Evaluating factors affecting customer churn at Telco, a fictitious company

Fouad Yared
Hunter College, Statistics 791 Independent Study, Fall 2020

Purpose

One of the many ways the economy has changed in recent years is through the rise of subscription-based businesses. According to Stanford Business, "In recent years, this market has grown by more than 100% a year, increasing from $57 million in sales in 2011 to $2.6 billion in 2016, according to one expert." A key metric for subscription-based companies is churn, or how many of its customers decide to leave a service. The purpose of this analysis is to evaluate factors that affect churn at Telco, a fictious company that provides home phone, internet, and internet-based services for its ~7,000 customers in California. (IBM provided the data as part of its Cognos Analytics platform.[3] It is publicly available on Kaggle.[1])

The Telco data includes several features on customers: their demographics and location, the services they are subscribed to, and their status with the company. Survival analysis, which looks at the time until an event occurs, will be used to better understand the relationship of various predictors against the tenure (in months) a customer has with the company. The event is Churn Value, a binary variable that looks at whether a customer decided to continue or discontinue their business with Telco that quarter. The outcome variable, Tenure, reflects the number of months a person has been a customer with the company. Other variables will serve as predictors. The data set consists of 7,043 customers.

Data Exploration

Bar plots and histograms were used to look at various aspects of the Telco data, as shown in the set of charts in the appendix. Most customers (73%) stayed with the company in the third quarter. Regarding tenure, most customers have been with the company for over twelve months, over 600 signed on in the previous month, and eleven signed on in the current month. Most customers are younger than 65, customers who have and do not have partners are evenly split, and the majority of customers have dependents. Customers tend to be neutral, satisfied, or very satisfied (scores 3 to 5) with the company and tend to sign month-to-month contracts. Phone service and fiber-optic internet service are common among most customers. The total charges for customers (approximately, the monthly charge multiplied by the tenure) are usually under $2,000 and the predicted customer lifetime value ranges between $4-$8,000. Although not shown in the graphs, internet-services (online security, streaming movies) were related to whether someone had purchased internet through the company.

Relationships between Variables

Numeric and categorical variables were evaluated separately to better understand their relationship with each other along with Churn Value and Tenure. Referencing the correlation plot found in the appendix, Churn Value had a high correlation with Churn Score (which predicts whether a customer will leave based on other covariates and gives a score from 0-100, where the higher the score the more likely the customer will churn). Total Revenue (lifetime revenue earned from a customer), Total Charges Present (lifetime amount charged to a customer), Monthly Charges, and Tenure all had a high correlation with one another.

To evaluate the relationship among categorical variables, mosaic plots and Cramer's V statistics were used. The mosaic plots compare counts of one categorical variable against counts that a customer will churn. When looking at satisfaction scores, we see that 100% of the

customers who said they were Very Unsatisfied or were Unsatisfied with the company (a “1” or “2”) left the company. Conversely, 100% of the customers who were Satisfied or Very Satisfied (a “4” or “5”) stayed with the company. The majority who scored the company as Average (a “3”) stayed. When looking at contract types, of those who have a month-to-month contract, 42% left the company compared to 7% of those with a one-year and 3% of those with a two-year contract. The rate of churn for certain marketing offers that customers accepted varied. Over 52% of those who accepted Offer E ended up leaving the company in the current quarter while only 6% of those who accepted Offer A left this quarter.

The bar plot shows the relationship of Churn Status and other binary nominal variables using Cramer’s V. Values for Cramer’s V range from 0 (where there is no association) to 1 (where one variable is completely determined by the other). Notably, Churn Reason is only provided for customers who left which is why there is a very strong relationship with Churn Status. Other variables have milder relationships and will be retained for modeling. Although not pictured, the Cramer’s V statistics were calculated for phone, internet, and internet-related variables. Many Internet-related variables (e.g., Online Security, Streaming Movies) only had values when the customer was provided internet from the company. Otherwise, the values for the internet-related variables were completely determined (as the Cramer’s V statistic was 1) when internet was not provided by the company. All internet-related variables were not included in modeling.

Evaluating Kaplan-Meier Survival Plots

Three KM Survival plots were evaluated using satisfaction scores, contract length, and offer accepted. Log rank tests were used to evaluate whether the curves were the same (the null hypothesis) or whether they were different. Since the p-values for all three KM plots were under 0.05, we conclude the survival curves in each of the plots are different from one another.

