Chapter 1

INTRODUCTION

1.1 Background of the Study

A business is only a business by the grace of its customers. Churn has become one of the ideal focuses of businesses that are seeking to stand for a very long time. Customer churn in business is the failure of a customer to subscribe to a particular business after expiration. Industries compete for customers and so signing out of your business is the signing in of another. This project seeks to identify and explain problems and their implications with solutions to the Telecel company in Kwame Nkrumah University of Science and Technology (KNUST).

Vodafone (Telecel) which was formerly called Ghana Telecom has worked for over 2 decades now. It is a telecommunication service provider with a great number of customers over the years. On 11th August 2016, Vodafone, now known as Telecel entered into an agreement with KNUST which was to be renewed every five years. In this agreement, every student in KNUST will be given a Vodafone sim with a monthly package which has been modified over the years. The study seeks at what end does students (default) customers lose interest in using this sim. At what point do customers (students) discontinue their usage of this sim? What drives this and how can this be solved?

1.2 Problem Statement

The study at the various places where the students (customers) reside and their limitations to the usage of the monthly packages. This also relates to the various levels of the students which is our main focus because most students tend to move from campus after their first year stay on campus. One legitimate question that can be asked is how well level 100 students use this sim with Wi-Fi available on campus. Will they continue to use this sim as they are moving out from campus knowing that the network is not stable outside campus? This and many more are the questions that the research work seeks to ask. The study in subsequent chapters will talk about the various strategies to solve this.

1.3 Research Objective

The objective of this work is to determine the factors contributing to student churn in KNUST.

1.4 Organization of the Study

There are five chapters in the research. The first chapter discusses the background information and the foundations for the research. The problem statement and goals are stated here. The second chapter of the study contains the literature review which includes several studies that

have been on the subject. The third chapter focuses on the methodology and data collection techniques. There are several ways of approaching this problem but the main focus here is by using a survival analysis modeling tool. The fourth chapter presents the models and their practical applications to the dataset. Detailed analysis is presented here with the aim of figures and diagrams to determine the optimal model on the research. The fifth chapter contains the study's conclusions as well as recommendations that can be made on the topic.

1.5 Significance of the Study

This research will go a long way to point out the problems faced by the Vodafone Network on campus and outside campus. A clinical example is that, the network is unstable outside campus and even on campus sometimes. When this problem is solved, and also when Vodafone decides to hand the school sim to first years who live off-campus alone, there will be efficient utilization and maximization of resources.

We will in the subsequent chapters review reports by other researchers on this same topic that will give detailed information on this project.

1.6 Limitations of the Study

The limitation of this research was that the data survey was done to accommodate students on the KNUST campus who were present during the college elections hence, the coverage was focused on only undergraduate students.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Mobile technology has become widely used, changing how students interact, communicate, and obtain information (Alqahtani & Altarifi, 2020). Acknowledging this change, telecom companies and colleges have teamed up to offer students exclusive mobile phone packages, including discounted SIM cards, data bundles, and call rates. For instance, Vodafone SIM cards are provided to students upon admission at Kwame Nkrumah University of Science and Technology (KNUST) thanks to a partnership with Telecel, formerly known as Vodafone Ghana.

Students often stop using these SIM cards due to various factors, leading to the loss of valuable customers. Understanding the factors driving student retention and churn of these SIM cards is vital for improving and enhancing network services. 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 of using the Vodafone SIM card. By applying survival analysis, this thesis will provide valuable insights into student retention and churn factors.

2.2 Customer Churn in Telecommunications

Customer churn, defined as a subscriber's discontinuation of a service or product, has a significant financial impact on telecommunications firms (Gupta et al., 2018). With annual churn rates in the telecom industry ranging from 20% to 40% (Liu et al., 2020), understanding and minimizing churn is important. Extensive research has been conducted using machine learning to explore factors influencing customer churn and develop strategies for customer retention. This section delves into prominent theories and models of customer retention and churn.

