Survival Analysis Project Report
Telco Customer Churn

August 12, 2024

Zakaria Bouhia	Eya Midouni	Maryam Teymouri	Oubeid Gharbi

1 Introduction

To enhance customer retention strategies, this project aims to apply survival analysis to Telco Customer Churn by IBM. We seek to investigate the timing of customer churn and uncover the key factors that influence it. By doing so, we will gain insights into customer behavior and develop targeted retention programs to reduce churn and boost customer loyalty.

Kaggle data: **https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data**

2 Methodology

This project will follow a structured methodology to analyze customer churn using survival analysis techniques. Initially, we will concentrate on configuring the R environment for our project.Then, data preparation and exploratory data analysis (EDA) to ensure the dataset is clean and suitable for modeling. Following data preparation, we will conduct EDA to understand the distribution of key variables and their relationships with churn.

Once the data is ready, we'll use several statistical techniques to analyze customer retention. First, we'll estimate survival functions for different customer groups using nonparametric methods to understand how long customers are likely to stay. Next, we'll compare these survival functions between different groups to see if there are any significant differences in churn rates. Finally, we'll apply semi-parametric Cox regression to explore how various factors affect customer survival and the likelihood of churn.

3 Exploratory Data Analysis (EDA)

Our analysis is rooted in a comprehensive understanding of the dataset, aiming to identify significant patterns and potential issues. The process can be broadly divided into two distinct stages:

Data Preparation :
This dataset provides detailed information on telecom customers, with each row representing a unique customer and each column capturing various attributes related to their service usage and demographics. The dataset contains 7,043 rows and 21 columns

While our dataset was generally well-structured with appropriate data types, we identified 11 missing values in the **TotalCharges** column. Instead of replacing these missing values, we decided to delete the customers with missing TotalCharges data. This approach ensures that our analysis is based on complete records, maintaining the integrity of the dataset without introducing any potential biases from imputation.

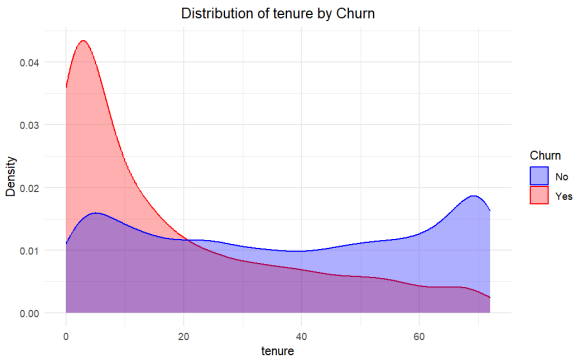
Data Analysis and Feature Engineering :

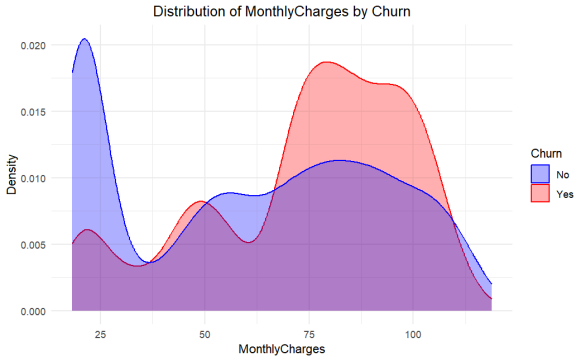
In this phase, we will focus on analyzing and visualizing the data to identify and retain the most significant variables for survival analysis. This involves examining the relationships between different features and customer churn, and assessing their impact. By performing EDA and feature engineering, we will clean and reduce the dataset, ensuring that we work with a streamlined set of important variables that enhance the effectiveness of our survival analysis. We conducted separate analyses for continuous and categorical factors to gain a deeper understanding of the dataset.

- **Continuous factors**:

we performed statistical analysis and visualizations, such as distributions, correlations, and summary statistics, to examine relationships and trends in variables like tenure, monthly charges, and total charges. You can go back to the code in order to see the full anaysis and visaulisation. The distribution analysis for both postiive and neagtive chrun for different Facotrs reveals:

- That customers with a short tenure (near 0 months) have a higher likelihood of churning, with churn probability decreasing as **tenure** increases. Long-term customers, who have been with the company for a longer period, are less likely to churn.
- Additionally, customers with higher **MonthlyCharges** are more likely to churn, particularly those with charges between 70 and 100 units, while those with lower monthly charges are less likely to churn. High monthly charges may contribute to dissatisfaction, suggesting the company should consider offering discounts or alternative plans to reduce churn.
- Furthermore, customers with lower **TotalCharges** are more likely to churn, consistent with the tenure findings, as these customers typically have shorter tenures and hence higher churn rates.