Variable Selection

Variables related to customer tenure or churn status were not included in the model. These include Total Charges, Total Revenue, Churn Score, and Customer Lifetime Value, among others. Instead of using tenure-related features like Total Charges when developing Cox models, Monthly Charges were used. Since Monthly Charges were retained in the model, individual line-items (e.g., Phone Service, Internet Service, Average Monthly Long Distance Charges, Streaming Music) were not included. As described in Table 1 below, twelve predictors were used in the model (rows 13 and 14 are the Event and Response variables, respectively).

Conducting backwards variable selection

Backwards variable selection was conducted using the selectCox() function in R. AIC was used as a determining factor. Five of the twelve predictors included in the model were removed. These are the four demographic predictors (Senior Citizen, Partner, Under 30, and Gender) and whether the customer has Paperless Billing. Stand-alone Cox models were fit using the remaining seven variables.

Table 1: Variables included in initial Cox model				
#	Type	Variable Name	Description	Variable Type
1	Pred	Senior Citizen	Indicates whether customer is 65 or older	Binary (Yes/No)
2	Pred	Partner	Indicates whether customer is married	Binary (Yes/No)
3	Pred	Under 30	Indicates whether someone in home is under 30	Binary (Yes/No)
4	Pred	Gender	Indicates customer's gender	Binary (Male/Female)
5	Pred	Contract	Indicates customer's current contract type	Categorical (Month-to-month, one year, two year)
6	Pred	Paperless Billing	Indicates if customer has chosen paperless billing	Binary (Yes/No)
7	Pred	Payment Method	Indicates how customer pays their bills	Categorical (automatic bank transfer, automatic credit card, mailed check, electronic check)
8	Pred	Number of Dependents Group	Indicates number of dependents that live with customer. Dependents can be children, parents, grandparents, etc	Numeric (0-10)
9	Pred	Number of Referrals Group	Indicates number of referrals customer has made	Numeric (0-4)
10	Pred	Offer	Identifies the last marketing offer the customer accepted, if applicable	Categorical (None, Offer A, B, C, D, and E)
11	Pred	Monthly Charges	Indicates customer's current total monthly charge for all services from the company	Numeric (18.25-118.75)
12	Pred	Satisfaction Score Group	A customer's overall satisfaction with the company, from 1 (very unsatisfied) to 5 (very satisfied)	Categorical (scores of 1 and 2 are grouped together; scores of 3-5 are grouped together)
13	Event	Churn Value	Whether the customer left the company this quarter or not	Binary ("1" means customer left company this quarter, "0" means customer stayed this quarter)
14	Resp	Tenure in Months	Total number of months the customer has been with the company	Numeric (0-72)

Table 2: Intercept values in initial Cox model			In post-variable selection Cox Model?
1	Senior Citizen	Not a Senior Citizen	N
2	Partner	Does not have a partner	N
3	Under 30	No one in home is under 30	N
4	Gender	Customer is female	N
5	Contract	Contract is for two years	Y
6	Paperless Billing	Does not have paperless billing	N
7	Payment Method	Uses automatic credit card for payment	Y
8	Number of Dependents Group	Numeric; no default value	Y
9	Number of Referrals Group	Numeric; no default value	Y
10	Offer	Accepted Offer A	Y
11	Monthly Charges	Numeric; no default value	Y
12	Satisfaction Score Group	Grouped satisfaction score of 3, 4, or 5 (Neutral to Satisfied to Very Satisfied)	Y
13	Churn Value	Not applicable	Event variable
14	Tenure in Months	Not applicable	Response variable

Fitting Cox models

Initially, two Cox models were fit: one was the full twelve variable model and the other was a reduced model with only the seven variables from backwards selection. Performing a log-likelihood test on the full and reduced models led to a p-value of 0.1236. Since the p-value for the log-likelihood test is >= 0.05, we use the reduced model. Table 2 shows the intercept values for variables in the Cox model before and after variable selection.

