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.

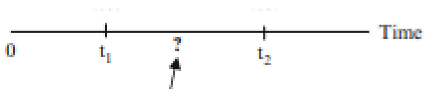

Figure 3.1: Interval Censoring

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

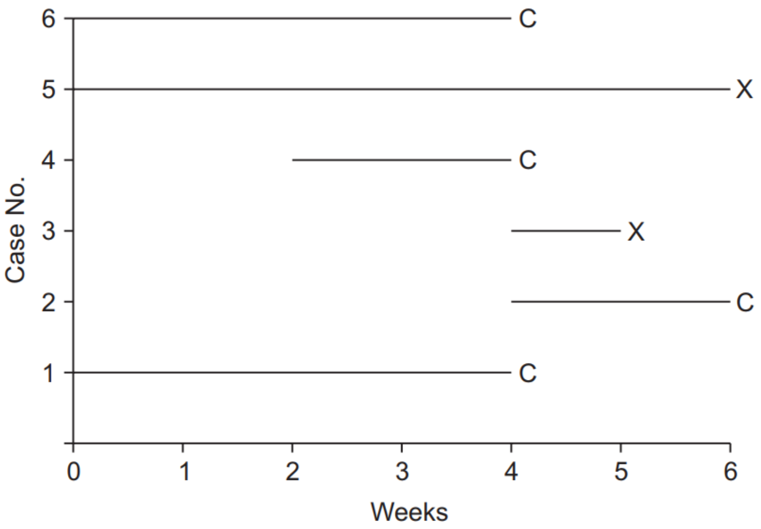

Figure 3.2: Censoring Graph

- 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 $\boldsymbol{t}$.

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

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.

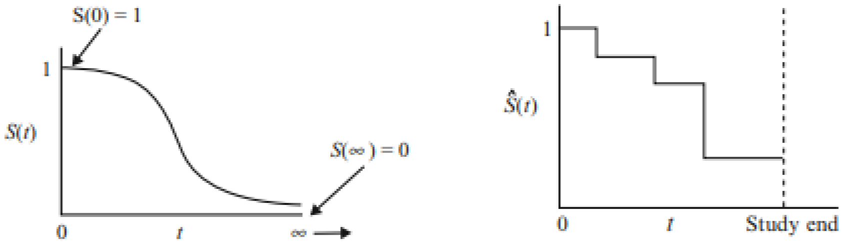

Figure 3.3: Survivor Curves

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.

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

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.