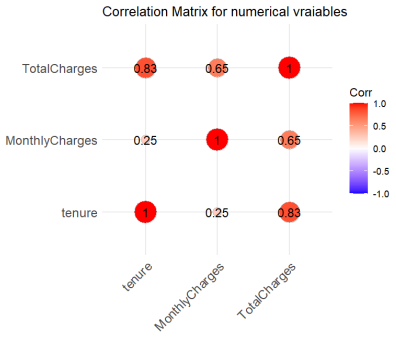




The correlation analysis shows the following relationships among the variables:

- **Tenure** and **TotalCharges** have a high positive correlation of 0.83, indicating a strong relationship where longer tenure generally leads to higher total charges.
- **MonthlyCharges** and **TotalCharges** also have a substantial positive correlation of 0.65, suggesting that higher monthly charges are associated with higher total charges, though the relationship is not as strong as with tenure.

  -**Tenure** and **MonthlyCharges** have a moderate positive correlation of 0.25, indicating that while there is some relationship between these two variables, it is weaker compared to their relationships with total charges.



Given these correlations, **Tenure** and **MonthlyCharges** are highly correlated with **TotalCharges**, justifying their retention for further analysis. These two factors effectively capture the variability in total charges and are crucial for understanding customer behavior related to churn.

- **Categorical factors**:

For categorical factors, we analyzed the frequency distributions and visualizations to explore the distribution of categories and their associations by **Churn**.

Our categorical data visualizations analysis highlights several key findings:

The **gender** does not significantly affect churn, as the distribution of churn is balanced between male and female. Additionally, **Senior** citizens seem less likely to churn compared to non-seniors, with the majority of churned customers not being senior citizens. Customers without a **Partner** or **Dependents** tend to have higher churn rates, suggesting that those with more personal responsibilities are more likely to stay. Customers lacking **PhoneService** or **MultipleLines** also show higher churn, possibly due to perceived lower value. Fiber optic users exhibit higher churn compared to DSL users or those without internet service. The absence of services like **OnlineSecurity**, **OnlineBackup**, **DeviceProtection**, and **TechSupport** is associated with higher churn, underscoring their role in retention. **StreamingTV** and **StreamingMovies**: movie services contribute to retention but are not decisive, as a notable proportion of users still churn. Long-term **Contract** are strongly linked to reduced churn, while monthly contracts are associated with increased churn. Additionally, for **PaymentMethod**, electronic billing and check payments correlate with higher churn rates, whereas automatic payment methods are associated with lower churn.

For deeper insights, we recommend revisiting the code provided for a more visualizations thorough analysis of categorical data.

To refine our analysis, we will use the Chi-Square χ^2 Test to identify significant associations between categorical factors and churn,which will help us focus on the most predictors with high impacts, streamlining our analysis and enhancing the accuracy of our survival analysis. The results showed that among the categorical variable tested **InternetService**, **OnlineSecurity**, **OnlineBackup**, **DeviceProtection**, **TechSupport**, **Contract**, and **PaymentMethod** had extremely low p-values, indicating a strong statistical association with churn. The factors with the most significant impact (lowest p-values) were identified, leading to a reduction in the number of variables for further analysis.

To advance our analysis, we calculated the correlation matrix after one-hot and labeling encoding the categorical factors. We found that OnlineSecurity, OnlineBackup, and DeviceProtection were highly correlated with each other. Therefore, we decided to retain only TechSupport for further analysis.

- **Conclusion**:

Through comprehensive data preparation, including addressing missing values, and detailed exploratory data analysis (EDA), we identified key continuous and categorical factors that significantly impact customer churn. By analyzing correlations and using the Chi-Square χ^2 Test, we refined our focus to the most influential factors, reducing the number of predictors to 10 (9 categorical and 1 continuous)as . This streamlined set of variables will be instrumental in enhancing the accuracy of our subsequent survival analysis, allowing for more precise modeling of customer churn behavior.

