Predicting Customer Churn in the Telecommunications

Abstract:

This study aims to utilize survival analysis techniques to forecast customer churn based on data gathered from a telecommunications company. Unlike traditional statistical methods such as logistics regression and decision trees, which are effective at predicting churn, survival analysis allows for predicting not just whether customers will churn, but also when they are likely to do so and how long they will remain with the company. The insights garnered from this research can assist telecommunications companies in gauging customer churn risk and timing, aiding them in optimizing their strategies for customer retention and resource allocation to reduce churn effectively.

Introduction:

In the telecommunications industry, customers have the freedom to switch between multiple service providers, driving intense competition. Customers seek personalized products and better services at competitive prices, while providers prioritize acquiring new customers. With an average annual churn rate of 15-25% and the cost of acquiring new customers significantly outweighing that of retaining existing ones by 5-10 times, customer retention has surpassed customer acquisition in importance. Retaining high-value customers has become a top priority for many established operators. To address this, telecommunications companies implement retention strategies aimed at keeping customers longer by offering tailored products and services. Consequently, churn reduction has become a key business objective for many operators, integrated into their overall business goals and strategies.

To effectively aid telecommunications companies in reducing churn, it's crucial not only to identify customers at high risk of churn but also to determine when these high-risk customers are likely to churn. This enables companies to optimize their marketing efforts by allocating resources where they're most

needed, aiming to prevent churn proactively. Essentially, by understanding which customers are at risk and when they might churn, telecommunications companies can design tailored communication and intervention programs promptly and efficiently.

Traditional statistical methods like logistic regression and decision trees are adept at predicting customer churn, but they struggle to forecast when customers will churn or how long they will remain with the company. In contrast, survival analysis, originally developed for handling survival data, emerges as a potent tool for predicting customer churn due to its efficiency and effectiveness in addressing these timing-related aspects.

Objectives:

This study has two primary objectives. Firstly, it aims to estimate the customer survival function and hazard function to understand customer churn patterns over the duration of their tenure. Secondly, it seeks to illustrate the application of survival analysis techniques in identifying high-risk customers and predicting when they are likely to churn.

Definitions and Exclusions:

This section clarifies some of the important concepts and exclusions used in this study.

In the telecommunications industry, churn broadly refers to the termination of a customer's telecommunications service, encompassing both instances initiated by the service provider and those initiated by the customer. For instance, service-provider initiated churn could result from reasons such as payment default leading to the closure of a customer's account. On the other hand, customer-initiated churn is more complex and can stem from various reasons. In this study, we focus solely on customer-initiated churn, which is defined through a set of cancel reason codes. These codes encompass

reasons such as dissatisfaction with call quality, opting for a more appealing pricing plan offered by a competitor, receiving misleading information during sales interactions, unmet customer expectations, billing discrepancies, relocation, changes in business circumstances, among others.

High-Value Customers:

The study focuses solely on customers who have received a minimum of three-monthly bills. High-value customers are defined as those with a monthly average revenue of at least $X over the past three months. In cases where a customer's initial invoice covers less than 30 days of service, their monthly revenue is adjusted to reflect a full month's revenue through proration.

Granularity – This study examines customer churn at the account level.

Exclusions – This study does not distinguish international customers from domestic customers. However, it is desirable to investigate international customer churn separately from domestic customer churn in the future. Also, this study does not include employee accounts since churn for employee accounts is not of a problem or an interest for the company.

Survival analysis and Customer Churn

Survival analysis comprises statistical techniques used to examine the incidence and timing of events, originally developed for longitudinal data analysis. Customer churn tracking serves as a fitting example of survival data. Such data present two challenging characteristics for conventional statistical methods: censoring and time-varying covariates.

In general, the survival function and hazard function are utilized to depict the status of customer survival throughout the observation period. The survival function indicates the probability of surviving beyond a specific time point, while the hazard function portrays the risk of the event (in this context, customer churn) occurring within a time interval after time t, given that the customer has survived up to time t. Consequently, the hazard function is often preferred in survival analysis as it aims to quantify the immediate risk of

customer churn happening at time t, considering that the customer has already survived up to that point.

In survival analysis, the ideal observation plan is prospective, where we commence observing a group of customers from a defined starting point (referred to as the origin of time) and track them over a considerable duration, documenting instances of customer churn as they occur. It's not mandatory for every customer to experience churn; those yet to churn are termed censored cases, while those who have already churned are considered observed cases. Beyond merely predicting the timing of customer churn, we aim to analyze how time-dependent covariates (such as customer calls to service centers, changes in plan types, billing options, etc.) influence both the incidence and timing of customer churn.

Sampling Strategy

On August 16, 2000, a sample of 41,374 active high-value customers was randomly chosen from a telecommunications company's entire customer base. These customers were monitored over the subsequent 15 months, making August 16, 2000 the starting point (origin of time) and November 15, 2001, the endpoint for observations. Throughout this observation period, instances of customer churn were documented. Each customer in the sample was assigned a variable, DUR, indicating the time of churn occurrence or, for censored cases, the last observed time, both measured from the origin of time (August 16, 2000). Another variable, STATUS, was employed to differentiate between censored and observed cases, with a value of 1 denoting observed cases and 0 representing censored cases. The survival data in this study are solely right censored, meaning all censored cases are assigned a DUR value of 15 months.