2.2.1 Theories and Models of Customer Churn

Several studies have analyzed customer retention in the telecommunications sector using various statistical models. Portela & Menezes (2011) utilized Accelerated Failure Time (AFT) models, identifying the log-logistic model as appropriate due to its low AIC, after the Cox PH model proved inadequate. Wong (2011) used Cox regression to explore churn predictors in Canadian wireless telecoms. Ahn et al. (2006) employed logistic regressions to study customer churn but faced limitations like missing data on account tenure and short data collection periods. Ocloo & Tsetse (2013) combined qualitative and quantitative methods in their study on Ghanaian mobile telecoms. Each model reviewed had weaknesses, such as ignoring survival times or making

distributional assumptions. This study aims to improve upon these models by using the Cox regression model and its extended version to determine the best approach for modeling telecom customer retention.

2.2.2 Application of Survival Analysis in Telecommunications

Several recent studies have successfully applied survival analysis to investigate customer churn in the telecommunications industry. These studies highlight the effectiveness of this approach in understanding customer churn dynamics and provide valuable insights for developing targeted retention strategies.

In a study titled "Predicting Customer Churn in a Telecommunications Company Using a Cox Proportional Hazards Model," Lee et al. (2018) used a survival analysis model to predict customer churn in a telecommunications business. Their research identified factors influencing the chance and lifetime of customers stopping the use of the network service over time. The study used survival analysis to analyze and predict the customers' risk based on:

- Service quality metrics: Network instability in some jurisdictions, duration of call credit and data, customer satisfaction, and loyalty to the provider.
- Call charges: High call charges leading customers to switch to other providers, suggesting reasonable pricing strategies.
- Customer satisfaction levels: Overall customer satisfaction with the service experience was a significant predictor of churn risk. Companies need to prioritize customer satisfaction to reduce churn.

By understanding these key drivers of churn risk, the telecommunications company could prioritize retention efforts and target interventions toward customer segments experiencing low service quality, high call charges, or dissatisfaction. This data-driven approach allows companies to improve customer lifetime value by mitigating churn and retaining satisfied customers.

Chen et al. (2020) studied customer churn in mobile network operators using a deep learning framework combined with survival analysis. They found that weak network coverage, heavy data usage exceeding plan limits, and specific customer demographics were significant predictors of churn. Improving network infrastructure, offering flexible data plans, and targeted retention campaigns were recommended strategies based on these findings.

Survival analysis is effective for studying customer churn by considering time and managing incomplete data. Its use in telecom shows promise. Applying it to student SIM card churn can reveal usage duration, factors affecting churn timing, and effective retention strategies. While previous research has explored customer churn in the telecommunications industry and the application of survival analysis for churn prediction, there are key gaps in understanding student

churn specifically related to school-provided SIM cards. This thesis aims to address these gaps and contribute valuable knowledge to this understudied area.

2.3 Research Gaps

This project aims to fill a research gap by studying student churn behavior related to school-provided SIM cards, an area with limited existing research. It acknowledges challenges in accessing usage data and proposes exploring alternative data sources. Additionally, it will investigate the impact of these SIM cards on student learning and academic experience, potentially uncovering valuable insights for retention strategies.

2.4 Project Contribution

This project will use survival analysis to study student churn with school-provided SIM cards at KNUST, addressing research gaps in this area. It aims to:

1. Estimate how long students use the SIM cards and identify factors influencing churn.
2. Investigate drivers of churn such as network coverage, data allowances, pricing competitiveness, student awareness, and brand loyalty.
3. Develop focused retention strategies based on these insights to enhance student adoption and usage of school-provided SIM cards.

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:

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

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

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

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.6

- (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 maximizing sample size, we use 0.5.
- Margin of Error (E): 5% since the confidence Level is 95%.

