Group 11 - Survival Analysis

Bryen PARAMAGAMSAN - Ammar TARIQ - Ayoge BASSEY - Mohamed El'Arabi TEFFAHI

Introduction

In this project, we perform a comprehensive statistical analysis of customer subscription data from a digital product offering that includes financial advisory services such as newsletters, webinars, and investment recommendations. The primary focus of this analysis is to understand the dynamics of customer retention and attrition over time. The dataset, sourced from Kaggle (available here), provides information on customer demographics, subscription details, and customer interactions with the company's call center.

The analysis is structured around three key statistical methods:

1. **Nonparametric Survival Analysis**: This method will be employed to estimate the survival function, which represents the probability that a customer will continue their subscription beyond a certain time point. The Kaplan-Meier estimator will be used for this purpose, providing insights into the typical lifespan of a subscription.
2. **Nonparametric Comparison of Groups**: We will compare the survival functions across different customer segments, such as demographic groups or subscription types. This comparison will help identify whether certain groups are more likely to cancel their subscriptions than others.
3. **Semi-parametric Cox Regression**: To further explore the factors influencing customer attrition, we will apply Cox proportional hazards regression. This model will allow us to assess the impact of various covariates, such as age, gender, and product type, on the risk of subscription cancellation.

The outcomes of this analysis will provide insights into customer behavior, which can be leveraged to enhance retention strategies and optimize customer support operations.

Dataset overview

The dataset is composed of four correlated tables, each providing different aspects of information related to customer subscriptions and interactions. These tables are:

1. Customer Cases (`customer_cases`)

This table records individual customer interactions with the company's call center. Each row represents a unique case, providing details about the interaction, including the date and time of the case, the channel through which the customer reached out, and the reason for the interaction.

- **Columns**:
  - `case_id`: Unique identifier for each customer case.
  - `date_time`: Timestamp of when the interaction occurred.
  - `customer_id`: Unique identifier for the customer involved in the case.
  - `channel`: The communication channel used for the interaction (phone or email).
  - `reason`: The purpose of the interaction (signup or support).

2. Customer Information (`customer_info`)

This table contains demographic information about the customers. Each row corresponds to a unique customer, providing details such as age and gender.

- **Columns**:
  - `customer_id`: Unique identifier for each customer.
  - `age`: Age of the customer.
  - `gender`: Gender of the customer.

3. Customer Product (`customer_product`)

This table captures customer subscription details. Each row represents a subscription to a product, including when the customer signed up and, when they canceled the subscription.

- **Columns**:
  - `customer_id`: Unique identifier for the customer.
  - `product`: Identifier for the subscribed product.
  - `signup_date_time`: Timestamp of when the customer signed up for the product.
  - `cancel_date_time`: Timestamp of when the customer canceled the subscription (if they canceled).

4. Product Information (`product_info`)

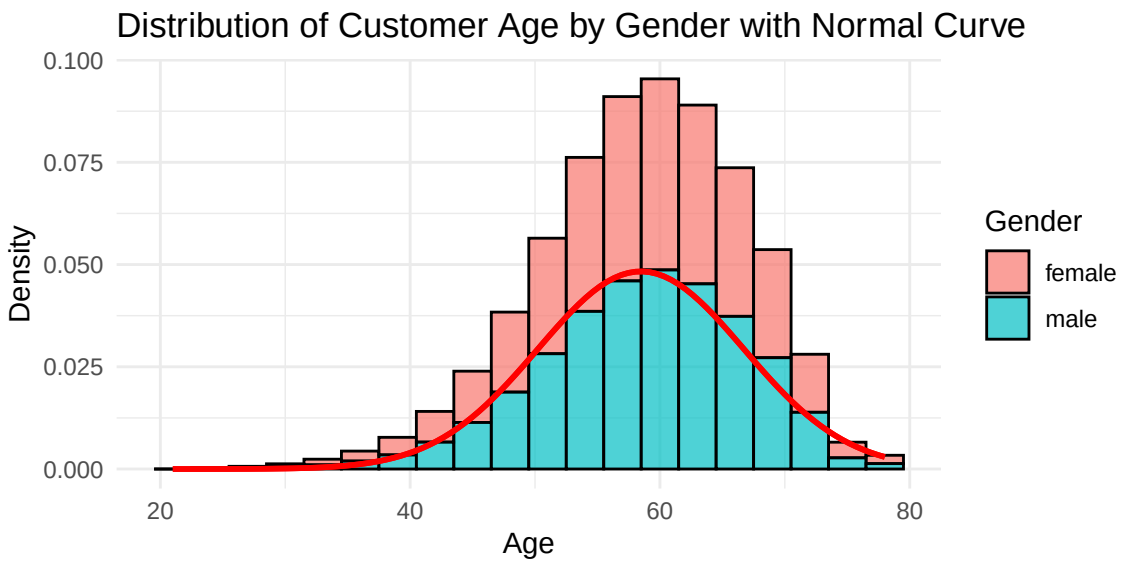
This table provides details about the products available for subscription. Each row corresponds to a unique product, including its pricing and billing cycle.

