**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).

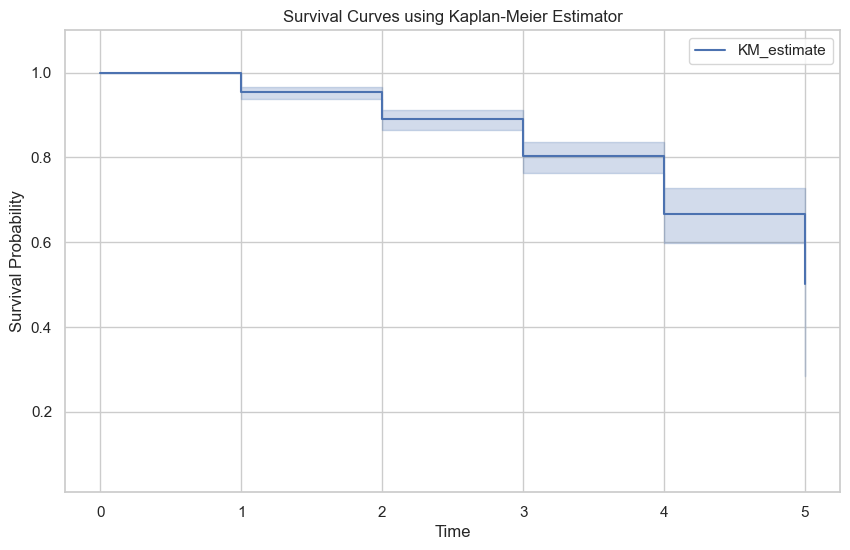
****

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:

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 |
| 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.

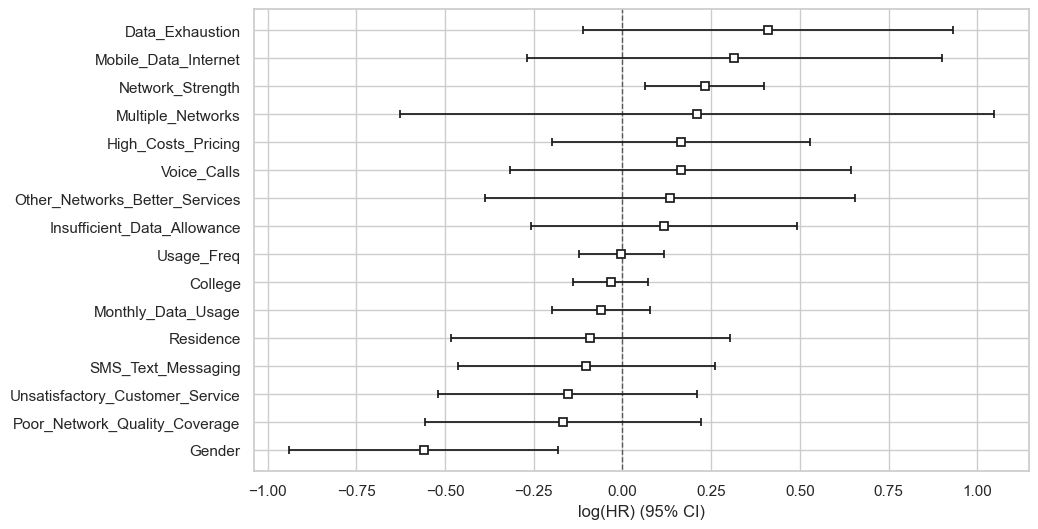
****

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 |
| 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.

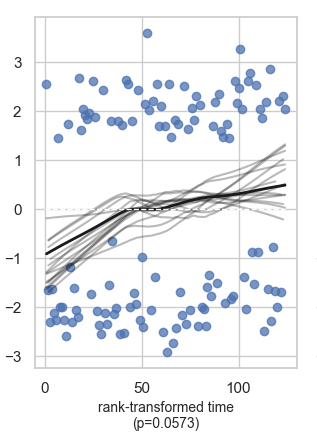
****

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.

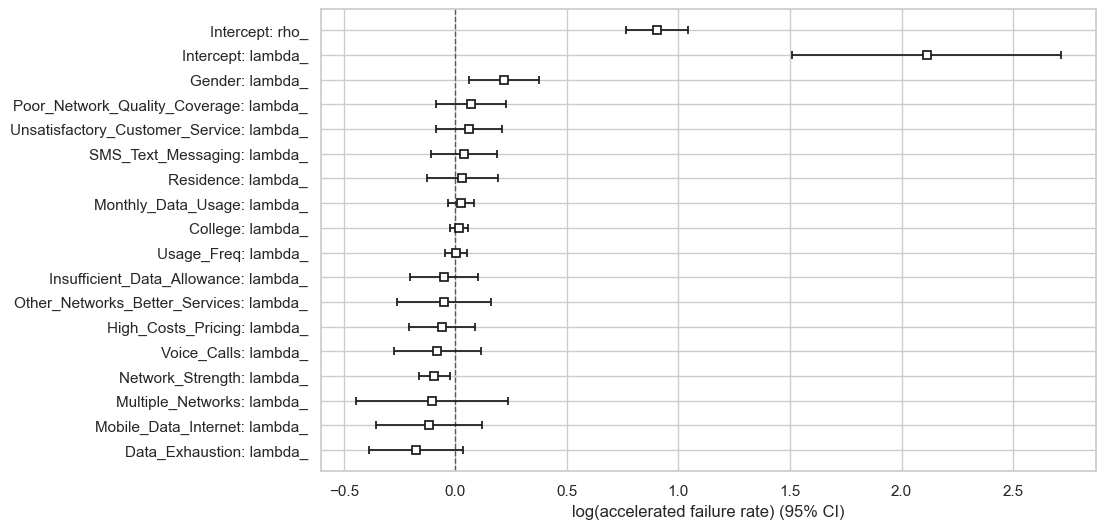
The intercept (lambda) with coefficient of 2.112 and p-value of 0.00005, indicates a highly significant baseline increase in time to event andthe intercept (rho) coefficient is 0.906 and a p-value of 0.00005 therefore indicating a highly significant baseline multiplicative effect on time to event.

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.