Data Sources:

There are four major data sources for this study: block level marketing and financial information, customer level demographic data provided through a third-party vendor, customer internal data, and customer contact records. A brief description of some of the data sources follows. Demographic Data – Demographic dada is from a third-party vendor. In this study, the following are examples of customer level demographic information: - Primary household member's age - Gender and marital status - Number of adults - Primary household member's occupation - Household estimated income and wealth ranking - Number of children and children's age - Number of vehicles and vehicle value - Credit card - Frequent traveler - Responder to mail orders - Dwelling and length of residence.

Demographic Data – Demographic dada is from a third-party vendor. In this study, the following are examples of customer level demographic information:

- Primary household member's age
- Gender and marital status
- Number of adults
- Primary household member's occupation
- Household estimated income and wealth ranking
- Number of children and children's age
- Number of vehicles and vehicle value
- Credit card - Frequent traveler
- Responder to mail orders
- Dwelling and length of residence

Customer Internal Data

The customer internal data sourced from the company's data warehouse comprises two main components. The first part pertains to customer details such as market channel, plan type, billing agency, customer segmentation code, ownership of other company products, disputes, late fee charges, discounts, promotions, additional lines, toll-free services, rewards redemption, billing disputes, and similar information. The second part encompasses the telecommunications usage data of customers.

- Weekly average call counts
- Percentage change of minutes
- Share of domestic/international revenue

Customer Contact Records – The Company's Customer Information System (CIS) holds comprehensive records of customer interactions, including calls made to service centers and correspondence sent to customers via mail. These records are categorized into various customer contact categories, such as general inquiries, requests for service changes, inquiries about canceling services, and others.

Modeling Process:

Model process includes the following four major steps.

Explanatory Data Analysis (EDA) –

Explanatory data analysis was conducted to prepare the data for the survival analysis.

Imported the data from the provided dataset.

	total_rech_data_6	total_rech_data_7	total_rech_data_8	total_rech_data_9	count_rech_2g_6	count_rech_2g_7	count_rech_2g_8	count_rech_2g_9	count_
count	25153.000000	25571.000000	26339.000000	25922.000000	25153.000000	25571.000000	26339.000000	25922.000000	251
mean	2.463802	2.666419	2.651999	2.441170	1.864668	2.044699	2.016288	1.781807	
std	2.789128	3.031593	3.074987	2.516339	2.570254	2.768332	2.720132	2.214701	
min	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
50%	1.000000	1.000000	1.000000	2.000000	1.000000	1.000000	1.000000	1.000000	
75%	3.000000	3.000000	3.000000	3.000000	2.000000	2.000000	2.000000	2.000000	
max	61.000000	54.000000	60.000000	84.000000	42.000000	48.000000	44.000000	40.000000	

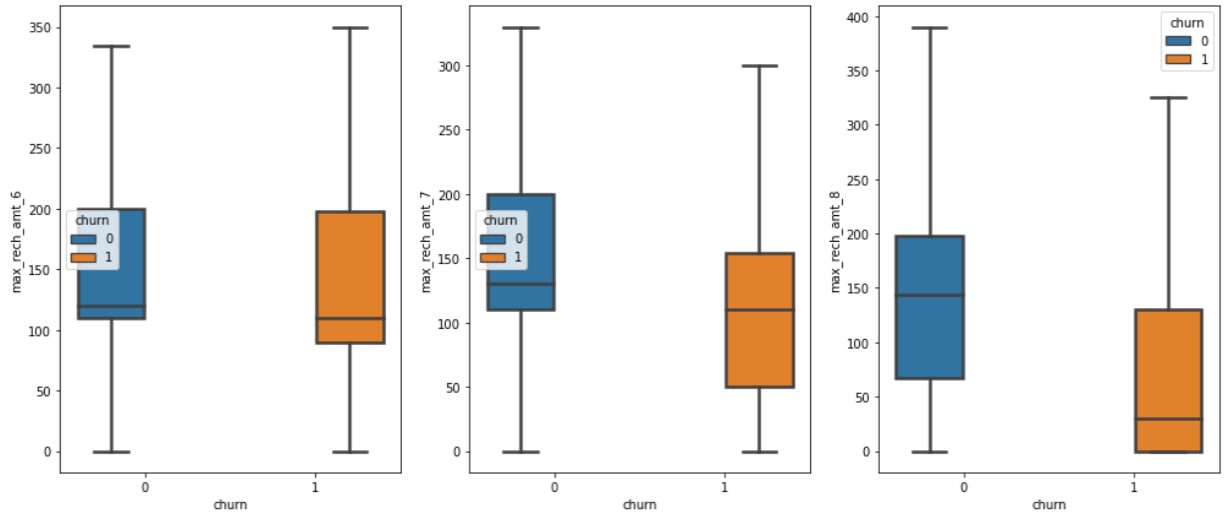
The customer failed to recharge due to missing recharge date and value information.