- **Columns**:
  - `product_id`: Unique identifier for each product.
  - `name`: Name of the product, they are two products in total.
  - `price`: Price of the product.
  - `billing_cycle`: The billing cycle of the product, indicating how often the customer is billed in number of months. It is either 1 for monthly, or 12 for annually.

Analysis of Age Distribution

The distribution of customer age is analyzed using three methods: a histogram with an overlaid normal distribution curve, a Q-Q plot, and the Shapiro-Wilk normality test. These methods help assess whether the age distribution follows a normal (Gaussian) distribution.

1. Analysis of Customer Age Distribution by Gender



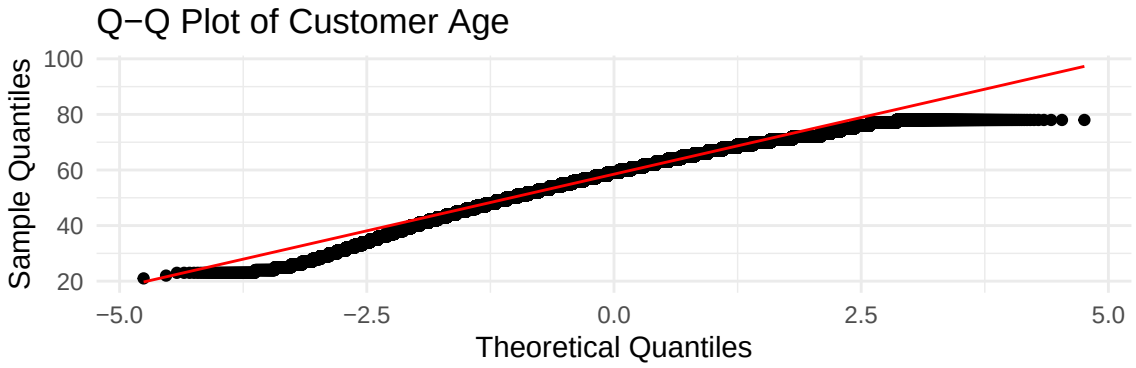
The histogram shows the distribution of customer ages for both male and female customers.

- **Central distribution**: Most customers are in the 50-60 age range, indicating that this middle-aged group is the most common among customers, suggesting the product or service appeals strongly to them.
- **Gender Proportions**: There are more female customers than male customers in the 50-65 age range, which may suggest the product or service is slightly more popular or retained better by females in this group. Males are also common in this range but in slightly lower numbers compared to females.
- **Age Extremes**: Fewer customers are under 40 or over 70, and the numbers are similar between males and females, indicating that the product or service is less appealing or relevant to these age groups.

2. Histogram with Normal Curve

The histogram shows the distribution of customer ages with a red curve representing the expected normal distribution, based on the mean and standard deviation of the data. Visually, the age distribution appears to be somewhat bell-shaped, which suggests it might approximate a normal distribution. However, there are notable deviations, particularly in the tails of the distribution.

3. Q-Q Plot



The Q-Q plot compares the quantiles of the observed age distribution to the quantiles of a normal distribution. In a perfectly normal distribution, all points would lie along the red diagonal line.

- **Observations**:
  - The points deviate from the line, particularly in the lower and upper quantiles, indicating that the age distribution is not perfectly normal.
  - The middle quantiles (representing the majority of the data) align relatively well with the line, suggesting that the central portion of the distribution is more normally distributed.

4. Shapiro-Wilk Normality Test

To statistically assess the normality of the age distribution, the Shapiro-Wilk test was performed on a random sample of 5000 customer ages.

