Using Survival Analysis To Predict Customer Churn

Daisy Shi

4/26/2021

Introduction

Customer churn occurs when customers or subscribers stop doing business with a company or service, also known as customer attrition. It is also referred as loss of clients or customers. One industry in which churn rates are particularly useful is the telecommunications industry, because most customers have multiple options from which to choose within a geographic location.This paper attempts to analyze the influence of various factors on customer churn through the research on the historical data of a certain operator's IBM customers, and further explores the correlation between these factors and customer churn through survival analysis model. At the same time, using description analysis and model analysis, this paper studies the influence of different factors on customer churn, grasps the situation of customer churn, and puts forward targeted suggestions on how to effectively prevent customer churn of IBM.

Data Overview

The data was downloaded from IBM Sample Data Sets for customer retention programs. The goal of this project is to predict behaviors of churn or not churn to help retain customers. Each row represents a customer, each column contains a customer's attribute.

Customers who left within the last month – the column is called Churn Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies Customer account information – how long they've been a customer, contract, payment method, paperless

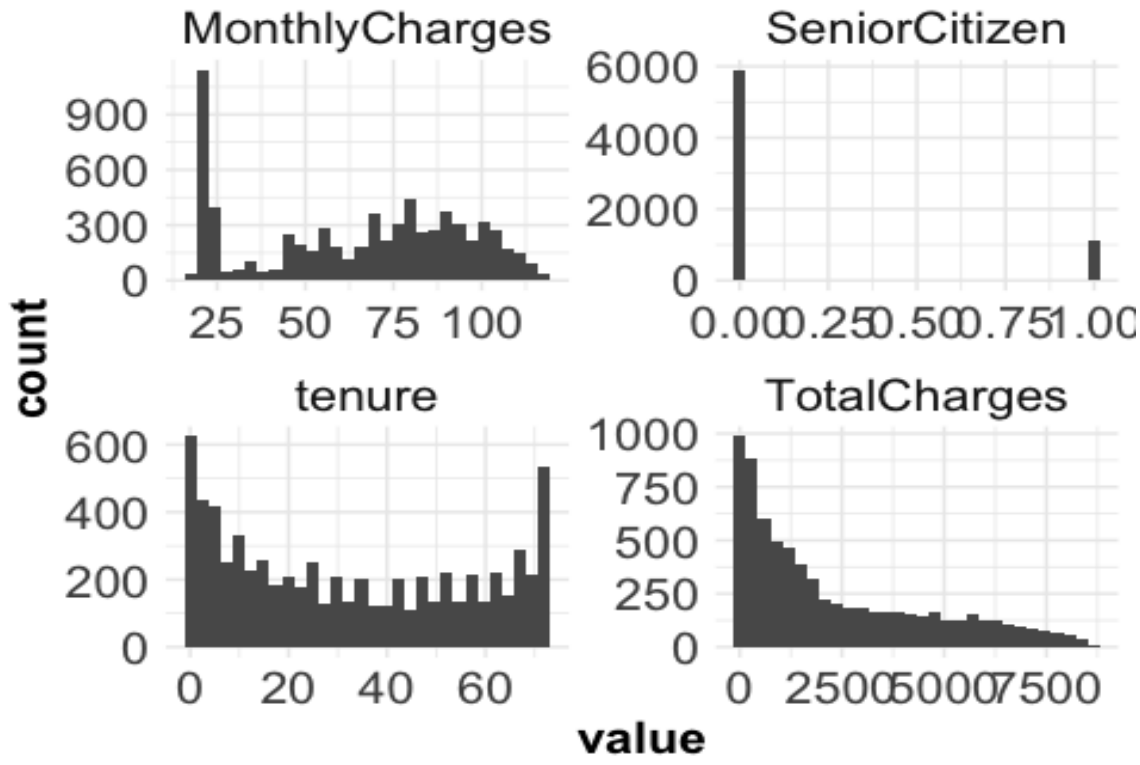
billing, monthly charges, and total charges Demographic info about customers – gender, age range, and if they have partners and dependents.