Interpreting Final Cox Model

Variables and variable-levels (for categorical variables) not included in the intercept can be found in the first column of Table 3 and their associated coefficients are found in the second column. The third column shows the hazard rate, which exponentiates the coefficients for each variable. The fourth and fifth columns show the standard error and z-scores, respectively. The sixth column shows p-values. (While all variables had significant p-values, two levels of categorical variables were not significant.) The seventh and eighth columns show the 2.5% and 97.5% confidence intervals for the hazard rate. Highlighted values consist of very high hazard rates, non-significant p-values greater than 0.05, and confidence intervals that include the baseline hazard ratio of 1 (which are related to p-values greater than 0.05). Overall, the length of a customer's contract, the marketing offer they accepted, and the customer's overall satisfaction play key roles in determining how long a customer stays with the company. The magnitude of each of these effects, when controlling for the effects of other variables, is described below.

The length of a customer's contract has the single largest effect on the hazard rate. Customers with a month-to-month contract have a hazard rate that is 16.27 times higher than those with a two-year contract. Similarly, customers with a one-year contract have a hazard rate that is 4.60 times higher than those with a two-year contract. In other words, the longer a customer's contract, the longer their tenure with the company.

The last marketing offer a customer accepted had the second largest effect on how long a customer stayed with the company. The base rate is for customers that accepted Offer A. The hazard rate for customers who accepted Offer E was 14.37 times as large as it was for those who accepted Offer A; the hazard rate for customers who accepted Offers C, D, and no offer were 2.19, 3.21, and 3.06 times as large as the hazard rate for those who accepted Offer A, respectively. (The hazard rate for those that accepted Offer B is not significantly different from the hazard rate for those who accepted Offer A, so the results will not be interpreted.) Details on the various offers were not included in IBM's data description, but it is likely that the terms of certain offers may favor shorter or longer tenures with the company.

The overall satisfaction of a customer also played a key role in determining how long a customer decided to stay with the company. Customers who stated they were unsatisfied or very unsatisfied had a hazard ratio of 9.88 times that of customers who said they were neutral, satisfied, or very satisfied with the company. Improving customer's satisfaction (or perception) would help retain customers.

We expect the hazard rate for customers that use an electronic check to increase by 54% and for those who use a mailed check to increase by 83%, compared to those who use an

automatic credit card. (The hazard rate for those who used an automatic bank transfer is not significantly different from the hazard rate for those who used an automatic credit card, so the results will not be interpreted.) The more effort a payment method requires (i.e., mailing a check requires more effort than an automatic payment), the more likely the customer will leave the company.

Regarding the numeric variables, a one unit increase in the number of dependents is associated with a 18% decrease in the expected hazard rate. Similarly, a one unit increase in the number of referrals is associated with a 22% decrease in the expected hazard rate. Customers with more dependents and referrals are more likely to stay with the company.

A one unit increase in Monthly Charges is associated with a 0.07% decrease in the expected hazard. To better understand how the amount paid each month affects tenure with the company, the Monthly Charges feature was split into four classes based on quartiles (<=$35, >$35-70, >$70-90, and >$90-120). (Since the Monthly Charges feature was changed, a new Cox model was fit. The set of coefficients for the reduced model is found in Table 4 below. The coefficients of the other variables are close to their previous coefficients.) Customers who pay $90-120 a month have a 35% lower hazard rate than customers who pay less than $35 a month. (The hazard rate for those who pay >$35-70 or >$70-90 a month is not significantly different from the hazard rate for those who pay less than $35 a month, so the results will not be interpreted.) There is not a clear relationship between how much a customer is charged each month and whether they will leave the company. The bar chart labeled “Median monthly charges by Tenure Length and Churn Status” shows there is a tendency for Monthly Charges to increase for each year a customer has been with the company. There is also a tendency for customers who leave to be paying more than customers who do not leave.

Evaluating Proportional Hazard Assumption

The null hypothesis for the Cox Model is that each variable has a constant hazard rate over time. The Proportional Hazard (PH) assumption for each of the variables are tested using the cox.zph() function in R. Values less than 0.05 reject the null hypothesis. Results are found in Table 5. The p-value for the PH assumption is below 0.05 for both the entire model (the Global line-item) and six of the seven variables in the final model. The violations to the PH assumptions are visualized in Chart 16 on Schoenfeld residuals. Time-related patterns are found in several of the plots, which are noticed as the number of months increases (the top-left plot on Beta residuals for the Contract length variable shows a curvilinear relationship with time).