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

$$n_0 \approx 383.82$$

$$n_0 \approx 384$$

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

$$n \approx 384.16$$

$$n \approx 384$$

- Design Effect$(\boldsymbol{DEFF})$: 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 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 The Concept of Survival Analysis

Survival analysis is a branch of statistics used to analyze time-to-event of 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 is death, disease incidence, or some other negative experience. There in are some special cases where the failure is a positive event and not a negative one.

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.

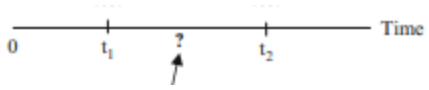

Figure 3.1: Interval Censoring

Interval-censoring incorporates both right-censoring and left-censoring. Left-censored data occurs whenever the value of t_1 is 0 and t_2 is a known upper bound on the true survival time. In contrast to the left-censored, right-censored data occurs whenever the value of t_2 is infinity, and t_1 is a known lower bound on the true survival time

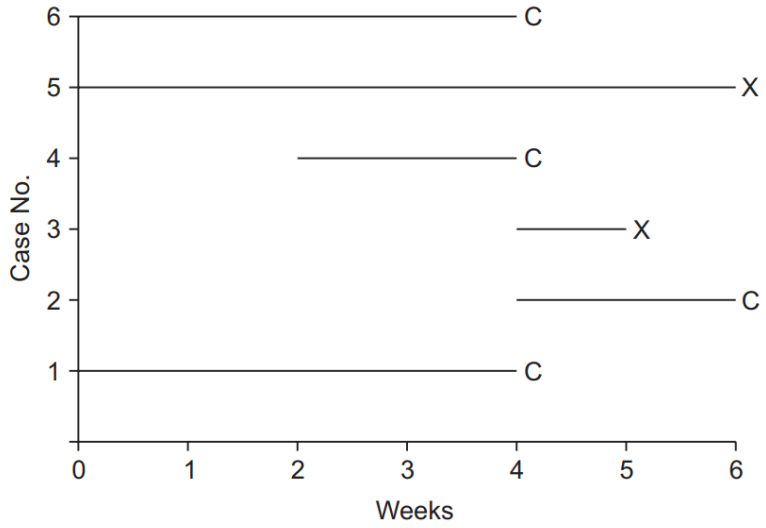

Figure 3.2: Censoring Graph

- C indicates censored data
- X indicates observed events

The 3.2 illustration shows 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 C). This could mean the participant dropped out of the study or the study ended before an event occurred for this individual. Case 2 and 4 is 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 X) 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

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

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

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

The first formula describes how the survival function $S(t)$ can be written in terms of an integral involving the hazard function. The formula stares 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 t to provide crucial summary information from the data.

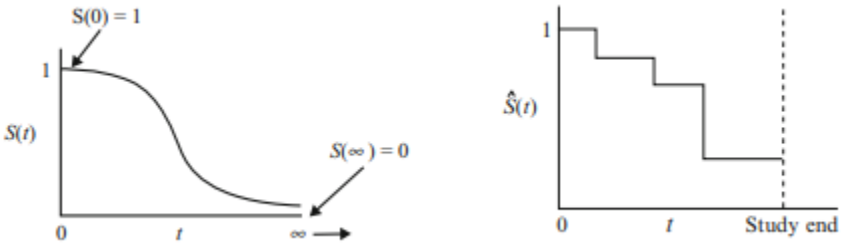

Figure 3.3: Survivor Curves

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 downwards 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 without a limit, the survivor curve will eventually reach 0.

When actual data is used, when 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).

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

$$h(t) \geq 0$$

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

The hazard function is also known as 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 denominator, $\triangle t$ denotes a small-time interval. By the division, a probability per unit 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 model. 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

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

Where:

- t is a time
- d_i the number of events (churn) at time ti
- $S(t)$ is the survival probability at time t
- n_i is the number of individuals at risk just before time ti

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.

$$h(t|x) = h_0(t)exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$$

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.
- $h_0(t)$ 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_j 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 CPH regression 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, approaches 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

RESULTS AND DISCUSSION

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.