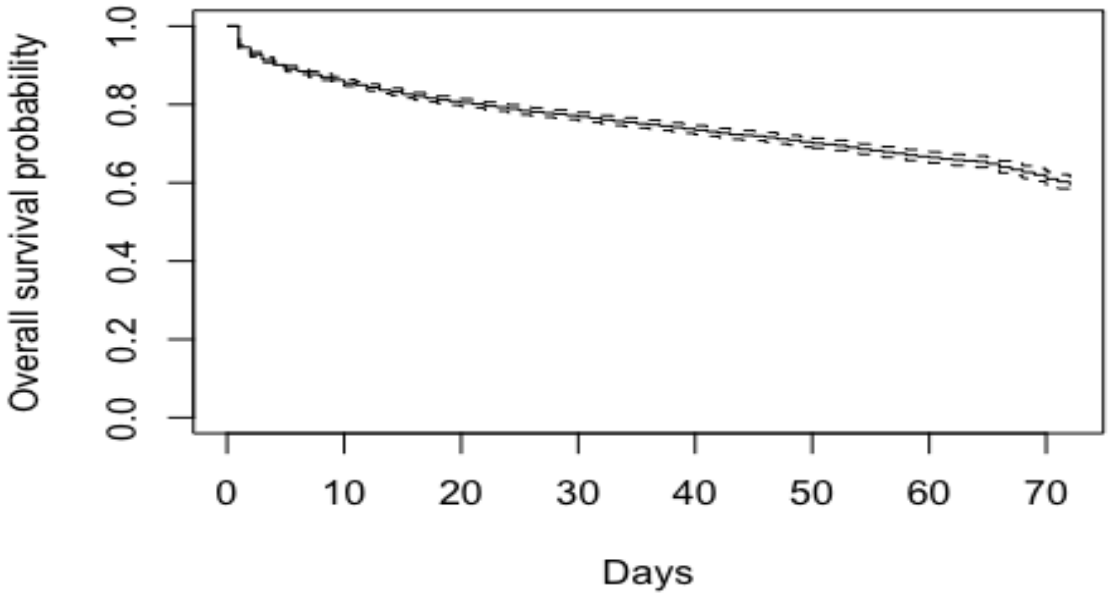
The raw data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target. TotalCharges has 0.156% missing value in the dataset. Other variables are listed as follows: customerID gender (female, male) SeniorCitizen (Whether the customer is a senior citizen or not (1, 0)) Partner (Whether the customer has a partner or not (Yes, No)) Dependents (Whether the customer has dependents or not (Yes, No)) tenure (Number of months the customer has stayed with the company) PhoneService (Whether the customer has a phone service or not (Yes, No)) MultipleLines (Whether the customer has multiple lines r not (Yes, No, No phone service) InternetService (Customer's internet service provider (DSL, Fiber optic, No) OnlineSecurity (Whether the customer has online security or not (Yes, No, No internet service) OnlineBackup (Whether the customer has online backup or not (Yes, No, No internet service) DeviceProtection (Whether the customer has device protection or not (Yes, No, No internet service) TechSupport (Whether the customer has tech support or not (Yes, No, No internet service) streamingTV (Whether the customer has streaming TV or not (Yes, No, No internet service) streamingMovies (Whether the customer has streaming movies or not (Yes, No, No internet service) Contract (The contract term of the customer (Month-to-month, One year, Two year) PaperlessBilling (Whether the customer has paperless billing or not (Yes, No)) PaymentMethod (The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))) MonthlyCharges (The amount charged to the customer monthly — numeric) TotalCharges (The total amount charged to the customer — numeric) Churn (Whether the customer churned or not (Yes or No))