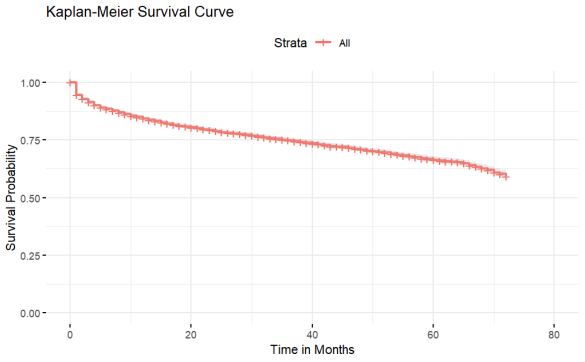
4 Survival Analysis

Survival analysis is a statistical approach used to analyze and predict the time until an event of interest occurs, such as customer churn in our example. After cleaning and reducing our dataset, we will apply several survival analysis methods. We will start with nonparametric estimation using the **Kaplan-Meier estimator** to estimate survival functions for different groups. Next, we will perform nonparametric comparisons to evaluate survival differences between these groups by applying the **Log_rank Test**. Finally, we will use the semi-parametric **Cox proportional hazards**

model to assess the impact of various factors on survival, allowing us to identify significant predictors of customer retention and churn.

- **Nonparametric estimation with Kaplan-Meier estimator** :

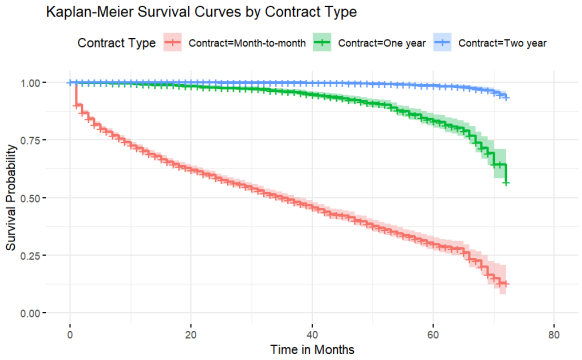
First, We applied the Kaplan-Meier estimation to the entire dataset to analyze the survival probability of all clients. In this analysis, we used the customer tenure as the time variable and churn as the event of interest. The Kaplan-Meier model was fitted to the data, and the resulting survival curve provides an overview of the survival probabilities over time for all customers.



The Kaplan-Meier survival curve shows the likelihood of customers staying with the company over time. Initially, all customers are loyal (survival probability = 1), but as time progresses, some leave, causing the survival probability to gradually decrease. By around 40 months, about 75% of customers remain. The shaded area around the curve represents the confidence interval, reflecting uncertainty in the estimates. Next, we applied the Kaplan-Meier estimation to analyze survival probabilities for multiple groups. This approach allowed us to compare how survival rates vary based on categorical factors.

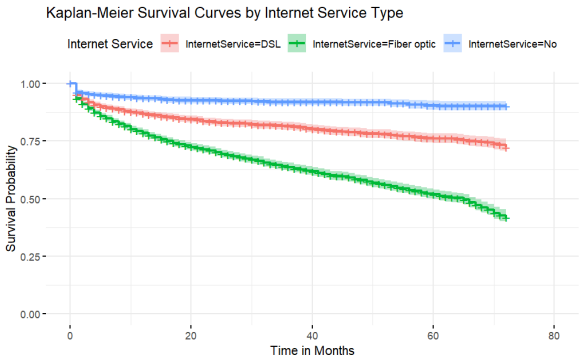
Here, we present some of the key results. For insights into other factors, please refer to the code provided.

Contract:



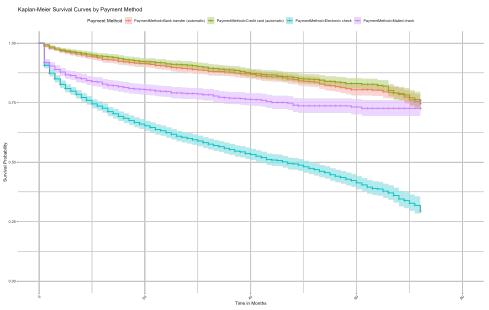
The Kaplan-Meier survival curve reveals that contract type significantly affects customer retention. Month-to-month contracts show the lowest retention probability and highest churn early on, while one-year contracts offer better retention but decline over time. Two-year contracts have the highest retention probability, with a more stable curve over the long term. Thus, longer contracts generally enhance customer retention.

InternetService:



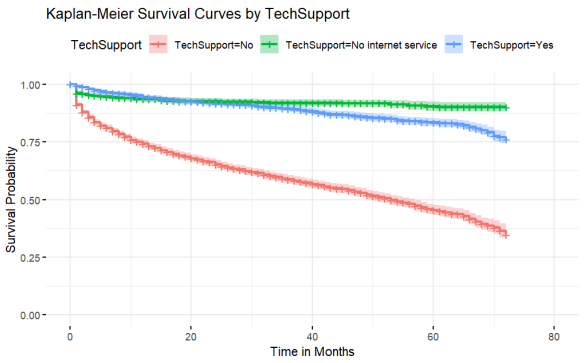
The Kaplan-Meier survival curve shows that customers with fiber optic internet service have the highest churn rate, while those without internet service exhibit the best retention, with a more stable survival probability over time.