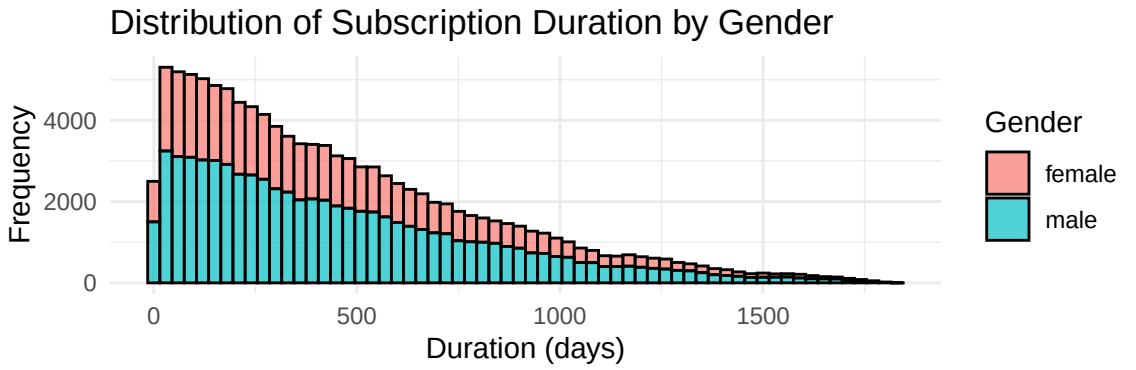
The Shapiro-Wilk test results indicate a p-value less than 0.05, suggesting that we can reject the null hypothesis that the data is normally distributed. This aligns with the visual observations from the Q-Q plot and histogram, where deviations from normality were observed, particularly in the tails.

5. Conclusion

The combination of visual and statistical methods suggests that while the age distribution has some characteristics of a normal distribution, particularly in its central range, it is not perfectly Gaussian.

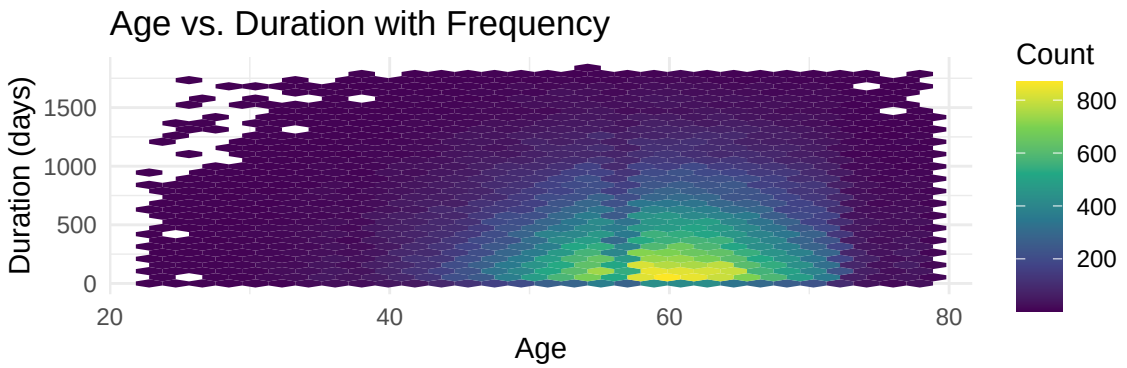
Analysis of Subscription Duration and Age by Gender

1. Distribution of Subscription Duration by Gender



- **Shape of Distribution**: The distribution of subscription duration is right-sided, indicating that most customers have shorter subscriptions.
- **Gender Differences**: Males tend to retain subscriptions slightly longer than females, particularly in the early subscription period (0-500 days). Both genders show a decline in frequency with increasing duration.

2. Age vs. Duration with Frequency (Hexbin Plot)



- **Concentration of Customers**: Customers aged 40-70 are most frequent and tend to have subscription durations up to 600-700 days.
- **Age Extremes**: Younger (<40) and older (>70) customers are less frequent and have a wider range of subscription durations.

3. Conclusion

- **Gender Insight**: Male customers may retain subscriptions slightly longer than females, particularly in the early stages.
- **Age Insight**: Middle-aged customers (40-70) are the most common and tend to have longer subscriptions.