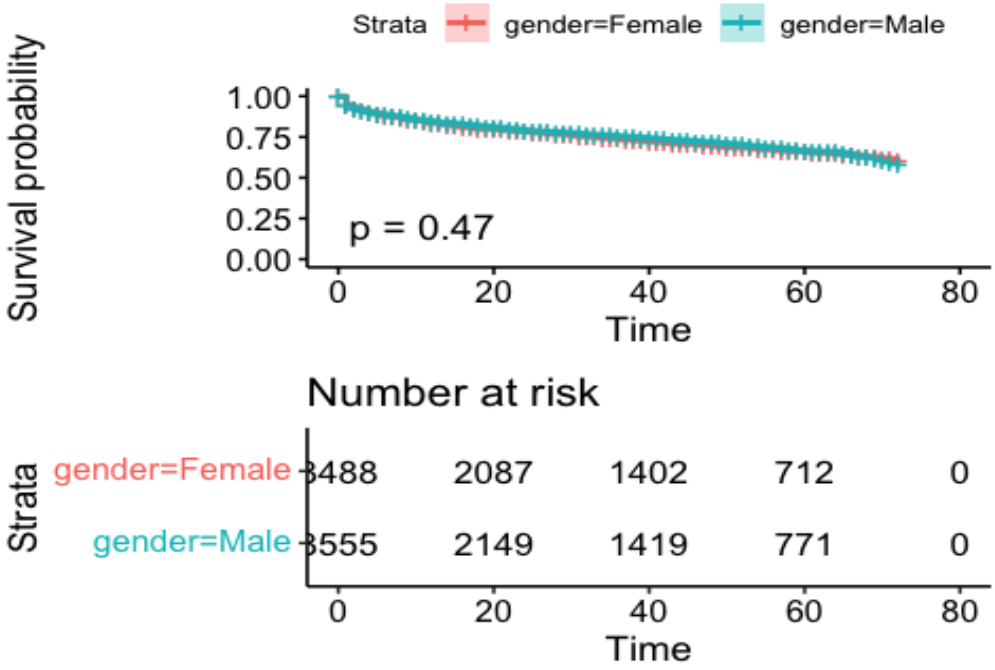
Methods and data diagnostics

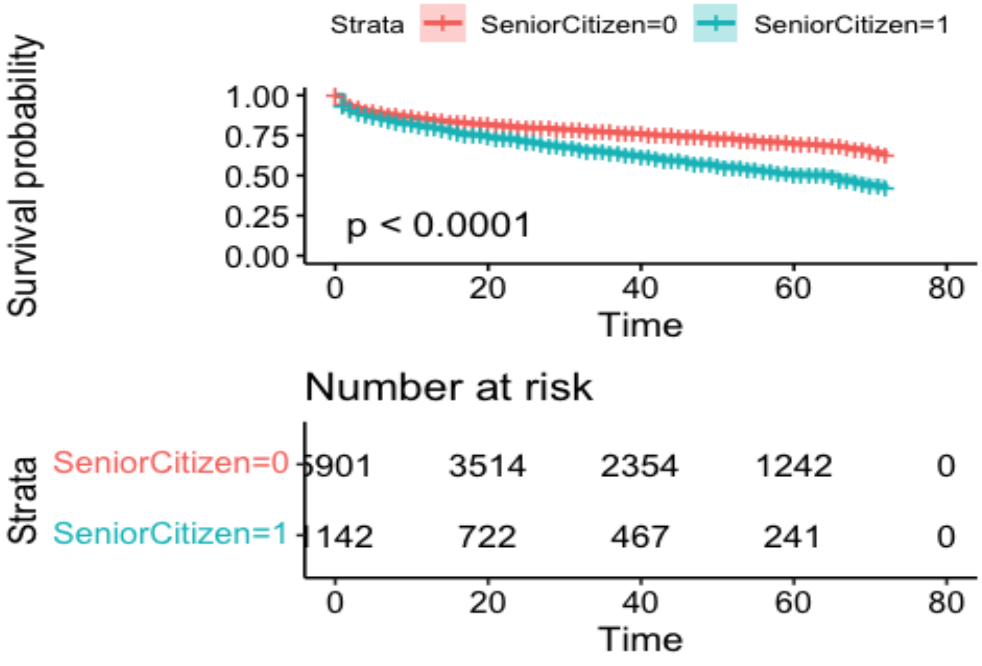
The question I want to study is what are the key factors that lead to the loss of customers. It turns out that retaining old customers is more important than developing and discovering new ones. According to the research results of Kotler, P. (1994), the cost of developing a new customer is 5 times that of retaining an old customer, and the interest rate brought by