PaymentMethod:



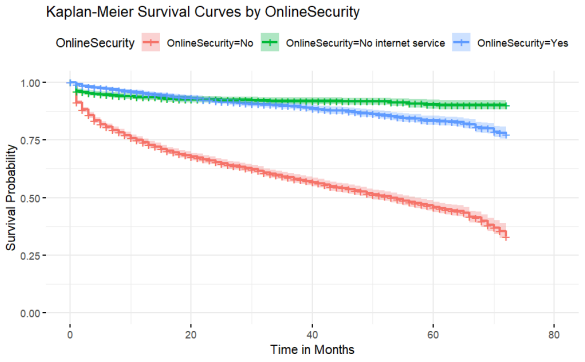
For PaymentMethod, the Kaplan-Meier survival curve shows that customers using electronic checks as their payment method cancel their subscriptions more quickly than those using other methods.

TechSupport:



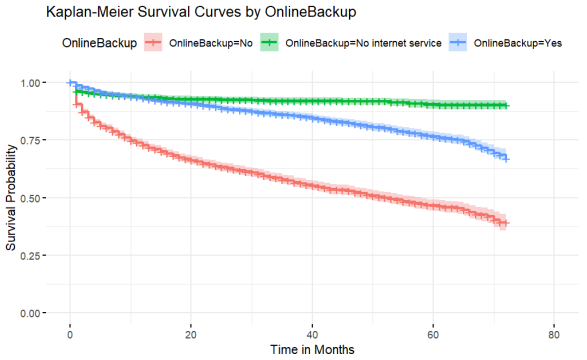
The Kaplan-Meier survival curve indicates that customers lacking technical support are more likely to churn, whereas those with technical support demonstrate better retention. Conversely, customers without internet service show the highest retention, reflecting a low churn rate.

OnlineSecurity:



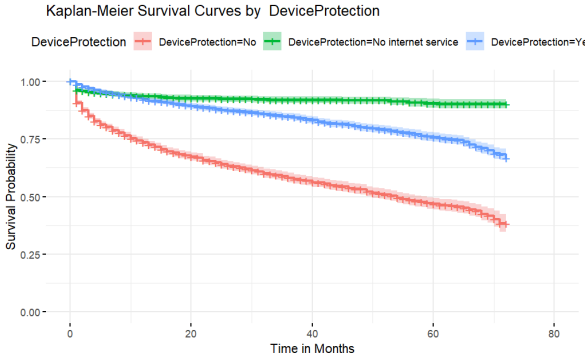
The Kaplan-Meier survival curve shows that customers without online security services have a higher churn rate and quicker decline in retention. In contrast, those with online security services maintain better and more stable retention. Customers without internet service exhibit the highest retention and lowest churn rate.

OnlineBackup:



The Kaplan-Meier survival curve shows that customers without online backup services experience higher churn rates with a rapid decline in retention, while those with online backup services have better and more stable retention. Customers without internet service have the highest retention and lowest churn rate.

DeviceProtection:



Same for device protection, The Kaplan-Meier survival curve shows that customers without device protection experience higher churn rates with a rapid decline in retention, while those with device protection have better and more stable retention. Customers without internet service have the highest retention and lowest churn rate.

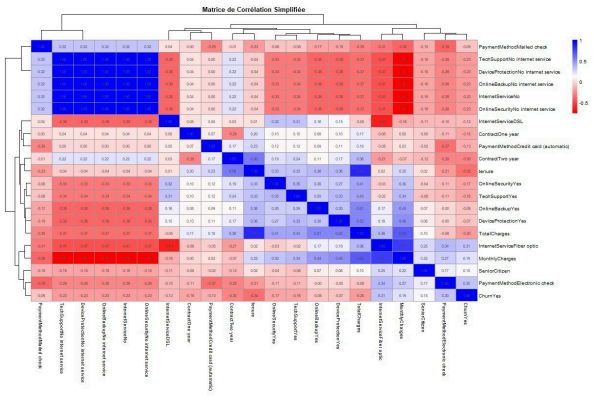
- **Non parametric comparison of two or more groups using Log-Rank Test**:

While the Kaplan-Meier estimation highlights the differences in survival probabilities, the comparison of groups using Log-Rank test assesses whether these differences are statistically significant or if they might be due to chance.