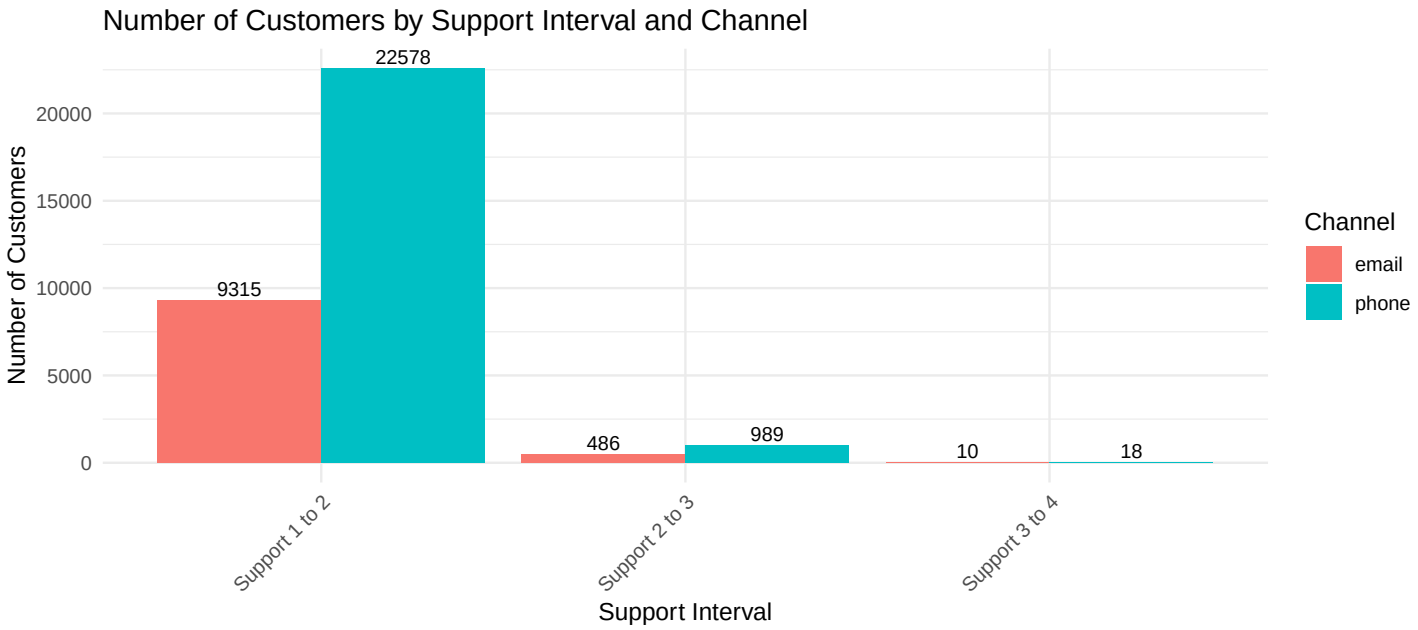
Analysis of Customer Support Contact Patterns

1. Visual Analysis

Distribution of Reasons by Channel

- The majority of customer interactions (both signup and support) occur through the phone channel, with email being significantly less utilized. This suggests that customers prefer real-time communication with the support.

Number of Customers by Support Interval and Channel

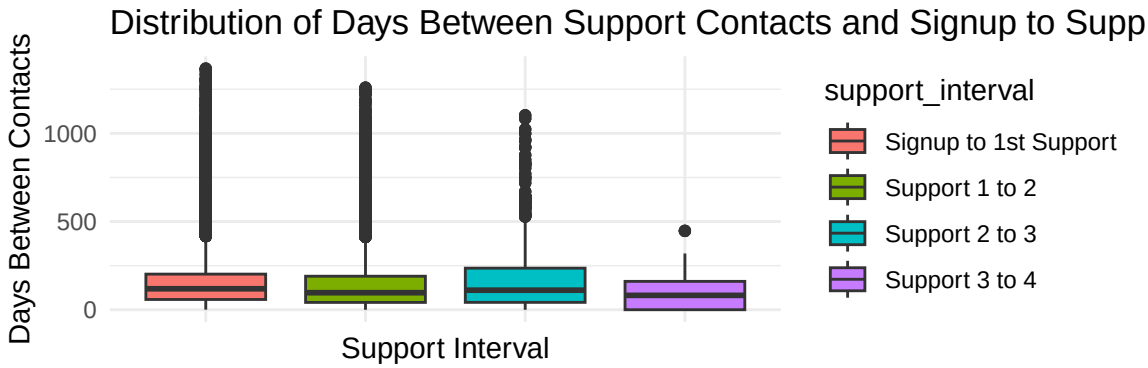


- The largest group of customers contacts support for the first time through the phone, and only a smaller fraction moves on to a second or third support interaction. By the time a third or fourth support interaction occurs, the number of customers decreases sharply, indicating that most issues are resolved within the first two interactions.

