded in 1986, is an Africa-focused telecommunication company that operates primarily in Africa and converges telecommunications with fintech, e-commerce, and 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, and 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 [2]. 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, West Africa Cable System (WACS), and the Africa Coast to Europe (ACE), resulting in MainOne and SAT3 going offline [5]. 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.

Previous studies have explored various factors that influence customer churn in a telecommunications company. 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

Regardless of efforts made by KNUST to partner with Telecel to provide affordable and accessible mobile communications services to the university's populace, churn still persists. There is a lack of understanding about the factors driving student churn and retention, making it difficult to develop effective strategies to address this issue [3]. This research aims to investigate the factors contributing to student churn and develop a survival analysis model for detecting at-risk students and design specific strategies to improve retention rates. It also seeks to enhance Telecel's services, improve student experience, and foster long-term relationships between Vodafone (Telecel) and KNUST.

1.3 Research Objectives

1.3.1 Main Objective

- To develop a survival analysis techniques model to detect at-risk students to improve student retention and reduce churn of Telecel at KNUST.

1.3.2 Specific Objectives

- Exploring the factors that influence student churn and retention.
- Using survival analysis models to identify students at risk of churning.
- Analyzing time-to-event data to understand the patterns and timing of student churn.
- Identifying strategies to improve retention rates and reduce churn.

1.4 Research Questions

- What factors influence student retention and churn for Telecel services at KNUST?

- What is the relationship between student demographics (age, gender, year of study) and churn behavior?
- How does the churn rate vary with different Telecel services (voice service, data service, internet service)?
- How does the quality of Telecel network and services (coverage, internet speed) influence student retention and churn?
- How can Telecel services be optimized to better meet the needs and preferences of KNUST students?

1.5 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, and less hassle from switching providers. Ultimately, it ensures a more stable and reliable service for students.

1.6 Structure 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.

1.7 Limitation of the Study

The limitation of this research was that the data survey only included students of KNUST who were present during the annual college elections. As a result, the coverage was limited to undergraduate students on the main campus.

Chapter 2

Literature Review

2.1 Introduction

2.2

2.3

2.4

2.5

2.6

Chapter 3

Methodology

3.1 Introduction

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

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 statistics were carried out, a pilot survey was held to grasp the scope of students' understanding to the topic. It was carried out using the residence of Otumfuo Osei Tutu II (popularly known as SRC hostel) on campus. Out of 150 questionnaires sent out, a total of 137 respondents were

returned. Most of the respondents were males and this could have been as a result of having females conduct the survey. This was done to improve the reliability of the data.

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). The survey targeted students across different academic levels (from level 100 to 600) during the annual college elections. The dataset includes a variety of questions, each capturing specific aspects relevant to our study with the aid of Google Forms.

The data collection process involved obtaining relevant information while ensuring confidentiality and ethical considerations.

3.5 Sample Size

The sample size of the research was determined through cluster sampling. Cluster sampling is a sampling technique in which the population is divided into groups known as clusters. A random sample is then selected from each cluster ensuring that, each cluster has an equal number of elements. It is therefore homogeneous within but heterogeneous around (clusters

are different from one another but the elements within it share common factors). This sampling is employed to determine the optimal sample size as it is a robust method that allows to efficiently estimate the population by selecting entire clusters (colleges within KNUST) rather than using individual students.

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

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

$$n = \frac{n_0}{1 + (\frac{n_0 - 1}{N})}$$

$$n_{cluster} = n \ * \ DEFF$$

Where:

- (N) is the population size.
- (n_0) is the sample size for simple random sampling
- (n) is the sample size for the population.
- $(n_{cluster})$ is the sample size of the clusters.
- (z)represents the critical value.
- (p) is the estimated proportion of the population (e.g., proportion of students with a certain behavior).
- $(DEFF)$ is the design effect, which accounts for the correlation among observations within the same cluster.
- (E) is the margin of error.

Parameter Justification:

- **Z-score** (z): A confidence level of 95% is chosen therefore resulting to a Z-score of 1.96.
- **Population**(N): The population size of KNUST students is about 85000.
- **Estimated Proportion** (p): To maximize sample size, we use 0.5.
- **Margin of Error** (E): 5% since the confidence Level is 95%.

$$n_0 = \frac{\left(1.96^2 * \ 0.5 \ * \ 0.5 \ \right)}{0.05^2}$$

$$n_0 \approx 383.82$$

$$n_0 \approx 384$$

$$n = \frac{384}{1 + (\frac{384-1}{85000})}$$

$$n \approx 384.16$$

$$n \approx 384$$

- **Design Effect** ($DEFF$): The design effect of 2 is used as the benchmark.

$$n_{cluster} = 384 \ * \ 2$$

$$n_{cluster} = 768$$

A cluster sample size of 768 students from the population of about 85000.