Conclusions

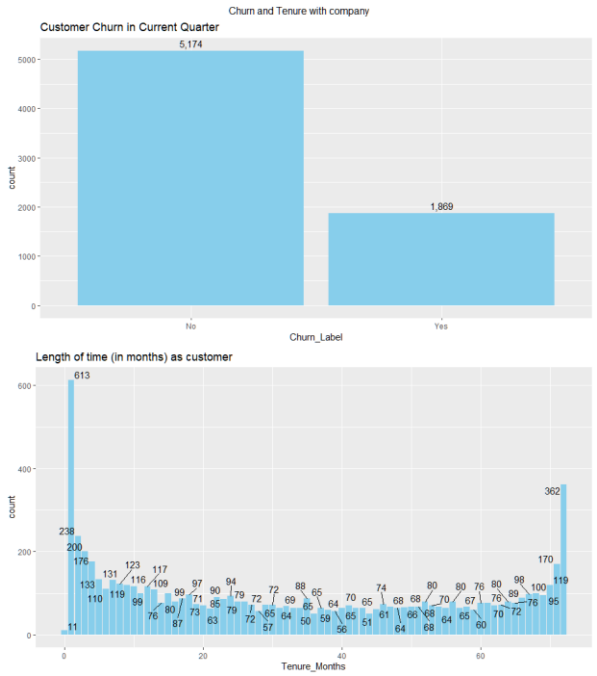
The findings of the exploratory analysis and KM curves were helpful in generating an idea of which variables may be important. Conducting variable selection with a-priori knowledge and backwards selection (using AIC) helped reduce the feature space, so the final Cox model only used seven variables. By interpreting the final model, the largest effects on tenure with the company are the contract length, the type of marketing offer a customer accepted, and the payment method. More work can be done to address violations to PH assumptions.

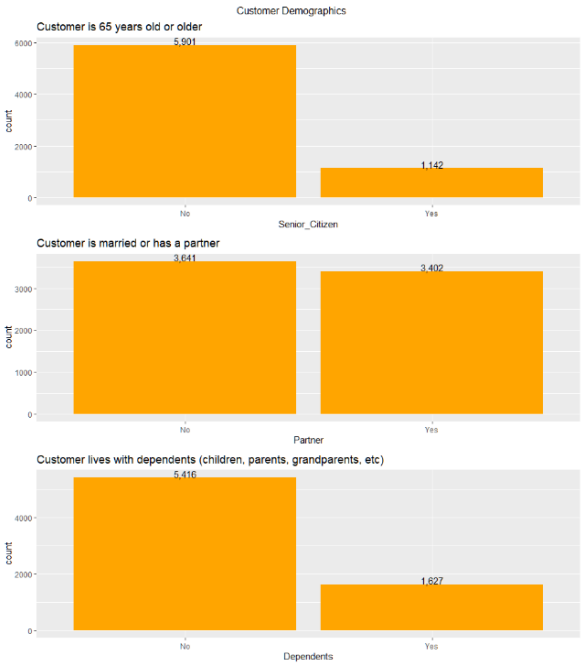
Table 3: Interpreting Seven-variable Cox Model (Monthly Charges are Numeric)							
Variable/Variable-level	coef	Hz	SE	z	Pval	Hz_2.5%	Hz_97.5%
Contract: Month-to-month	2.8	16.27	0.2	16.7	0.00	11.7	22.6
Contract: One year	1.5	4.60	0.2	8.9	0.00	3.3	6.5
Payment Method: Bank transfer (automatic)	0.1	1.11	0.1	1.1	0.27	0.9	1.3
Payment Method: Electronic check	0.4	1.54	0.1	5.8	0.00	1.3	1.8
Payment Method: Mailed check	0.6	1.83	0.1	6.6	0.00	1.5	2.2
Number of Dependents Group	-0.2	0.82	0.0	-4.1	0.00	0.7	0.9
Number of Referrals Group	-0.3	0.78	0.0	-11.6	0.00	0.7	0.8
Offer: None	1.1	3.06	0.2	6.2	0.00	2.1	4.4
Offer: Offer B	0.1	1.15	0.2	0.7	0.49	0.8	1.7
Offer: Offer C	0.8	2.19	0.2	3.7	0.00	1.5	3.3
Offer: Offer D	1.2	3.21	0.2	5.8	0.00	2.2	4.8
Offer: Offer E	2.7	14.37	0.2	13.6	0.00	9.8	21.1
Monthly Charges	0.0	0.99	0.0	-6.4	0.00	1.0	1.0
Satisfaction Score Group: Rated 1 or 2	2.3	9.88	0.1	37.7	0.00	8.8	11.1