Customer Flow from Signup to Support and Cancellation

Average Days Between Support Contacts and Signup to Support

- On average, the time between signup and the first support contact is about 150 days, while subsequent support interactions tend to occur at shorter intervals. The reduced time between later support contacts might suggest unresolved issues or increasing customer frustration.



- However, when plotting the distribution of days between support contacts, it shows significant variability, particularly between the first and second interactions. The decrease in variability for later support contacts suggests that customers who need repeated support tend to do so in quicker succession.

2. Conclusion

- **Channel Preference**: Phone is the preferred channel for support.
- **Support Contact**: Most customer issues are addressed within the first or second support contact, but those requiring repeated support tend to escalate their interactions, with shorter intervals between contacts.
- **Timing**: A significant portion of customers are making their first contact for support within 150 days of signup.

Analysis of Product

They are two products : Product one, costing 1200 (currency unknown) and billed annually ; Product two, costing 125 and billed annually. If you subscribe to product two for a year, it's going to be more expensive than product one. They are more people subscribed to product one than product two, maybe because of the price advantage offered by the annual billing.

Kaplan-Meier Survival Analysis

Stratified by product

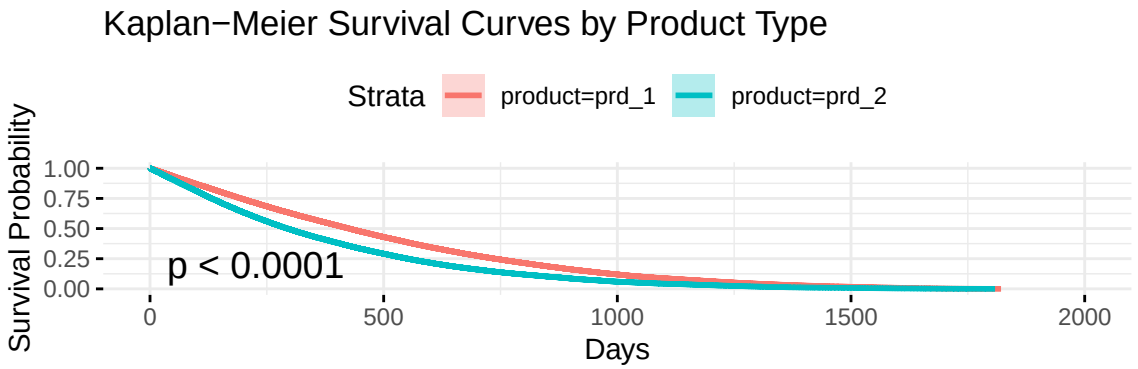
1. Introduction

In this analysis, we used the Kaplan-Meier estimator to estimate the survival probability of customers who subscribed to different product types (`prd_1` and `prd_2`). The survival probability represents the likelihood that a customer remains subscribed over time. The goal is to compare the retention rates between these product types.

2. Methodology

- **Survival Object Construction**:
  - We created a survival object, where the `duration` represents the time (in days) until a customer cancels their subscription, and the `cancellation` variable indicates whether the cancellation occurred (`Yes`) or (`No`).
- **Kaplan-Meier Estimator**:
  - We fitted the Kaplan-Meier estimator, stratifying by `product` to estimate and compare the survival curves for `prd_1` and `prd_2`.
- **Log-Rank Test**:
  - A log-rank test was performed to assess whether the differences in survival between the product types are statistically significant.

3. Results



- **Survival Probability**:
  - The survival curves show that the probability of a customer remaining subscribed decreases over time for both product types.
  - **Product** `prd_2` generally has a higher survival probability compared to **Product** `prd_1`, indicating better retention for `prd_2`.
- **Statistical Significance**:
  - The p-value from the log-rank test (`p < 0.0001`) indicates a statistically significant difference in survival between the two product types, suggesting that the observed differences are unlikely due to chance.
- **Confidence Intervals**:
  - The confidence intervals around the survival curves are narrow, providing a high degree of confidence in the estimated survival probabilities.
- **Number at Risk**:
  - The "Number at risk" table shows the number of customers still subscribed at various time points for each product, highlighting the decline in subscriptions over time.

4. Conclusion

The Kaplan-Meier analysis reveals that customers subscribed to `prd_2` have a better retention rate compared to those subscribed to `prd_1`. The significant difference in survival probabilities between the two product types suggests that `prd_2` may offer greater value or satisfaction to customers, leading to longer subscription duration.