Since there are 6 clusters from which each represents the colleges, the sample size is evenly allocated across the clusters. Each cluster would have approximately 128 students.

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 are conducted on a complete and representative dataset. One of the critical preprocessing 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 structured 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.

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. This could be anything from the failure of a mechanical part to the occurrence of a disease all the way to the death of a patient. 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. There are some special cases where the failure is a positive event and not a negative one.

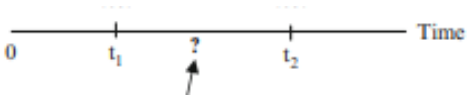

Figure 3.1: Interval Censoring

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 occurs 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

It occurs when a subject has some loss to follow up or the study ends before the event of interest occurs. The lifetime is known to exceed a certain value meaning that, true survival time is equal to or greater than the observed survival time.

3.8.2 Left-Censored

It occurs when the event of interest has occurred before the study starts, and thus the exact survival time is known only to be less than a certain value indicating that the true survival time is less than or equal to the observed survival time.

3.8.3 Interval-Censored

It can occur if a subject's true but unobserved survival time is within a certain known specified time interval. Figure 3.1: Interval Censoring

Interval-censoring incorporates both right-censoring and left-censoring. Left-censored data occurs whenever the value is 0 and is a known upper bound on the true survival time. In contrast to the left-censored, right-censored data occurs whenever the value of is infinity, and is a known lower bound on the true survival time.

- C indicates censored data

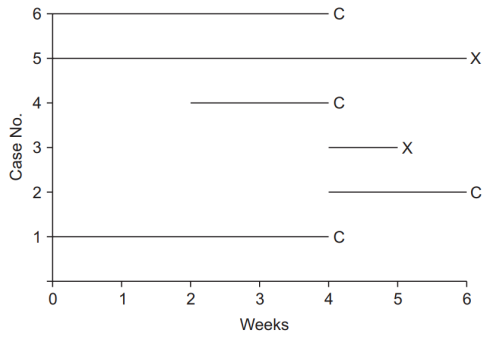

Figure 3.2: Censoring Graph

- X indicates observed events

Figure 3.2 illustrates the different types of censoring in a survival analysis study conducted over 6 weeks. Case 1 and Case 6 are right-censoring. The participant was followed from the start of the study until about 4 weeks, at which point they were censored (indicated by). This could mean the participant dropped out of the study or the study ended before an event occurred for this individual. Cases 2 and 4 are another instance of right censoring, but their participants joined the study later and still did not experience the event of interest. Case 3 and Case 5 show an observed event (indicated by) occurring at around 4 weeks and 6 weeks respectively. This is not censored data, as the event of interest was observed within the study period.

3.9 Fundamental Concepts of Survival Analysis

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

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

These functions are complementary in understanding the dynamics of survival data, namely survival function and hazard function. They are the

building blocks for every survival analysis. They clearly have a defined relationship between the two. The $S(t)$ can be derived from the $h(t)$ complement one another such that, one can be derived from the other.

The first formula describes how the survival function $S(t)$ can be written in terms of an integral involving the hazard function. The formula states that $S(t)$equals the exponential of the negative integral of the hazard function $h(t)$ between integration limits of 0 and t. The second formula describes how the hazard function $h(t)$ can be written in terms of a derivative involving the survival function. The formula states that $h(t)$ equals minus the derivative of $S(t)$ with respect to t divided by $S(t)$.

3.9.1 Survival Function $S(t)$

The survival function $S(t)$ is also known as the survival probability function gives the probability a person survives longer than some specified time t.

$$S(t) = P(T > t)$$

The survival function is very fundamental to a survival analysis as it helps in determining survival probabilities for different values of time to provide crucial summary information from the data.

