**CHAPTER 3 Keywords**

Kaplan-Meier Estimator, Cox Proportional Hazards Model, Random Survival Forest, Concordance Index, covariates

**METHODOLOGY**

**3.1 Introduction**

The chapter’s structure includes the methods as well as the procedures followed for the analysis to predict student churn of Vodafone (Telecel) within the context of KNUST. A comprehensive explanation of the models as well as 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 in student churn prediction.

**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 towards 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.

**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 comprises of 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 clusters have equal number of elements. It is therefore homogeneous within but heterogenous 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:

**Where:**

* is the population size.
* is the sample size for simple random sampling
* is the sample size for the population.
* is the sample size of the clusters.
* represents the critical value.
* is the estimated proportion of the population (e.g., proportion of students with a certain behavior).
* is the design effect, which accounts for the correlation among observations within the same cluster.6
* is the margin of error.

**Parameter Justification:**

* **Z-score**  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** : To maximizing sample size, we use 0.5.
* **Margin of Error**  5% since the confidence Level is 95%.
* **Design Effect** The design effect of 2 is used as the benchmark.

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 Preprocessing**

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. Missing values were only encountered in the feedback column as it was the only open question. This step ensures 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 within a variable. By converting categorical data into numeric form, we enabled the application of statistical and machine learning models that require numerical inputs. The `LabelEncoder` package from Python's `sklearn.preprocessing` module facilitated this transformation.

The transformation structured the dataset to facilitate survival analysis. Then, the data was organized to include essential components such as survival time, event indicators, and relevant covariates to the variables.

**3.7 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.

* **Survival Time** The time from a well-defined starting point (such as diagnosis or treatment) to the occurrence of the event of interest (like death or failure).

**3.7. 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.

**3.7.1 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.

**3.7.2 Left-Censored**

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.

**3.7.3 Interval-Censored**

It can occur if a subject’s true (but unobserved) survival time is within a certain known specified time interval.



**Table 3.1**

* indicates censored data
* indicates observed events

The 3.1 illustration shows the different types of censoring in a survival analysis study conducted over 6 weeks.

**Case 1:** This is 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.

**Case 2**: This is another instance of right censoring, but this participant was followed for a longer period, until about 6 weeks before being censored.

**Case 3:** This shows an observed event (indicated by ) occurring at around 4 weeks. This is not censored data, as the event of interest was observed within the study period.

**Case 4:** Right censoring occurring at 4 weeks.

**Case 5:** This represents a participant who was in the study for the entire 6-week period without experiencing the event of interest. This is not censored data.

**Case 6:** This is the same as case 1. This is also right censoring.

1. **Survival Function**  Also known as the survival probability function, it gives the probability that a subject survives beyond a specified time t.
2. **Hazard Function** Represents the instantaneous rate of failure at time t, given survival up to that time. It is defined as the probability that an event occurs at time t, given that the subject has survived up to time t.
3. **Cumulative Hazard Function**  The cumulative hazard at time t is the integral of the hazard function up to time t. It represents the total hazard experienced up to time t.

**3.8 Types of Survival Analysis**

**3.8.1 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.

**3.8.2 Semi-Parametric Methods**

These methods are primarily represented by the Cox Proportional Hazards Model, which is the most widely used semi-parametric method in survival analysis. The term "**semi-parametric**" refers to the fact that this model combines parametric elements (the effect of covariates) with non-parametric elements (the baseline hazard function).

**3.8 .3 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:

* **Kaplan-Meier Estimator**: Used to estimate the survival function from observed survival times.
* **Nelson-Aalen Estimator**: Used to estimate the cumulative hazard function.
* **Log-Rank Test:** Used to compare the survival distributions of two or more groups.

**3.7 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 be the survival probability at time. The estimator is computed as

Where:

* is a time
* the number of events (churn) at time
* is the survival probability at time
* ​ is the number of individuals at risk just before time

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 . Here, would be the time at which the first churn event occurred, would be 1 (since one churn event occurred), and would be the total number of students at that time

**3.8 Cox Proportional (Cox PH) Hazards Model**

The Cox Proportional Hazards model is a popular semi-parametric model for survival analysis. It models the relationship between the survival time and a set of predictor variables, assuming a proportional hazard rate.

where:

* is the hazard function, i.e., the instantaneous rate of the event occurring at time t given the predictor variables x.
* is the baseline hazard function, representing the hazard for individuals with all predictor variables equal to zero
* are the coefficients for the predictor variables

The coefficients 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. The Hypothesis of Proportional Hazard assumption is as follows:

H0: The Assumption of Proportional Hazard is fulfilled

H1: the assumption of proportional Hazard is not fulfilled

**3.9 Random Survival Forests**

Random Survival Forests extend the traditional random forest algorithm to the survival analysis setting. They are an ensemble method that combines multiple decision trees to improve predictive performance and handle censored data.

Random Survival Forests use a similar structure to traditional random forests but with modifications to handle right-censored data and to predict survival probabilities. The predicted survival probability at a specific time for a new instance can be computed as:

where:

* ) is the predicted survival probability
* is the total number of trees in the forest
* is the predicted survival probability from the tree

Each tree is constructed using a bootstrapped sample of the data, and the splitting criteria are based on survival-specific metrics like the log-rank statistic or the log-rank score.

For this model, we would need to build multiple decision trees and calculate the predicted survival probability for each tree.

**3.10 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 order or rank the predicted risks of individuals based on their actual event times. In survival analysis, the goal is often to predict the time until a specific event occurs, such as death, relapse, or failure. The Concordance Index evaluates whether the model's predicted risks align with the observed event times.

**Mathematical Computation:**

**Step 1: Define Pairs of Individuals**

* Create all possible pairs of individuals from the dataset.
* For each pair, compare their predicted risk scores and event times

**Step 2: Calculate Concordant and Discordant Pairs**

* A pair of individuals is concordant if the ordering of their predicted risks aligns with the ordering of their event times:
* A pair is discordant if the ordering of predicted risks is opposite to the ordering of event times.
* Pairs where event times are equal

**Step 3: Compute the Concordance Index**

The Concordance Index (C) is calculated as:

Interpretation:

* ranges from 0 to 1, where:
  + 0.5 indicates random guessing (no predictive ability)
  + 1 indicates perfect discrimination (perfect predictive ability
* A value above 0.5 suggests that the model has predictive ability better than random chance.
* Higher values indicate better model performance and more accurate risk predictions.