	total_rech_data_7	date_of_last_rech_data_7
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN
5	NaN	NaN
6	NaN	NaN
7	NaN	NaN
10	NaN	NaN
11	NaN	NaN
12	NaN	NaN
13	NaN	NaN

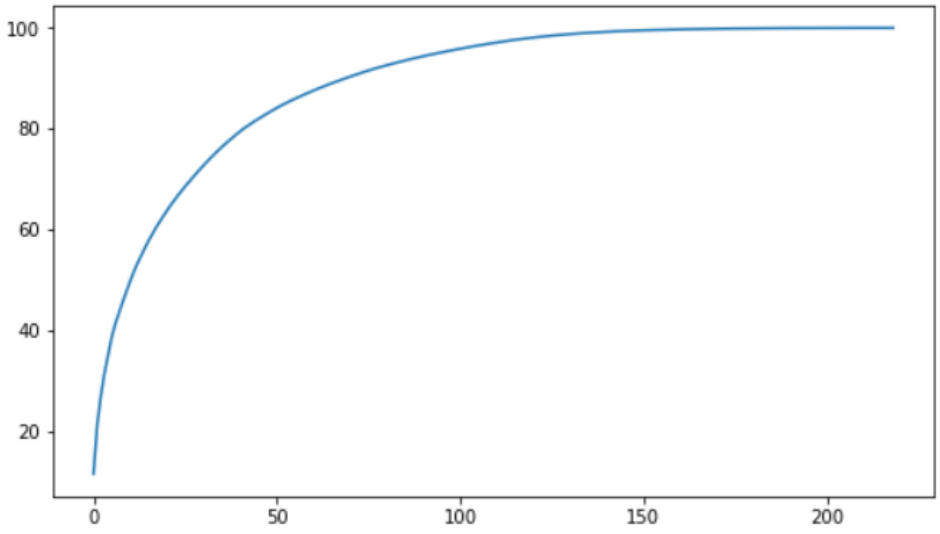
Eliminate variables that contain a significant number of missing values, exceeding a specified threshold.

	features	include
0	loc_og_t2o_mou	True
1	std_og_t2o_mou	True
2	loc_ic_t2o_mou	True
3	arpu_6	True
4	arpu_7	True
...	...	...
207	aon	True
208	aug_vbc_3g	True
209	jul_vbc_3g	True
210	jun_vbc_3g	True
211	sep_vbc_3g	True

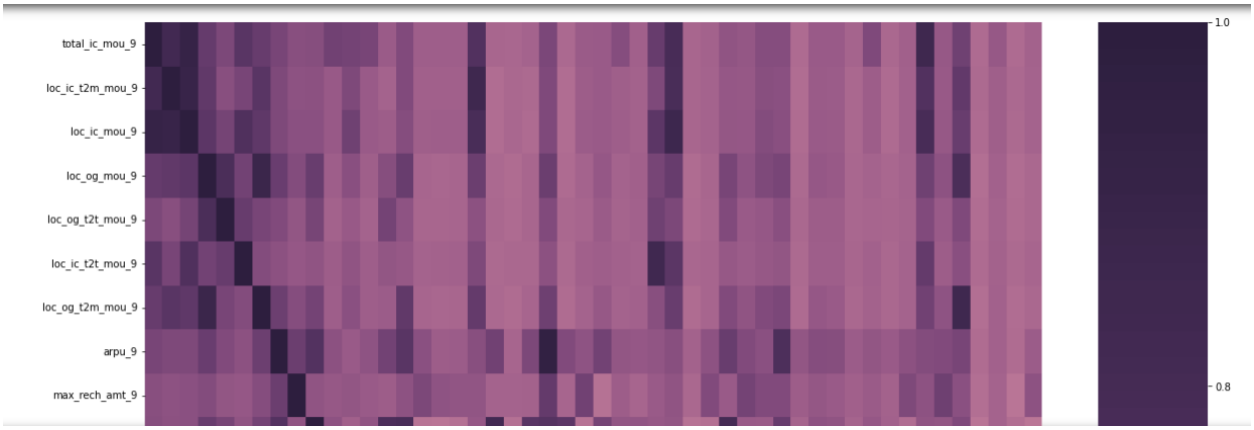
Examine the total recharge amount within the dataset through thorough analysis. This involves scrutinizing the sum of all recharge amounts across the data, providing insights into overall spending patterns or trends related to recharging among customers.



Create a plot showcasing the variance of features within the dataset. This plot will display how much each feature's values deviate from the mean, giving an indication of the spread or dispersion of data across different variables.



Generate a plot illustrating the correlation between different features in the dataset. This plot will visually depict the strength and direction of the relationships between pairs of variables, providing insights into potential patterns or associations within the data.



Conclusion:

This study introduces survival analysis as a potent statistical method for forecasting customer churn. By sorting customers based on their predicted survival probabilities, significant proportions of churners can be captured within the top deciles, with the top two deciles accounting for 45-60% and the top five deciles capturing nearly 85-90% of churners. These insights aid telecommunications companies in comprehending the risk and timing of customer churn throughout their tenure. Ultimately, the study facilitates the customization of marketing communications and customer treatment programs, enabling companies to strategically time their marketing interventions.