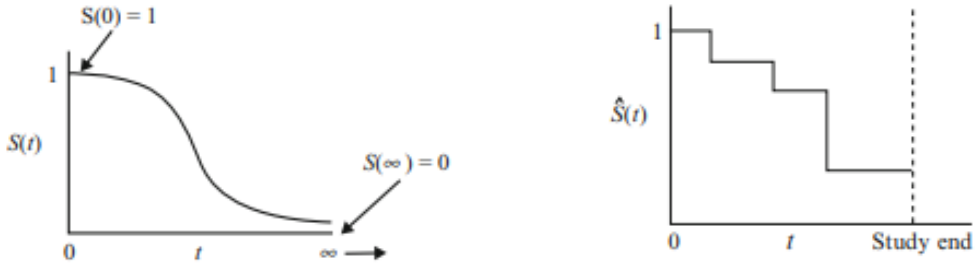

Figure 3.3: Survivor Curves

The figure on the left in Figure 3.3 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 time 0, $S(t) = 1$ and when $t = \infty, S(\infty) = 0$. At, the start of the study, no event occurs and it is assumed that if the study period is without a limit, the survivor curve will eventually reach 0. When actual data is used, the survivor curve does not

result to a smooth curve but rather a step function. The step function is illustrated on the right in Figure 3.3. Since the study period is never infinite in duration as there may be competing risks for failure, it is possible that not everyone obtains 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.

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).

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

$$h(t) \geq 0$$

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.

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 while depending on whether time is measured in days, weeks, months, or years.

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

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.$H(t)$ provides a straightforward cumulative measure of risk or failure over time.

3.10 Approaches in Survival Analysis

There are various approaches to determine the type of survival model to use. Each approach has its strengths and weaknesses, and the choice typically involves balancing statistical assumptions, data characteristics, and the complexity of the survival patterns to model.

3.10.1 Parametric Methods

These methods assume that the survival times follow a specific statistical distribution. Common parametric survival models are exponential, Weibull, log-normal and gamma models. They estimate parameters using maximum likelihood estimation or Bayesian methods.

3.10.2 Non-Parametric Methods

These methods include approaches that make minimal assumptions about the form of the survival distribution. Common non-parametric methods in survival analysis are the Kaplan-Meier Estimator, Nelson-Aalen Estimator and Log-Rank Test.

3.10.3 Semi-Parametric Methods

Semi-parametric models combine parametric elements (the effect of covariates) with non-parametric elements (the baseline hazard function). It is primarily represented by the Cox Proportional Hazard Model.

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. Let $S(t)$ be the survival probability at time. The estimator is computed as

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

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
- 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 for survival analysis by Sir David Cox (1924). It models the relationship between the survival time and a set of predictors.

$$h\left(t|x\right) = h_0(t)exp(_1x_1 +_2 x_2 + +_px_p)$$

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.

H_0: The Assumption of Proportional Hazard is fulfilled

H_1: The Assumption of Proportional Hazard is not fulfilled

3.13 Accelerated Failure Time (AFT)

In situations where the Cox proportional 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 is following a Weibull, lognormal or exponential distribution. The aim of an AFT is to account for the influence of multiple covariates on the survival time by either accelerating or decelerating it.

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

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
- x_i are the covariates

3.14 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a measure of the relative quality of statistical models for a given dataset. It aids in model selection in survival analysis. It contributes by penalizing models with more parameters to avoid the problem of overfitting. The lower the AIC value, the better the model is considered to fit the data.

$$AIC = 2k - 2ln(L),$$

Where:

- k is the number of parameters in the model.
- L is the likelihood of the model

3.15 Concordance Index in Survival Analysis

The Concordance Index, often referred to as the C-index or Harrell's C-index, is a statistical metric used to evaluate the performance of models in survival analysis. It assesses how well a model discriminates between subjects in terms of their event times and predicted risks.

The Concordance Index measures the model's ability to correctly rank the predicted risks of individuals based on their actual event times. The Concordance Index evaluates whether the model's predicted risks align with the observed event times.

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

A C value above 0.5 suggests that the model has predictive ability better than random chance. A higher C value implies a better model performance and more accurate risk predictions.

Chapter 4

Analysis and Findings

4.1 Introduction

This section presents the study's findings and discusses what each output means towards the research goals. It includes easy-to-read tables, graphs, and computer results based on the methods used. The analysis closely examines details towards 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 Data Description

The dataset has a shape consisting of 768 rows and 18 columns from a sample in KNUST.

4.3 Model Building

The lifelines package played a pivotal role in this section by providing essential survival analysis models in Python. These models are crucial for analyzing data where the time student churn event is important.

Field Name	Description
Gender	
College	
Churn	
Level	

Table 4.1: Caption

4.3.1 Kaplan Meier (KM) Curve

A Kaplan-Meier curve, also known as a survival curve, is a statistical tool used in survival analysis to estimate the survival function from timeline data. It provides a way to visualize the proportion of individuals surviving over time, taking into account censored data (individuals who have not experienced the event by the end of the observation period).