Stratified by Age Group

1. Introduction

We did the same as previously but focused on age instead of products.

2. Results

- **Survival Probability**:
  - The survival curves show that the probability of a customer remaining subscribed decreases over time across all age groups.
  - There is a slight variation among the age groups, with younger age groups (e.g., 18-29) initially showing lower survival probabilities compared to older age groups, although these differences diminish over time.
- **Statistical Significance**:
  - The p-value from the log-rank test (`p = 0.013`) indicates that the differences in survival probabilities across age groups are statistically significant, though the practical differences between the groups are small.

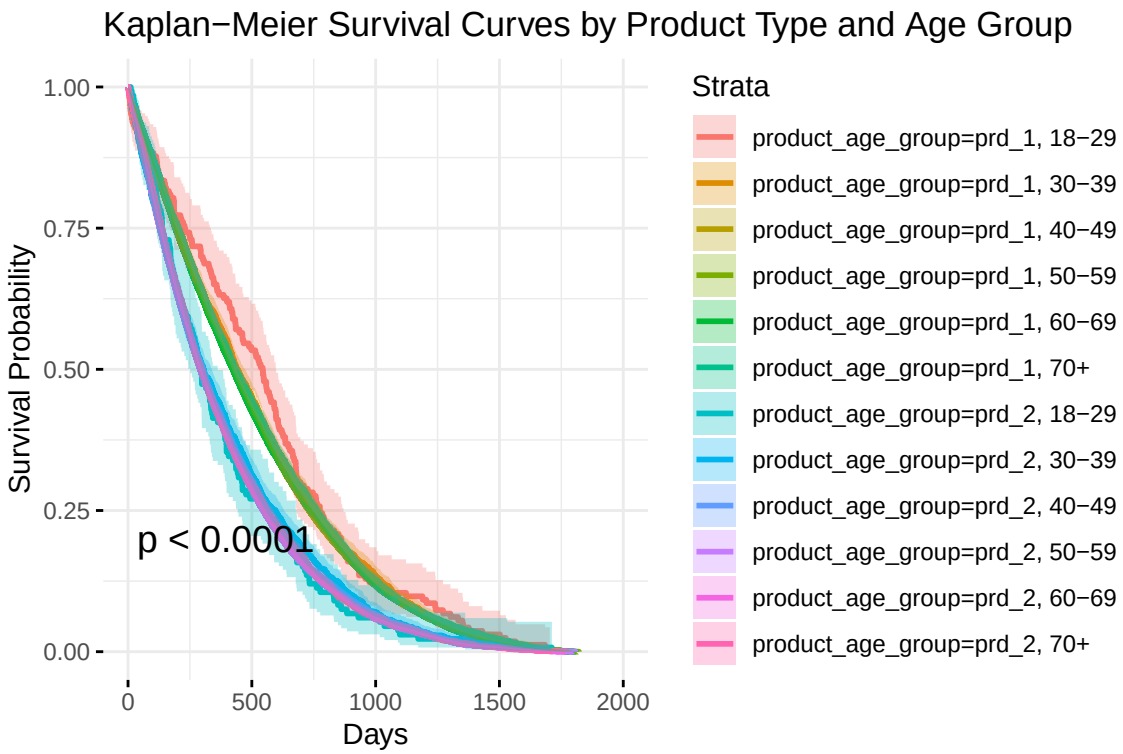
3. Conclusion

The Kaplan-Meier analysis stratified by age group reveals some differences in retention rates among different age groups, with younger customers showing slightly lower retention initially. However, these differences are not substantial, suggesting that age alone is not a strong predictor of retention.

Stratified by Age Group and Product

1. Introduction

We did the same as previously but focused on age and products.



2. Results

- **Survival Probability**:
  - Within each product type, there is some variation among the same age group (e.g., 18-29). For example, for product 1, younger age groups have higher survival probabilities than for product 2.

3. Conclusion

The Kaplan-Meier analysis, stratified by both product type and age group, reveals that **Product** `prd_2` generally has better retention across all age groups compared to **Product** `prd_1`. While age differences do exist, the product type appears to have a more substantial impact on customer retention. The statistically significant p-value reinforces the importance of considering both product type and age group in retention strategies.

Analysis of Nonparametric Comparison of Two or More Groups

1. Introduction

In this analysis, we performed a nonparametric comparison of survival distributions across multiple groups using the log-rank test. The groups were defined by combinations of product type and age group, and the goal was to assess whether there were statistically significant differences in the survival distributions among these groups.

