**4 Analysis**

**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. The table below describes the dataset:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| **Gender** | The students' gender. |
| **College** | The specific college within the university. |
| **Churn status** | Indicates whether the student has stopped using the school sim. |
| **Level** | The academic level of the student. |
| **Residence** | Whether the student lives on-campus or off-campus. |
| **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** | Whether the student makes voice calls. |
| **Mobile data usage** | Whether the student uses mobile data. |
| **SMS text messaging** | Whether the student sends SMS texts. |
| **Data exhaustion** | Whether the student experiences data exhaustion. |
| **Other networks** | Whether the student uses other networks. |
| **Poor network quality and coverage** | Whether the student experiences poor network quality. |
| **Insufficient data allowance** | Whether the student's data allowance is insufficient. |
| **Unsatisfactory customer service** | Whether the student is dissatisfied with customer service. |
| **High costs and pricing** | Whether the student finds the pricing high. |
| **Monthly data usage** | Amount of data used monthly. |

**4.3.1 Kaplan Meier 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).

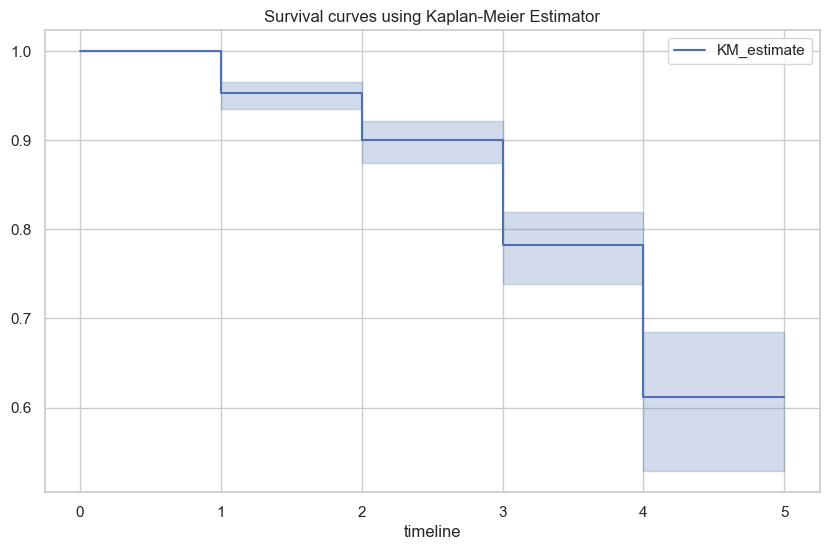
****

Table 4.1:

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 0 to 5 levels or years. The y-axis, likely representing survival probability, ranges from 0.6 to 1.0.

The KM estimate line graph shows the survival probability at various points along the timeline. 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 with a certain level of confidence.

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, 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**

| **Event Time** | **Number at Risk** | **Number of Event** | **Survival Probability** | **Lower CI** | **Upper CI** |
| --- | --- | --- | --- | --- | --- |
| 0 | 768 | 0 | 1.000000 | 1.000000 | 1.000000 |
| 1 | 768 | 36 | 0.953125 | 0.935608 | 0.965963 |
| 2 | 732 | 29 | 0.900775 | 0.874728 | 0.921650 |
| 3 | 703 | 39 | 0.782092 | 0.738667 | 0.819197 |
| 4 | 664 | 17 | 0.611636 | 0.528495 | 0.684555 |
| 5 | 647 | 0 | 0.611636 | 0.528495 | 0.684555 |

Table 4.2:

The provided Kaplan-Meier estimator output in table 4.1 summarizes the survival analysis 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 and thereby become censored. 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 732 for the 2nd year after 36 students churned at the end of the year. Subsequently, the survival probability declines gradually from 1 to 0.611636 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

**4.4 Cox PH Hazard**

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

| **Variables** | **Coefficient** | **Exp(Coefficient)** | **S.E(Coefficient)** | **Z** | **P** |
| --- | --- | --- | --- | --- | --- |
| Gender | -0.27 | 0.76 | 1.12 | -1.38 | 0.17 |
| College | -0.03 | 0.97 | 1.10 | -0.43 | 0.67 |
| Residence | -0.11 | 0.90 | 1.38 | -0.48 | 0.63 |
| SIM\_Usage | 0.63 | 1.87 | 5.41 | 1.15 | 0.25 |
| Usage\_Freq | 0.04 | 1.04 | 1.17 | 0.66 | 0.51 |
| Network\_Strength | 0.01 | 1.01 | 1.25 | 0.06 | 0.95 |
| Voice\_Calls | 0.08 | 1.08 | 1.94 | 0.26 | 0.79 |
| Mobile\_Data\_Internet | 0.05 | 1.05 | 1.80 | 0.18 | 0.86 |
| SMS\_Text\_Messaging | 0.14 | 1.15 | 1.72 | 0.69 | 0.49 |
| Data\_Exhaustion | 0.08 | 1.08 | 1.80 | 0.31 | 0.76 |
| Multiple\_Networks | -0.38 | 0.69 | 1.25 | -1.23 | 0.22 |
| Other\_Networks\_Better\_Services | 0.44 | 1.55 | 2.78 | 1.45 | 0.15 |
| Insufficient\_Data\_Allowance | -2.12 | 0.12 | 0.21 | -7.13 | <0.005 |
| Unsatisfactory\_Customer\_Service | -0.83 | 0.44 | 0.71 | -3.33 | <0.005 |
| High\_Costs\_Pricing | -1.45 | 0.23 | 0.37 | -6.23 | <0.005 |
| Monthly\_Data\_Usage | -0.07 | 0.93 | 1.11 | -0.80 | 0.42 |

