# Methodology

## Introduction

This chapter discusses the building of these models and how they were applied to predict student churn at Vodafone (Telecel) in KNUST. In this third chapter, the methodologies employed in the modeling process are delved into focusing on data processing and analysis techniques. An in-depth exploration of the mathematical concepts and theorems central to our research will be examined. The primary methodology used is survival analysis and this chapter elaborates on its underlying concepts, assumptions, theorems, and relevant techniques.

The chapter describes survival analysis in general, steps in survival analysis, the concept of censoring, and the specific models—the Kaplan-Meier Estimator, the Cox Proportional Hazards Model, and the Random Survival Forest—are discussed in detail. These models are applied to predict student churn at Vodafone (Telecel) within the context of KNUST.

The chapter structure includes a discussion on data collection methods followed by data pre-processing and analysis techniques. We then provide comprehensive explanations of the Kaplan-Meier Estimator, the Cox Proportional Hazards Model, and the Random Survival Forest. Mathematical formulations and interpretations are presented. Finally, we compare model performance using the Concordance Index shedding light on their predictive capabilities and practical applications in student churn prediction.

## Data Collection

The data used in this study comes from the telecommunication domain in KNUST. The primary data source for this research was a survey conducted among students at Kwame Nkrumah University of Science and Technology (KNUST). During the annual college elections, the survey targeted students across different academic levels (from levels 100 to 600). The dataset comprises 18 variables each capturing specific aspects relevant to our study.

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

For further details please refer to the figure below.

| **Variable** | **Description** |
| --- | --- |
| Gender | The students’ gender. |
| College | The specific college within the university. |
| Churn status | Is the student still using the sim |
| Level | The academic level of the student. |
| Residence | Campus or off-campus residence. |
| SIM usage | Whether the student uses a Vodafone SIM card. |
| Usage frequency | Frequency of SIM usage. |
| Network strength | Quality of the network (on a scale). |
| Voice calls | The student’s voice calls usage. |
| Mobile data usage | The student’s mobile data usage. |
| SMS text messaging | The student’s SMS usage. |
| Data exhaustion | The student’s data exhaustion experience. |
| Other networks | Usage of multiple networks. |
| Poor network coverage | Network signal experience |
| Insufficient data allowance | The student’s data allocation satisfaction. |
| Dislike of Customer service | Satisfaction with customer service. |
| High costs and pricing | The student’s pricing review. |
| Monthly data usage | Amount of data used monthly. |

## Sample Size Calculation

The sample size of the research was determined through cluster sampling. Cluster sampling is employed to determine the optimal sample size as it is a robust method that allows us to efficiently estimate population parameters by selecting entire clusters (colleges within KNUST) rather than individual students.

Here is the calculation of the sample size of the study:

Where:

* is the required sample size.
* represents the critical value corresponding to the desired confidence.
* is the estimated proportion of the population with a specific characteristic (e.g. proportion of students with a certain behavior or likeness).
* is the design effect that accounts for the correlation among observations within the same cluster.
* is the desired margin of error (expressed as a proportion).

Parameter Justification:

* Confidence Level (): We choose a confidence level of 95% to determine the range within which our estimate is likely to fall.
* Estimated Proportion (): If you don’t have an estimate, you can use for maximum variability (which yields the largest sample size).
* Margin of Error (): We use 5% since our confidence Level is 95%.
* Design Effect (): This accounts for the clustering effect. For simplicity, we assume a design effect of 2 since our population is not too massive.

Calculation:

Where:

* = sample size
* = z-score (1.96 for 95% confidence level)
* = estimated proportion (0.5 maximizing sample size)
* = design effect (2 accounting for cluster sampling)
* = margin of error (0.05 or 5%)

We therefore obtain a sample size of 768 students from our population of about 85,000.

Cluster Size: Since there are 6 clusters the sample size is evenly allocated across clusters. Each cluster would have approximately 128 students.

## 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 a 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 crucial pre-processing tasks involved the transformation of categorical variables into numeric format. This was achieved using label encoding, a technique that assigns unique integer labels to each category within a variable. By converting categorical data into a 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.

## Survival Analysis

Survival analysis is a branch of statistics used to analyze time-to-event data where the primary interest lies in the time until the occurrence of an event of interest. This could be anything from the failure of a mechanical part, the occurrence of a disease, to the death of a patient. Here are some key terminologies used in survival analysis:

* **Survival Time (T)**: 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).
* **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. There are two main types:
  + **Right-censoring**: This occurs when a subject is lost to follow-up or the study ends before the event occurs. The exact survival time is known only to be greater than a certain value.
  + **Left-censoring**: 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.
* **Survival Function (S(t))**: Also known as the survival probability function, it gives the probability that a subject survives beyond a specified time .
* **Hazard Function (h(t))**: Represents the instantaneous rate of failure at time given survival up to that time. It is defined as the probability that an event occurs at time given that the subject has survived up to time .
* **Cumulative Hazard Function (H(t))**: The cumulative hazard at time is the integral of the hazard function up to time . It represents the total hazard experienced up to time .

## Types of Survival Analysis

### 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.

### 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).

### 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.

## Kaplan-Meier Estimator

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.

Where:

* is a time point
* 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.

## Cox Proportional Hazards Model

The Cox Proportional Hazards model (Cox PH) 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 given the predictor variables .
* 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.

## 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 -th 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.

## 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:**

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.
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
3. Compute the Concordance Index:

**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.