Field Name	Description
Gender	Gender of the student
College	College of the student belong to
Churn	Whether the student has churned or not
Level	Academic level of the student
Residence	Whether the student lives on-campus or off-campus
Usage_Freq	Frequency of Vodafone network usage
Network_Strength	Strength of the Vodafone network
Voice_Calls	Usage of voice calls
Mobile_Data_Internet	Usage of mobile data for internet
SMS_Text_Messaging	Usage of SMS text messaging
Data_Exhaustion	Whether the student uses the entire 5GB in a month
Multiple_Networks	Whether the student uses multiple networks
Other_Networks_Better_Services	Whether other networks provide better services
Poor_Network_Coverage	Whether the student experiences poor network coverage
Insufficient_Data_Allowance	Whether the student finds data allowance insufficient
Unsatisfactory_Customer_Service	Whether the student is dissatisfied with customer service
High_Costs_Pricing	Whether the student finds Vodafone's pricing high
Monthly_Data_Usage	Monthly data usage of a student in gigabytes

Table 4.1: Data description of the features in the questionnaire

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.

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).

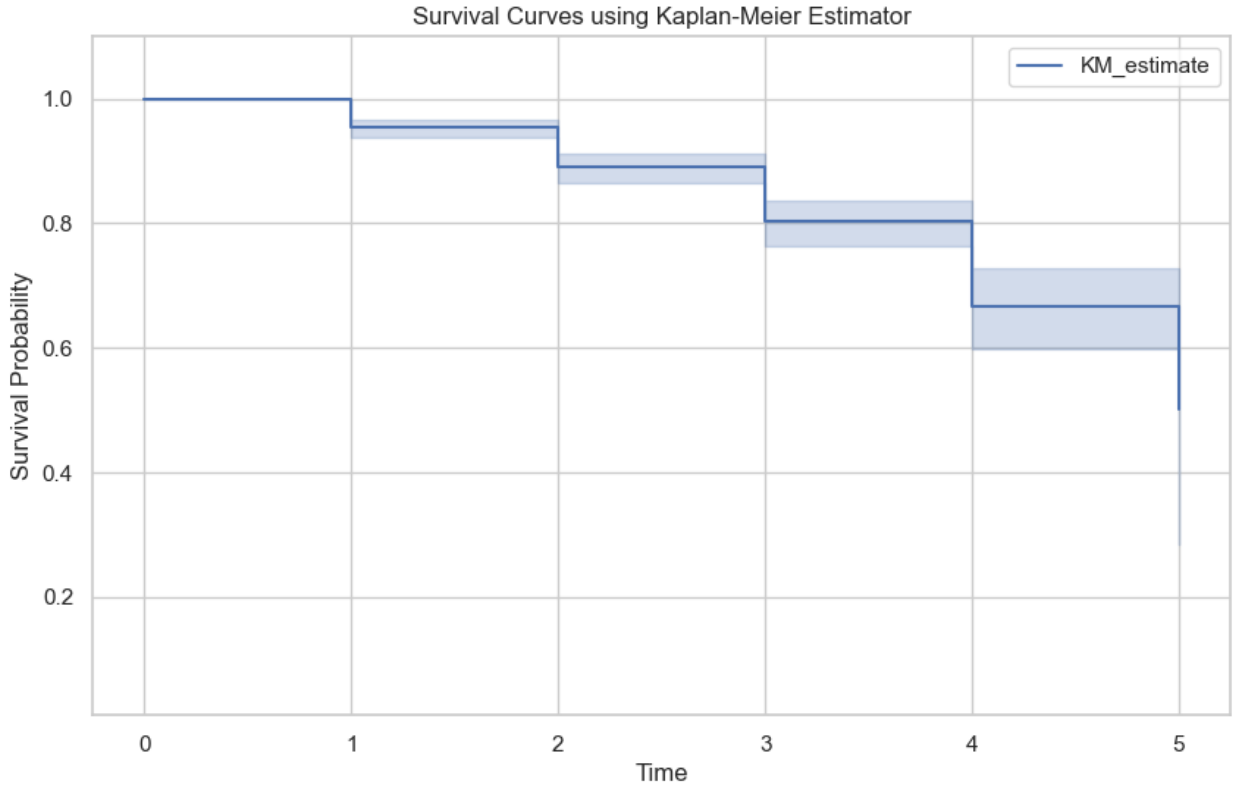

Figure 4.1: Kaplan Meier Curve of students from levels 100-500

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.