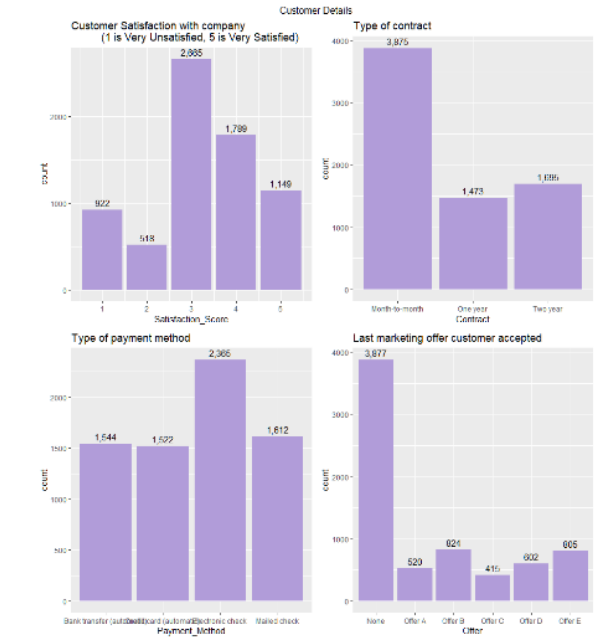
Table 4: Interpreting Seven-variable Cox Model (Monthly Charges split into quartiles)							
Variable/Variable-level	coef	Hz	SE	z	Pval	Hz_2.5%	Hz_97.5%
Contract: Month-to-month	2.8	16.01	0.2	16.6	0.00	11.5	22.2
Contract: One year	1.5	4.61	0.2	8.9	0.00	3.3	6.5
Payment Method: Bank transfer (automatic)	0.1	1.11	0.1	1.1	0.25	0.9	1.3
Payment Method: Electronic check	0.4	1.52	0.1	5.6	0.00	1.3	1.8
Payment Method: Mailed check	0.7	1.94	0.1	7.3	0.00	1.6	2.3
Number of Dependents Group	-0.2	0.83	0.0	-3.9	0.00	0.8	0.9
Number of Referrals Group	-0.2	0.78	0.0	-11.6	0.00	0.7	0.8
Offer: None	1.1	3.12	0.2	6.3	0.00	2.2	4.5
Offer: Offer B	0.2	1.20	0.2	0.9	0.38	0.8	1.8
Offer: Offer C	0.8	2.18	0.2	3.7	0.00	1.4	3.3
Offer: Offer D	1.2	3.22	0.2	5.8	0.00	2.2	4.8
Offer: Offer E	2.7	14.77	0.2	13.7	0.00	10.1	21.7
Monthly Charges Group: >$35-70	0.0	0.98	0.1	-0.2	0.82	0.8	1.2
Monthly Charges Group: >$70-90	0.1	1.06	0.1	0.7	0.48	0.9	1.3
Monthly Charges Group: >$90-120	-0.4	0.66	0.1	-4.4	0.00	0.5	0.8
Satisfaction Score Group: Rated 1 or 2	2.3	9.68	0.1	37.5	0.00	8.6	10.9