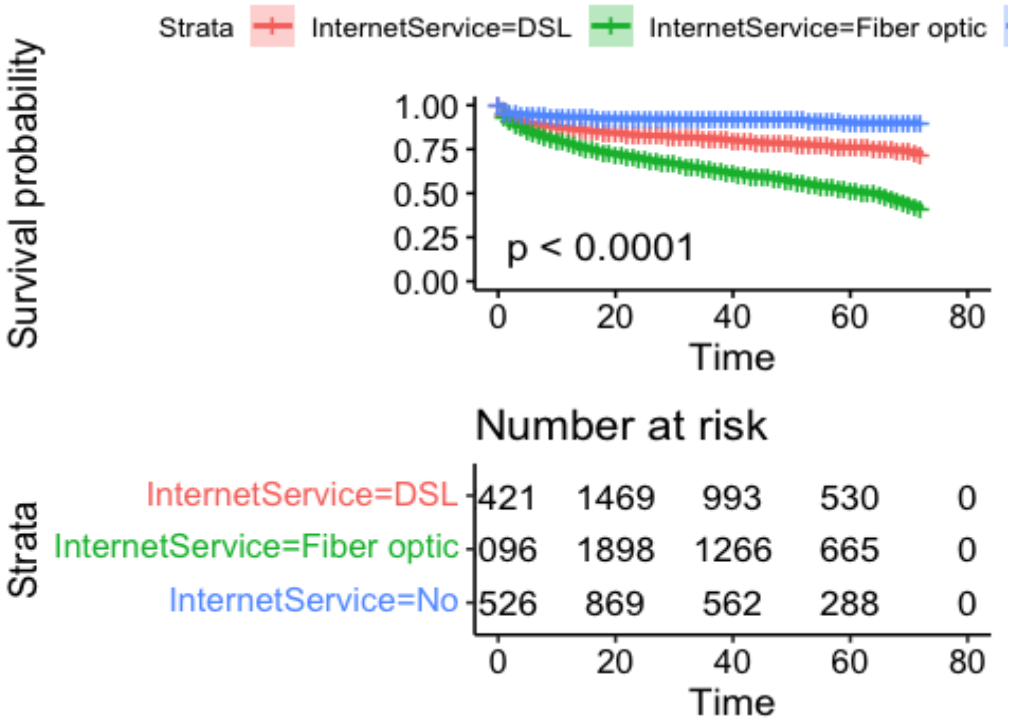
retaining an old customer is 16 times that of developing a new customer. Therefore, reducing customer loss is equal to reducing profit loss with less cost. From this, we know that based on the maintenance and management of customer relationship, how to retain customers is very important to the survival and development of the company. A good operator must know how to retain its customers. So how to effectively predict the potential loss of customers and improve the efficiency of customer retention has become our concern. At present, researches on customer churn prediction are very extensive, such as decision tree algorithm, which is the most widely used. Decision tree algorithm classification accuracy is higher, the modeling also relatively simple, high classification accuracy, and it can export more straightforward classification rules, but it also has some shortcomings, in addition, many experts and scholars put forward using Logistic regression and other research method and artificial neural network model, however, due to the particularity of the customer churn data set itself, The use of these methods is worth discussing. In view of this, according to the basic concept of survival analysis method, this paper establishes a survival analysis model of customer churn scenario, and makes a specific analysis by using historical relevant data to construct a customer churn model. I think survival analysis will be effective for guiding enterprises to manage customer churn.By establishing Cox model and Kaplan-Meier analysis of the relationship between various factors and customer loss, I will try to find out the main reasons for customer loss, and give solutions and suggestions for these problems.Visualization techniques are also used to present and analyze key information. Likelihood ratio test is preformed to check the model comparison.