The Kaplan-Meier survival curve provided is a tool to estimate the probability that students will remain enrolled over a given time period. The x-axis represents the timeline, which in this ranges from 1 to 5 levels or years. The 0 indicates when the study began. The y-axis represents survival probability and ranges from 0 to 1.

The KM estimate line graph shows the survival probability at various points along the study. 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. For instance, if there's a notable step down at a particular time point on the x-axis (3-4), it might indicate a period when students are more likely to leave and thus could be a target for retention efforts.

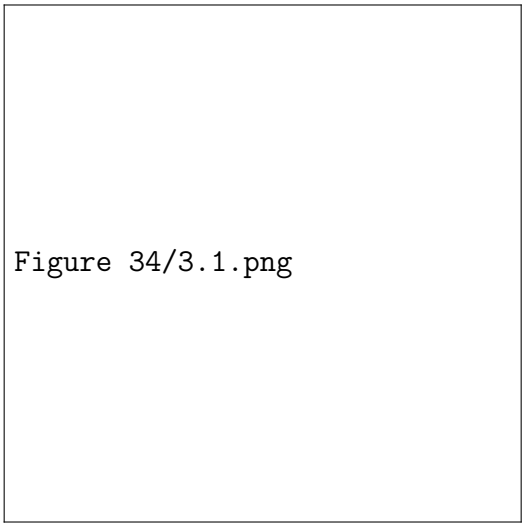

Figure 4.1: Kaplan Meier Curve of students from levels 100-500

4.3.2 Kaplan Meier Analysis

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

Initially, at time 0 (when students initially start the academic year), all 768 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 decrease. This can be seen for example, in the 2^{nd} level where the initial 768 students from the beginning of the 1^{st} year decreased to 733 for the 2^{nd} year after 35 students churned at the end of the year. Subsequently, the survival probability declines gradually from 1 to 0.500583 by the 5^{th} year. The confidence intervals (Lower Cl and Upper CI) provide ranges within which the true survival probabilities lie with a certain level of confidence.

4.3.3 Cox Proportional Hazard (COX PH)

The analysis was continued by doing the Cox Proportional Hazard modeling. In the Cox Proportional Hazard modeling, there are two things that are done, namely partial testing for each predictor variable and testing the Cox

Proportional Hazard assumption.

The Cox Proportional hazards regression model can be expressed as such:

$$\begin{aligned} h(t,x) = h_0(t)\exp(\ &-0.56 \cdot \text{Gender} - 0.03 \cdot \text{College} - 0.09 \cdot \text{Residence} \\ &- 0 \cdot \text{Usage\_Freq} + 0.23 \cdot \text{Network\_Strength} \\ &+ 0.16 \cdot \text{Voice\_Calls} \\ &+ 0.32 \cdot \text{Mobile\_Data\_Internet} - 0.1 \cdot \text{SMS\_Text\_Messaging} \\ &+ 0.41 \cdot \text{Data\_Exhaustion} - 0.21 \cdot \text{Multiple\_Networks} \\ &+ 0.13 \cdot \text{Other\_Networks\_Better\_Services} \\ &- 0.17 \cdot \text{Poor\_Network\_Quality\_Coverage} \\ &- 0.12 \cdot \text{Insufficient\_Data\_Allowance} \\ &- 0.15 \cdot \text{Unsatisfactory\_Customer\_Service} \\ &+ 0.16 \cdot \text{High\_Costs\_Pricing} - 0.06 \cdot \text{Monthly\_Data\_Usage}) \end{aligned}$$

Furthermore, the results of parameter estimation and partial testing are presented in the Table below.

The coefficient and exp(coefficient) columns provide information about the relationship between each independent variable and the dependent variable. A positive coefficient or exp(coefficient) >1 (such as Network Strength and Voice Calls) indicates that an increase in the independent variable is associated with an increase in the odds of the outcome. These increase the hazard (risk) of churn. A higher value of these variables is associated with a higher likelihood of churn.

Conversely, a negative coefficient or exp(coefficient) < 1 (such as Poor Network Quality Coverage and Unsatisfactory Customer Service) suggests a decrease in the odds of the outcome. These decrease the hazard (risk) of churn. A higher value of these variables is associated with a lower likelihood of churn. The p-value column helps assess the statistical significance of each independent variable. A low p-value (typically < 0.05) indicates that the variable is likely to have a meaningful impact on the outcome. This can be seen in Network Strength and Gender.

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 student churn. 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 student churn.

The further a coefficient is from zero, the stronger the effect of that covariate on the hazard of churn.

4.3.4 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.