2. Methodology

- **Log-Rank Test (`survdiff`)**:
  - The log-rank test was used to compare the survival curves across the different `product_age_group` combinations. This test assesses whether there are significant differences in the time to cancellation between the groups.
  - The test statistic is a chi-square statistic, with the degrees of freedom equal to the number of groups minus one.
- **Pairwise Comparisons (`pairwise_survdiff`)**:
  - In addition to the overall log-rank test, pairwise comparisons were conducted to identify specific pairs of groups that showed significant differences in their survival distributions.

3. Results

- **Overall Log-Rank Test**:
  - The chi-square statistic for the overall log-rank test was 2841, with 11 degrees of freedom.
  - The p-value was less than 2e-16, indicating that there are significant differences in the survival distributions among the different product and age group combinations.
- **Chi-Square Contributions**:
  - The contributions to the chi-square statistic varied across the groups, with some groups (e.g., `prd_1, 50-59` and `prd_2, 50-59`) contributing more to the overall statistic, indicating that these groups had larger discrepancies between observed and expected events.

- **Pairwise Comparisons**:
  - Pairwise comparisons revealed that most of the significant differences in survival probabilities were between the `prd_1` and `prd_2` groups, across all age groups.
  - For instance, the comparison between `prd_1, 18-29` and `prd_2, 18-29` showed a highly significant p-value ($p = 0.00039$), indicating that these two groups have significantly different survival distributions.
  - Within `prd_1`, differences between age groups were less pronounced, with most pairwise comparisons showing non-significant p-values (e.g., `prd_1, 18-29` vs. `prd_1, 30-39` with $p = 0.41476$).
  - On the other hand, significant differences were observed between age groups within `prd_2`, especially when comparing younger and older age groups (e.g., `prd_2, 30-39` vs. `prd_2, 50-59` with $p = 0.08772$).

4. Conclusion

The nonparametric comparison using the log-rank test indicates significant differences in customer retention between different product types and age groups. The largest discrepancies were observed between the two product types (`prd_1` and `prd_2`), with `prd_1` generally showing better survival (retention) rates across all age groups.

- **Product Comparison**:
  - There is a clear difference in survival distributions between `prd_1` and `prd_2` across all age groups, with `prd_1` consistently showing better retention.
- **Age Group Comparison**:
  - Within each product type, differences in survival distributions across age groups are less pronounced for `prd_1`. However, `prd_2` shows more variation, especially when comparing younger to older age groups.

This analysis provides evidence that product type significantly impacts customer retention, and age group also plays a role, particularly within the context of `prd_2`. These insights suggest that retention strategies may need to be adapted to the product but also to specific age demographics.

Cox Proportional Hazards Model Analysis

1. Introduction

In this analysis, we fit a Cox Proportional Hazards model to evaluate the impact of product type, age, gender, and support count on customer survival (i.e., the time until cancellation of the subscription). The survival probability represents the likelihood that a customer remains subscribed over time, adjusting for the specified covariates.

2. Methodology

- **Cox Model Construction**:
  - We constructed a Cox model using `product` type, `age`, `gender`, and `support_count` as covariates. The `duration` represents the time (in days) until a customer cancels their subscription, and the `cancellation` variable indicates whether the cancellation occurred (`Yes`).
- **Exploratory Attempts**:
  - We initially attempted to include the total cost paid by a customer up until cancellation as a covariate in the model. However, this caused significant overfitting, as indicated by a substantial increase in the concordance index, but with unrealistic predictions and unstable hazard ratios.
  - To mitigate overfitting, we also explored using a regularized Cox model with an elastic-net penalty (a mix of L1 and L2 regularization). Despite this, the model still overfitted when including the total cost, leading us to exclude this variable from the final model.
- **Model Fit and Interpretation**:
  - The model was fit on a dataset of 112,485 observations, with 396,447 observations deleted due to missing data (ongoing subscriptions). The output includes the coefficients, hazard ratios (`exp(coef)`), standard errors, and p-values for each covariate.