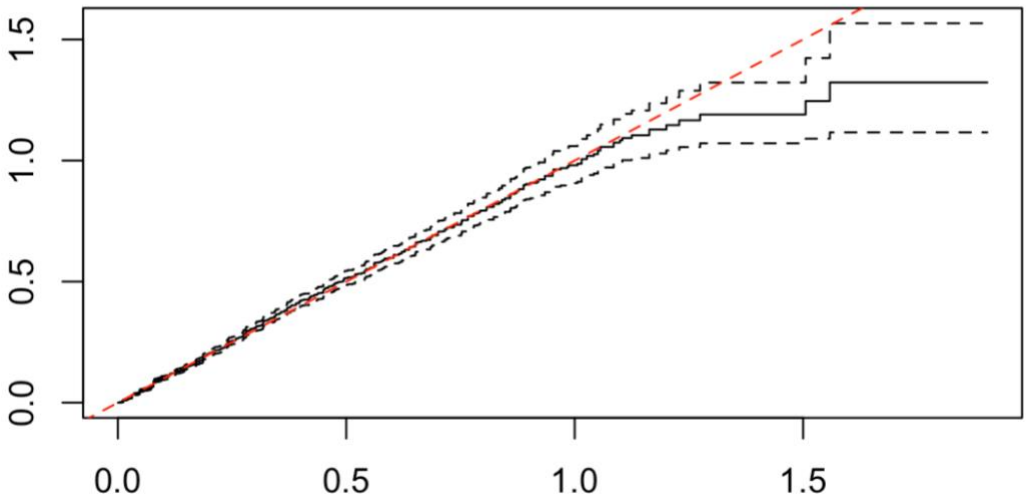

MonthlyCharges
SeniorCitizen
tenure
TotalCharges
count
value
900
600
300
0
6000
4000
2000
0
600
400
200
0
1000
750
500
250
0
25 50 75 100
0.000.250.500.751.00
0 20 40 60
0 2500500075000


Overall survival probability
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Days


Strata
gender=Female
gender=Male
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.47
0
20
40
60
80
Time
Number at risk
Strata
gender=Female
gender=Male
3488
2087
1402
712
0
3555
2149
1419
771
0
0
20
40
60
80
Time


Strata
SeniorCitizen=0
SeniorCitizen=1
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
20
40
60
80
Time
Number at risk
Strata
SeniorCitizen=0
SeniorCitizen=1
5901
3514
2354
1242
0
1142
722
467
241
0
0
20
40
60
80
Time

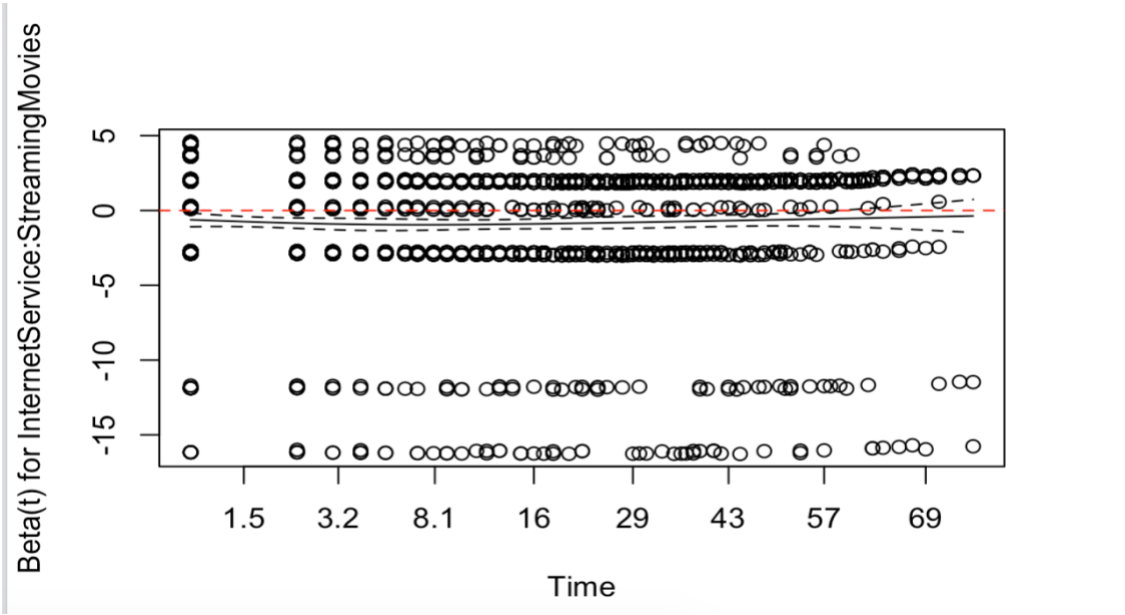

Strata
InternetService=DSL
InternetService=Fiber optic
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
20
40
60
80
Time
Number at risk
Strata
InternetService=DSL
InternetService=Fiber optic
InternetService=No
421
1469
993
530
0
096
1898
1266
665
0
526
869
562
288
0
0
20
40
60
80
Time