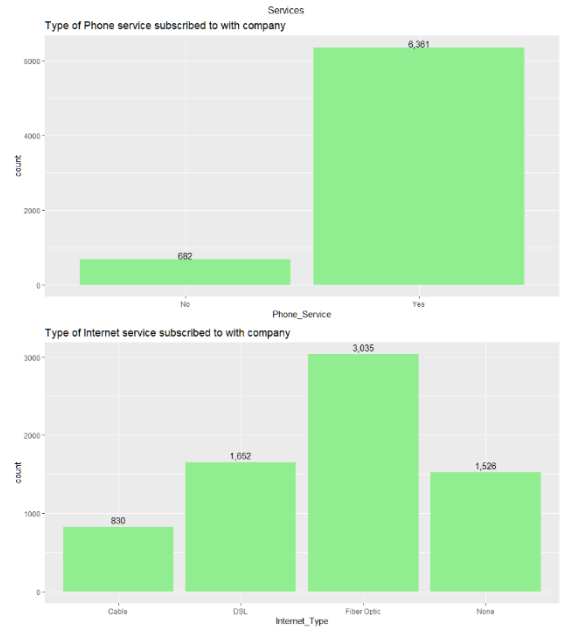
Table 5: Evaluating PH assumptions for Seven-variable Cox Model			
Variable	chisq	df	p
Contract	59.10	2	0.00
Payment Method	16.19	3	0.00
Number of Dependents Group	18.89	1	0.00
Number of Referrals Group	0.50	1	0.48
Offer	667.58	5	0.00
Monthly Charges	24.28	1	0.00
Satisfaction Score Group	57.52	1	0.00
GLOBAL	851.79	14	0.00

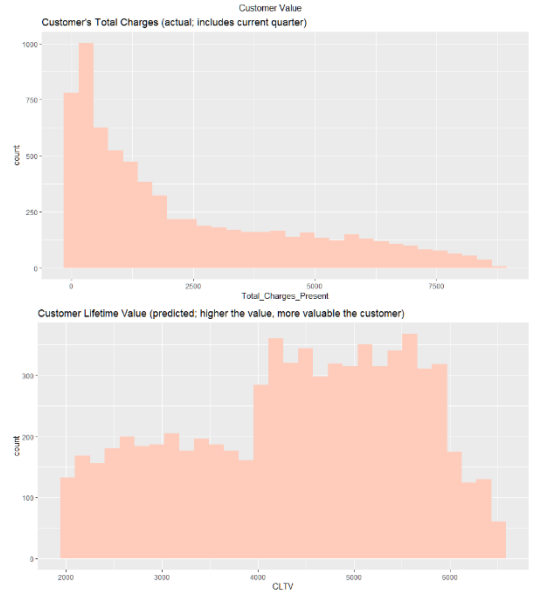
Appendix

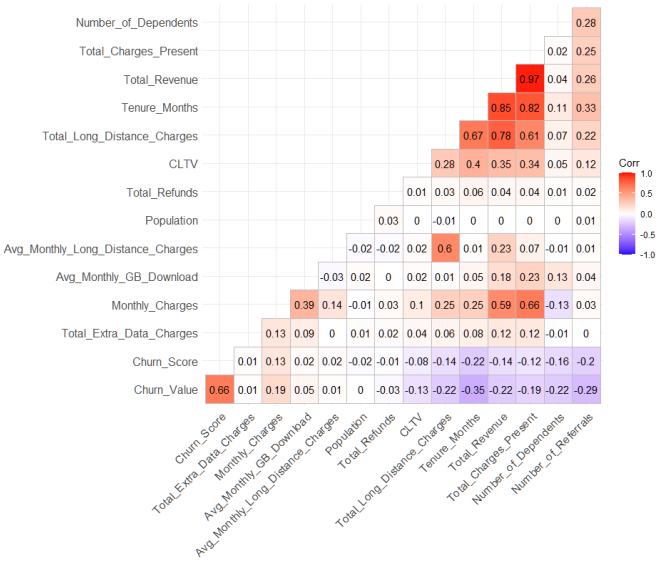

Churn and Tenure with company
Customer Churn in Current Quarter
5,174
1,869
count
No
Yes
Churn_Label
Length of time (in months) as customer
613
362
Tenure_Months