**Table 4.3:**

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) indicates that an increase in the independent variable is associated with an increase in the odds of the outcome. Conversely, a negative coefficient (or exp(coefficient) < 1) suggests a decrease in the odds of the outcome.

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.

1. **Model Summary:**
   * The summary of the Cox model provides coefficients (coef), hazard ratios (exp(coef)), and p-values for each predictor variable:
     + **Positive Coefficients (e.g., SIM\_Usage, Other\_Networks\_Better\_Services):** These increase the hazard (risk) of churn. A higher value of these variables is associated with a higher likelihood of churn.
     + **Negative Coefficients (e.g., Insufficient\_Data\_Allowance, High\_Costs\_Pricing, Unsatisfactory\_Customer\_Service):** These decrease the hazard (risk) of churn. A higher value of these variables is associated with a lower likelihood of churn.
2. **Proportional Hazards Assumption Check:**
   * This test checks if the impact of the predictor variables on the hazard rate is constant over time. Variables violating this assumption might need further investigation or transformation.

* **Significant Predictors:**
  + **Insufficient\_Data\_Allowance:** Strong negative impact on churn, with a very low p-value, indicating it is a significant predictor.
  + **High\_Costs\_Pricing:** Also a strong negative impact with high significance.
  + **Unsatisfactory\_Customer\_Service:** Another significant negative impact on churn.
* **Other Predictors:**
  + **SIM\_Usage and Other\_Networks\_Better\_Services** have positive coefficients, indicating they increase the risk of churn but are less significant compared to the top predictors mentioned.

 **Insufficient\_Data\_Allowance, High\_Costs\_Pricing, and Unsatisfactory\_Customer\_Service** having a significant impact on churn is realistic. Customers often leave service providers due to inadequate data plans, high costs, and poor customer service.

*  **SIM\_Usage and Other\_Networks\_Better\_Services** increasing churn risk is also realistic. High SIM usage may indicate heavy users who are sensitive to service quality, and availability of better services from competitors is a common reason for churn.

Furthermore, the Cox Proportional Hazard regression model can be written as follows:

Cox Ph assumption Test

|  |  | **test\_statistic** | **p** | **-log2(p)** |
| --- | --- | --- | --- | --- |
| College | km | 0.62 | 0.43 | 1.22 |
| rank | 0.21 | 0.65 | 0.62 |
| Data\_Exhaustion | km | 0.00 | 0.96 | 0.05 |
| rank | 0.08 | 0.78 | 0.35 |
| Gender | km | 1.01 | 0.32 | 1.66 |
| rank | 0.70 | 0.40 | 1.31 |
| High\_Costs\_Pricing | km | 3.15 | 0.08 | 3.72 |
| rank | 2.82 | 0.09 | 3.43 |
| Insufficient\_Data\_Allowance | km | 0.25 | 0.62 | 0.70 |
| rank | 0.00 | 0.98 | 0.03 |
| Mobile\_Data\_Internet | km | 0.93 | 0.33 | 1.58 |
| rank | 0.76 | 0.38 | 1.38 |
| Monthly\_Data\_Usage | km | 0.39 | 0.53 | 0.91 |
| rank | 0.17 | 0.68 | 0.55 |
| Multiple\_Networks | km | 0.28 | 0.59 | 0.75 |
| rank | 0.01 | 0.92 | 0.12 |
| Network\_Strength | km | 0.09 | 0.76 | 0.40 |
| rank | 0.40 | 0.53 | 0.93 |
| Other\_Networks\_Better\_Services | km | 7.23 | 0.01 | 7.13 |
| rank | 5.09 | 0.02 | 5.37 |
| Residence | km | 0.21 | 0.64 | 0.64 |
| rank | 1.04 | 0.31 | 1.70 |
| SIM\_Usage | km | 0.67 | 0.41 | 1.27 |
| rank | 0.97 | 0.32 | 1.62 |
| SMS\_Text\_Messaging | km | 0.06 | 0.81 | 0.30 |
| rank | 0.06 | 0.81 | 0.31 |
| Unsatisfactory\_Customer\_Service | km | 2.60 | 0.11 | 3.23 |
| rank | 1.22 | 0.27 | 1.90 |
| Usage\_Freq | km | 1.55 | 0.21 | 2.23 |
| rank | 1.90 | 0.17 | 2.57 |
| Voice\_Calls | km | 0.00 | 0.95 | 0.07 |
| rank | 0.48 | 0.49 | 1.04 |