3. Results

- **Covariate Effects**:
  - **Product Type**: The `prd_2` product type has a significant positive coefficient (`coef = 0.3264`), with a hazard ratio of `1.386`. This indicates that customers who subscribed to `prd_2` are 38.6% more likely to cancel their subscription at any given time compared to those who subscribed to `prd_1`, holding other factors constant.
  - **Age**: The effect of age on survival is small but statistically significant (`p = 0.00887`), with a hazard ratio of `1.008`. This suggests that older customers have a slightly higher risk of cancellation, though the effect size is minimal.
  - **Gender**: The effect of gender, specifically being male, shows a non-significant trend (`p = 0.08798`), with a hazard ratio of `1.010`. This indicates that male customers have a slightly higher likelihood of cancellation compared to female customers, but this effect is not statistically significant.
  - **Support Count**: The `support_count` variable was found to have a minimal and statistically insignificant effect on survival (`p = 0.67431`), with a hazard ratio of `1.001`. This suggests that the number of times a customer contacted support does not substantially affect their likelihood of cancellation.
- **Model Performance**:
  - **Concordance Index (C-index)**: The model's concordance index is `0.546`, which indicates a slightly better-than-random ability to predict customer survival based on the covariates. However, this value suggests that the model has limited predictive power.

  - **Likelihood Ratio Test**: The likelihood ratio test yields a highly significant result (`p < 2e-16`), indicating that the model as a whole is significant and that the covariates contribute to the model's ability to predict survival.
- **Proportional Hazards Assumption**:
  - The global test of the proportional hazards assumption (`GLOBAL p < 2e-16`) indicates that there may be some violation of the proportional hazards assumption in this model, particularly for the `product` and `support_count` variables.

4. Conclusion

The Cox Proportional Hazards model indicates that the product type (`prd_2`) has a significant impact on the likelihood of customer cancellation, with customers of `prd_2` being more likely to cancel than those of `prd_1`. Age also has a minor effect on survival, with older customers being slightly more likely to cancel. Gender and support count do not significantly affect the likelihood of cancellation in this model.

Considerations Regarding Total Cost:

- Attempts to include the total cost paid by a customer until cancellation led to model overfitting, with an unrealistic increase in predictive performance. Even after applying regularization techniques like the elastic-net Cox model and standardizing the data (filtering extreme values, applying log transformation), the issue persisted, leading to the exclusion of this variable in the final model.

Overall Assessment:

While the model identifies some important predictors of customer cancellation, the low concordance index and potential violation of the proportional hazards assumption suggest that the model may have limitations in accurately predicting customer survival. Further exploration with alternative modeling approaches or inclusion of time-varying covariates may be necessary to improve model performance.

Conclusion

This project provided a comprehensive analysis of customer retention and attrition using various survival analysis techniques, including Kaplan-Meier survival analysis, nonparametric comparison of groups using the log-rank test, and Cox proportional hazards regression.

Key Findings

- **Customer Retention Patterns**:
  - The Kaplan-Meier survival analysis indicated that customers subscribing to `prd_2` have a higher risk of cancellation compared to those subscribing to `prd_1`.
  - Age was found to have a minor effect on the survival probability, with older customers being slightly more likely to cancel their subscriptions. However, the effect size was minimal.
  - The analysis revealed that support contacts had an insignificant impact on customer survival, suggesting that the frequency of contacting support does not strongly influence customer attrition.
- **Nonparametric Group Comparisons**:
  - The log-rank tests highlighted significant differences in survival distributions between different product types and age groups. Specifically, `prd_1` consistently showed better retention across all age groups compared to `prd_2`.
  - Pairwise comparisons indicated that age group differences within each product type were more pronounced for `prd_2`, especially when comparing younger and older customers.
- **Cox Proportional Hazards Model**:
  - The Cox model confirmed the significant impact of product type on customer retention, with `prd_2` customers being more likely to cancel their subscriptions.
  - Attempts to include the total cost paid by customers as a covariate resulted in model overfitting, even after applying regularization techniques like elastic-net Cox models. This suggests that the cost may introduce complexity that requires further exploration with different modeling approaches.
  - Despite the model's ability to identify significant predictors, the low concordance index and potential violations of the proportional hazards assumption indicate limitations in the model's predictive accuracy.

Recommendations

- **Retention Strategies**: The findings suggest that targeted retention strategies should focus on customers of `prd_2`, as they exhibit a higher likelihood of cancellation. Additionally, the company should consider tailoring interventions based on age demographics, particularly for younger customers within `prd_2`.
- **Model Improvements**: Given the limitations observed in the Cox model, particularly with the inclusion of total cost, future analyses could explore alternative survival models or time-varying covariates to capture the dynamics of customer retention more accurately. Further investigation into the reasons behind overfitting when including cost-related variables could also yield valuable insights.

In summary, while this analysis provided valuable information into customer retention patterns, there is still room for improving the predictive model.