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.

Event Time	Number at Risk	Number of Censored	Survival Probability	Lower Confidence Interval	Upper Confidence Interval
0	768	0	1.000000	1.000000	1.000000
1	768	35	0.954427	0.937099	0.967065
2	733	37	0.890799	0.864030	0.912565
3	696	33	0.802521	0.763671	0.835681
4	663	17	0.667443	0.597011	0.728409
5	646	2	0.500583	0.285044	0.682828

Table 4.1: Kaplan Meier Estimate Analysis

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 2nd level where the initial 768 students from the beginning of the 1st year decreased to 733 for the 2nd year after 35 students churned at the end of the year. Subsequently, the survival probability declines gradually from 1 to 0.500583 by the 5th 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 Hazard regression model can be expressed as such:

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

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

Variable	Coefficient	EXP(Coefficient)	SE(Coefficient)	P
Gender	-0.56	0.57	0.20	<0.005
College	-0.03	0.97	0.05	0.51
Residence	-0.09	0.91	0.20	0.65
Usage_Freq	-0.00	1.00	0.06	0.96
Network_Strength	0.23	1.26	0.08	<0.005
Voice_Calls	0.16	1.18	0.23	0.48
Mobile_Data_Internet	0.32	1.37	0.27	0.24
SMS_Text_Messaging	-0.10	0.90	0.18	0.58
Data_Exhaustion	0.41	1.51	0.28	0.14
Multiple_Networks	0.21	1.23	0.42	0.62
Other_Networks_Better_Services	0.13	1.14	0.25	0.59
Poor_Network_Quality_Coverage	-0.17	0.85	0.19	0.38
Insufficient_Data_Allowance	0.12	1.12	0.19	0.54
Unsatisfactory_Customer_Service	-0.15	0.86	0.18	0.40
High_Costs_Pricing	0.16	1.18	0.18	0.36

Variable	Coefficient	EXP(Coefficient)	SE(Coefficient)	P
Monthly_Data_Usage	-0.06	0.94	0.07	0.38

Table 4.2: Detailed Cox PH analysis

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 Nerwork_Strength and Gender.