The Log-Rank test results reveal several significant patterns in customer churn.**SeniorCitizens** have a significantly lower rate of churn compared to non-senior citizens, indicating they are less likely to leave the service. Customers with longer **Contract** commitments, such as one-year or two-year contracts, exhibit lower churn rates, indicating that they are more likely to remain with the service compared to those with short-term contracts. **PaymentMethod** analysis shows that customers using electronic checks are more prone to churn than those using other payment methods. Additionally, for **InternetService**, fiber optic customers have a higher likelihood of churning compared to DSL or no internet service users. Offering **TechSupport**, **OnlineSecurity** and **OnlineBackup** services is associated with lower churn rates, suggesting that these features contribute to better customer retention. Similarly, **DeviceProtection** correlates with reduced churn, as customers with this service are less likely to cancel.

- **Semi-parametric Cox regression**:

By comparing the different survival curves, we observed considerable similarity among them, suggesting potential correlations between variables that could affect the Cox regression analysis. To address this, we calculated the correlation matrix after one-hot and label encoding the categorical factors. We found that OnlineSecurity, OnlineBackup, and DeviceProtection were highly correlated. As a result, we decided to retain only TechSupport for further analysis.



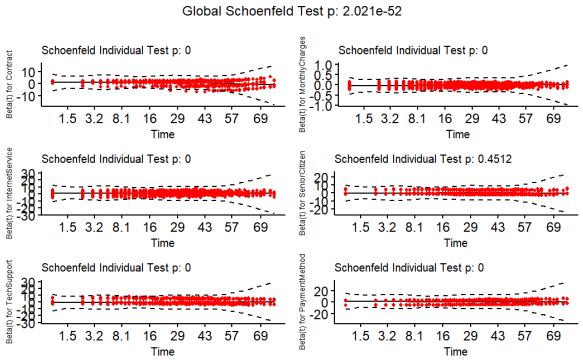
We fitted a Cox proportional hazards model to estimate the impact of various factors on the risk of customer churn. The model includes Contract, MonthlyCharges, InternetService, SeniorCitizen, TechSupport, and PaymentMethod to assess how each factor influences the hazard of churn.

The summary of the model provide insights into the significance and effect of these predictors on customer retention. The findings indicate that longer-term contracts, especially two-year agreements, significantly lower the risk of churn, with the strongest effect observed for the longest contracts. Fiber optic internet service is associated with a higher churn risk, likely due to dissatisfaction or competitive factors, while customers without internet service tend to exhibit lower churn rates, possibly because they are less likely to switch providers. Higher monthly charges are linked to a slight reduction in churn, suggesting that customers who pay more might be more satisfied or committed. Providing tech support is also shown to significantly decrease churn. Additionally, customers who use electronic or mailed checks are more likely to churn compared to those using automatic payment methods, which may be due to the convenience of automatic payments. The p-values from the Cox model confirm that all variables, except for SeniorCitizen and PaymentMethod: Credit Card (automatic), significantly influence churn risk. This indicates that the impact of being a senior citizen on churn is less significant when other factors are accounted for, highlighting the importance of various predictors in understanding customer retention.

- **Schoenfeld residuals**:

Schoenfeld residuals are used to test the proportionality assumption in the Cox model. In this analysis, residuals close to zero suggest that the assumption is met for certain covariates, such as ContractOne year and MonthlyCharges, indicating their effects on the hazard are proportional. Larger deviations from zero

may indicate issues with the proportionality assumption for other covariates..



- **Conclusion**:

Our survival analysis suggests that to boost customer retention, focus on encouraging longer-term contracts, addressing issues with fiber optic services, enhancing tech support, and incentivizing automatic payment methods.

5 Conclusion

This project focused on enhancing customer retention by applying survival analysis to Telco Customer Churn data. Our analysis revealed several key insights into factors influencing customer churn. Customers with longer tenures tend to exhibit lower churn rates, indicating that long-term loyalty is a significant factor in retention. Conversely, higher monthly charges are associated with increased churn, suggesting potential dissatisfaction or financial strain among customers with higher bills. Contract type plays a crucial role, with longer-term contracts (one-year and two-year) markedly reducing churn compared to month-to-month contracts. Additionally, providing services like tech support, online security, and backup is associated with improved retention, while fiber optic internet service is linked to higher churn rates. Payment methods also influence churn, with electronic checks correlating with higher churn rates compared to automatic payment methods. By leveraging these insights, Telco can refine its retention strategies, focusing on promoting longer-term contracts, enhancing service offerings, and encouraging the use of automatic payments to better manage customer churn and boost overall loyalty.