0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5

```
coxph(formula = Surv(tenure, is_churn) ~ Partner + PhoneService +
    InternetService * StreamingMovies + Contract + PaymentMethod,
    data = telco_df)

  n= 7043, number of events= 1869
```

	coef
PartnerYes	-0.62748
PhoneServiceYes	-0.16445
InternetServiceFiber optic	0.41625
InternetServiceNo	-0.35843
StreamingMoviesNo internet service	NA
StreamingMoviesYes	-0.25047
ContractOne year	-1.89399
ContractTwo year	-3.67663
PaymentMethodCredit card (automatic)	-0.06357
PaymentMethodElectronic check	0.66942
PaymentMethodMailed check	0.64931
InternetServiceFiber optic:StreamingMoviesNo internet service	NA
InternetServiceNo:StreamingMoviesNo internet service	NA
InternetServiceFiber optic:StreamingMoviesYes	0.01901
InternetServiceNo:StreamingMoviesYes	NA



Main Results for Survival Analysis

By estimating survival curves with Kaplan-Meier Method we find 60 months probability 0f survival is 66.4%, 95% confidence interval between [65%-68%]. Ho: Lifetime distributions under Senior Citizen and Not Senior Citizen are the same. Ha: Lifetime distributions under Senior Citizen and Not Senior Citizen are not the same. Kaplan-Meier test Indicate Senior Citizen Has Higher Churn Probability. Ho: Lifetime distributions under Payment Method are the same. Ha: Lifetime distributions Payment Method are not the same. Kaplan-Meier test Indicate Payment Method with Electronic Check Has Higher Churn Probability. Ho: Lifetime distributions under Internet Service are the same. Ha: Lifetime distributions Internet Service are not the same. Kaplan-Meier test Indicate Internet Service for Fiber Has Higher Churn Probability. Ho: Lifetime distributions under Contract Type are the same. Ha: Lifetime distributions Contract Type are not the same. Kaplan-Meier test Indicate Month to Month Contract Has Higher Churn Probability. Ho: Lifetime distributions under Gender are the same. Ha: Lifetime distributions Gender are not the same. Kaplan-Meier test Indicate Gender Is Not Significant and There Is No Difference Between Gender. And the Log-Rank Test Comparing Survival Times Between Groups, we identified the streaming tv and contract type are also significant variables.

Cox Proportional-Hazard (PH): Likelihood ratio test indicate variables are significant. Using Likelihood ratio to compare different models, and based on the likelihood test result, Likelihood ratio test indicate the p-value is significant, the following variables are selected. “Partner+PhoneService+InternetService*StreamingMovie+Contract+PaymentMethod”. We notice that Interaction Effect Appears for Steaming Movie, Steaming TV And Internet Service. Also, Proportional Hazards and Schoenfeld Residual Indicated Violation of All Variables. H0: the effect of the j-th explanatory variable is constant over time (i.e. proportional hazards) Ha: the effect is not constant over time (i.e. non-proportional hazard). Based on the significant p-value, we can reject the null hypothesis and states that variables violated and has the effect is not constant over time except variable Phone Service in the model.

Conclusion

Churn is indeed high, we note that, holding other variables constant, the average churn risk of telephone service with Fiber users is 1.52 times that of telephone service without Fiber users. Staying with the firm for 60 weeks, the survival probability for senior citizen is 0.5091 and non-senior citizen is 7.01e-01. The average churn risk of Payment Method with Electronic check users is 1.95 times that of non-Electronic check users. Main driver is weather the client is Partner or not, and if they have Internet Service, also Payment Method they choose as well as if they are Senior Citizen. To reduce customer churn probability, I would suggest Increase retention by tying long-term contracts with discount rates for customers who choose to use Internet fiber services and those who choose to Electronic Check.

Works Cited

Kotler, P. (1994). Reconceptualizing marketing: an interview with Philip Kotler. European Management Journal, 12(4), 353-361.