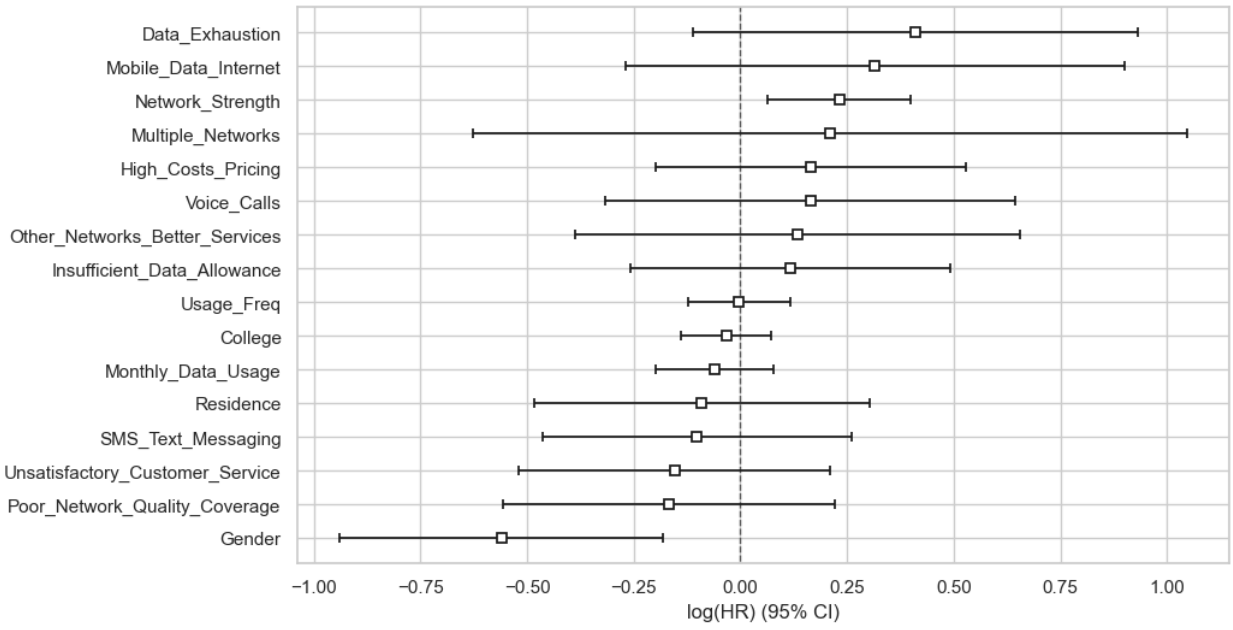

Figure 4.2: Coefficients of Cox PH

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 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.

Covariates	Test statistic	p
College	0.26	0.61
Data_Exhaustion	0.28	0.60
Gender	0.25	0.62
High_Costs_Pricing	3.61	0.06
Insufficient_Data_Allowance	1.45	0.23
Mobile_Data_Internet	0.21	0.64
Monthly_Data_Usage	0.10	0.75
Multiple_Networks	1.48	0.22
Network_Strength	1.43	0.23
Other_Networks_Better_Services	0.39	0.53
Poor_Network_Quality_Coverage	0.60	0.44
Residence	0.09	0.77
SMS_Text_Messaging	1.13	0.29

Covariates	Test statistic	p
Unsatisfactory_Customer_Service	0.26	0.61
Usage_Freq	0.03	0.86
Voice_Calls	1.60	0.21

Table 4.3: Cox PH assumption test

The Cox Proportional Hazard method has a weakness which is that the proportional hazard assumption must be met. In Table 4, it can be seen that the covariate meet the assumptions as none of the covariate 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.

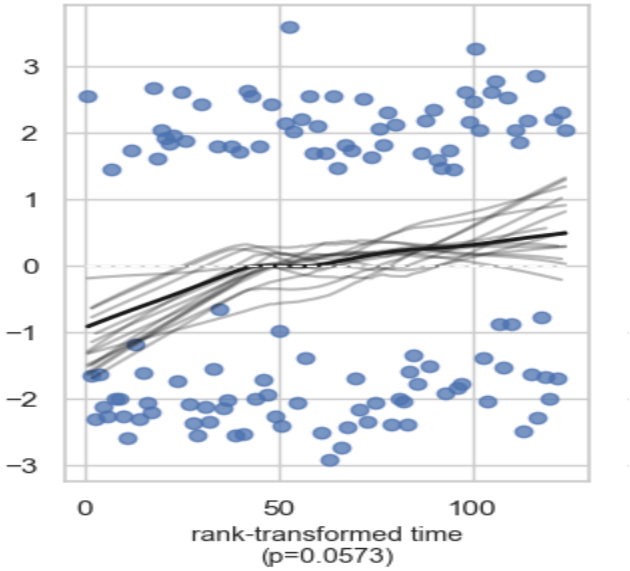

Figure 4.4: Schoenfeld residuals of High_Costs_Pricing

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 for making 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