The Cox Proportional Hazard method has a weakness which is that the proportional hazard assumption must be met. In Table 4.3, it can be seen that the covariate meets the assumptions as none of the covariates have p-values below 0.05, which suggests that there is strong evidence for the proportional hazard's assumption for any single covariate.

This means that, based on this test, the assumption that the hazard ratios are constant over time holds for these covariates.

4.3.5 Schoenfeld Residuals for High Costs Pricing

Despite all covariates having p-values above 0.05, indicating no strong evidence against the proportional hazard's assumption, High Costs Pricing ($p = 0.06$) tends to be very close to the threshold of 0.05 thus implying that, the Schoenfeld residual plot for this covariate is necessary to determine if there is any visible pattern over time.

Schoenfeld residuals are a diagnostic tool used in survival analysis to test the proportional hazards assumption of the Cox Proportional Hazards model. They are the differences between observed event times and the expected event times, under the model, at each event time.

Based on the visual inspection of the Schoenfeld residuals plot for High Costs Pricing, there does not appear to be a strong violation of the proportional hazard's assumption since the p-value for High Costs Pricing was not less than the threshold (0.05). The visual inspection further supports that there might not be a significant violation as there is no obvious trend

or pattern in the residuals over time, which suggests that the proportional hazards assumption might hold for this covariate.

4.4 Accelerated Failure Time (AFT)

The Accelerated Failure Time (AFT) model is often used when the Cox Proportional Hazard (PH) model does not hold. Although the null hypothesis of Cox PH was accepted, the study still uses the AFT in order to understand the direct effect of covariates on survival time and to make predictions about survival times.

- Weibull AFT Fitter: AIC = 815.516
- Lognormal AFT Fitter: AIC = 822.647
- Loglogistic AFT Fitter: AIC = 820.250

Among these models, the Weibull AFT Fitter has the lowest AIC value with 815.516, thereby providing the best fit to the data compared to the Lognormal and Loglogistic models. This means that, based on the AIC criterion, the Weibull distribution is the most appropriate for modeling your survival data in this context.

4.4.1 Weibull Model

Just like the Cox PH analysis, the coefficient and exp(coefficient) columns in Table 4.4 provide information about the relationship between the covariates to churn. A positive coefficient or exp(coefficient) > 1 (such as Residence and Poor Network Quality Coverage) indicates that an increase in the covariate is associated with an increase in the time to the event. These increase the time to churn. A positive coefficient in a Weibull AFT model means the covariate increases the time to the event, indicating a protective effect against the event occurring sooner.

A negative coefficient or exp(coefficient) < 1 (such as Data Exhaustion and Network Strength) suggests a decrease in the time to the event. These decrease the time to churn. A higher value of these covariates is associated with a higher likelihood of churn. The p-value column helps assess the statistical significance of each covariate. A low p-value (typically < 0.05) indicates

that the covariate is likely to have a meaningful impact on churn. This can be seen in Network Strength and Gender.

The intercept (lambda) with a coefficient of 2.112 and p-value of 0.00005, indicates a highly significant baseline increase in time to event and the intercept (rho) coefficient is 0.906 and a p-value of 0.00005 therefore indicating a highly significant baseline multiplicative effect on time to event.

The boxes (squares) in the plot represent the estimated coefficients for each factor in the Weibull AFT mode in Figure 4.5. These coefficients indicate the effect size of each covariate on the accelerated failure rate (churn). The positive values suggest that the covariate speeds up the churn rate, while negative values suggest it slows down the event (decreases the churn rate). 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 Model Comparison

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

Based on the comparison of the C-Index values and AIC in Table 4.5, it is known that the Weibull model shows a substantially lower AIC, indicating a better overall fit compared to the Cox PH model for churn modeling of telecommunication. The concordance index of the Weibull is slightly higher than that of the Cox PH thus implying that, the Weibull has a better predictive ability.

Chapter 5

Conclusion

Bibliography

[1] Bandim, M. (2022). Growth of Ghana's Telecommunications Industry. *Journal of African Telecommunications*, 45(2), 123-135.

[2] Ghana Web. (2023). Underwater Fiber Optic Cable Outages Impact Telecommunications in West Africa. Retrieved from Ghana Web.

[3] Kapur, R. (2018). Factors Driving Customer Churn in Telecom Industry. *International Journal of Telecom Studies*, 22(4), 321-334.

[4] Koranchirath, A. (2024). Customer Retention Strategies in Competitive Markets. *Telecom Review*, 30(1), 56-72.

[5] AP News. (2023). Underwater Cable Disruptions Affect West Africa's Internet Services. Retrieved from AP News.