Covariates	Coefficient	EXP(Coefficient)	P
College	0.017	1.017	0.440
Data_Exhaustion	-0.176	0.839	0.102
Gender	0.218	1.244	0.006
High_Costs_Pricing	-0.06	0.941	0.423
Insufficient_Data_Allowance	-0.050	0.951	0.521
Mobile_Data_Internet	-0.118	0.889	0.335
Monthly_Data_Usage	0.026	1.026	0.374
Multiple_Networks	-0.104	0.901	0.546
Network_Strength	-0.093	0.911	0.008
Other_Networks_Better_Services	-0.052	0.950	0.633
Poor_Network_Quality_Coverage	0.071	1.073	0.378
Residence	0.032	1.032	0.698
SMS_Text_Messaging	0.039	1.040	0.600
Unsatisfactory_Customer_Service	0.063	1.065	0.406
Usage_Freq	0.002	1.002	0.936
Voice_Calls	-0.080	0.923	0.421
Intercept (lambda)	2.112	8.266	0.00005
Intercept(rho)	0.906	2.471	0.00005

Table 4.4 Weibull AFT

Just like the Cox PH analysis, the coefficient and exp(coefficient) columns in table 4.4 provides 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 covariate 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 covariabe is likely to have a meaningful impact on churn. This can be seen in Network_Strength and Gender.

The intercept (lambda) with 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.

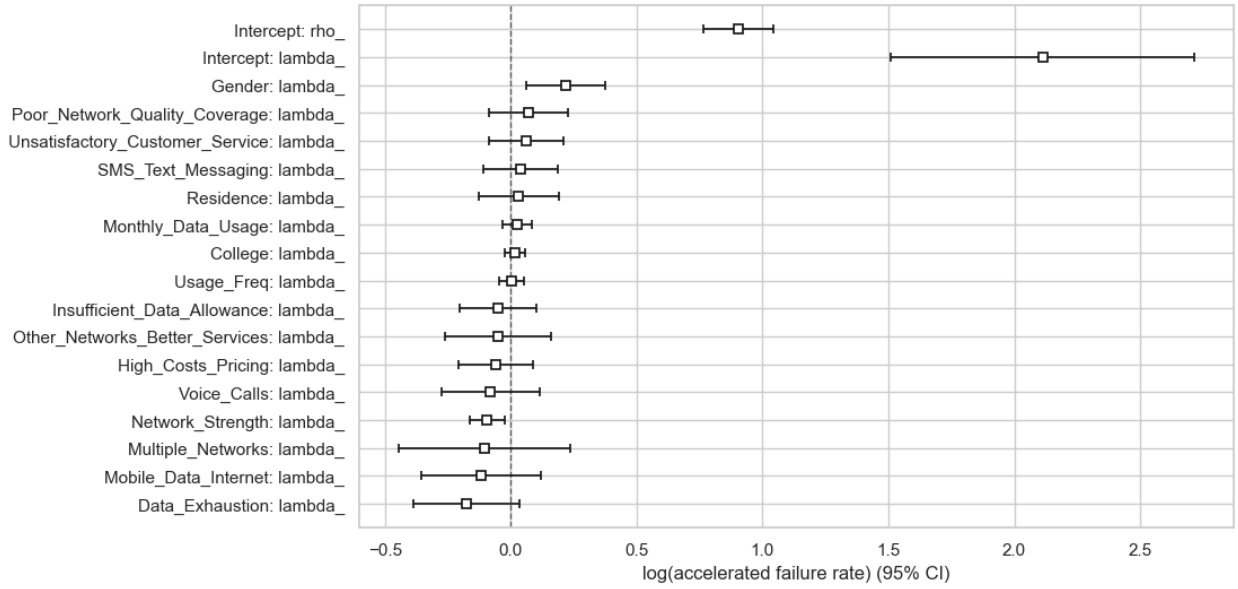

Figure 4.5: Weibull Coefficients

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 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.

Model	Concordance	AIC
Weibull	0.624	815.516
Cox PH	0.62	1479.47

Table 4.5: Model Comparison

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 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.