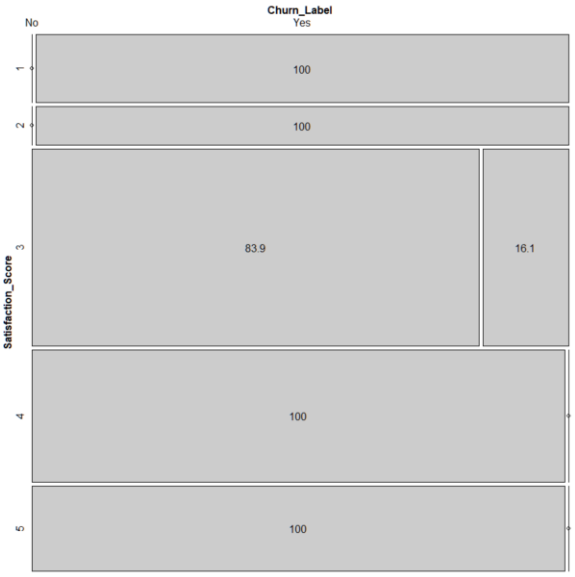

Customer Demographics
Customer is 65 years old or older
5,901
1,142
Senior_Citizen
Customer is married or has a partner
3,641
3,402
Partner
Customer lives with dependents (children, parents, grandparents, etc)
5,416
1,627
Dependents

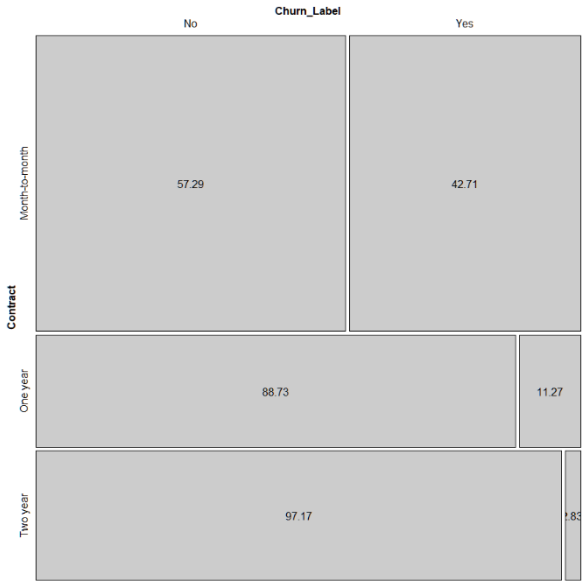

Customer Details
Customer Satisfaction with company
(1 is Very Unsatisfied, 5 is Very Satisfied)
2,885
1,799
1,149
922
518
Satisfaction_Score
Type of contract
3,875
1,473
1,695
Month-to-month
One year
Two year
Contract
Type of payment method
2,365
1,544
1,522
1,612
Payment_Method
Last marketing offer customer accepted
3,877
924
805
602
529
415
None
Offer A
Offer B
Offer C
Offer D
Offer E
Offer

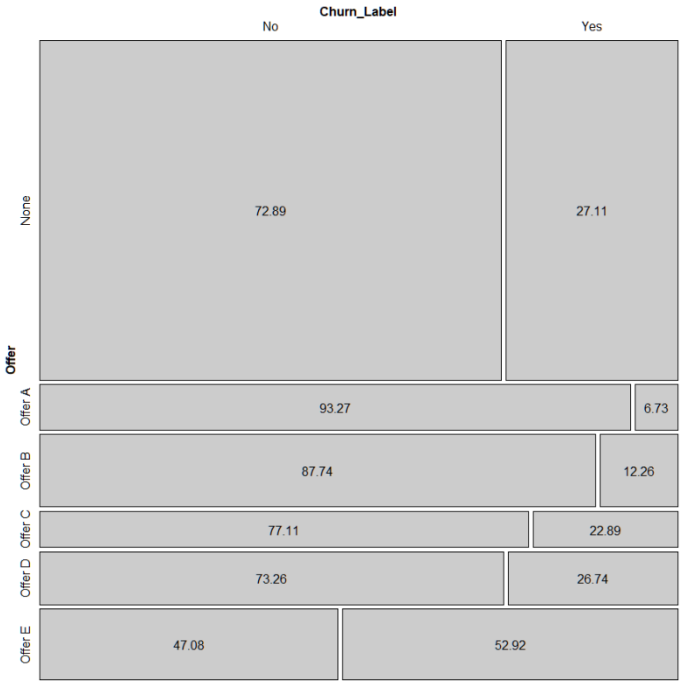

Services
Type of Phone service subscribed to with company
6,361
682
No
Yes
Phone_Service
Type of Internet service subscribed to with company
3,035
1,652
1,526
830
Cable
DSL
Fiber Optic
None
Internet_Type

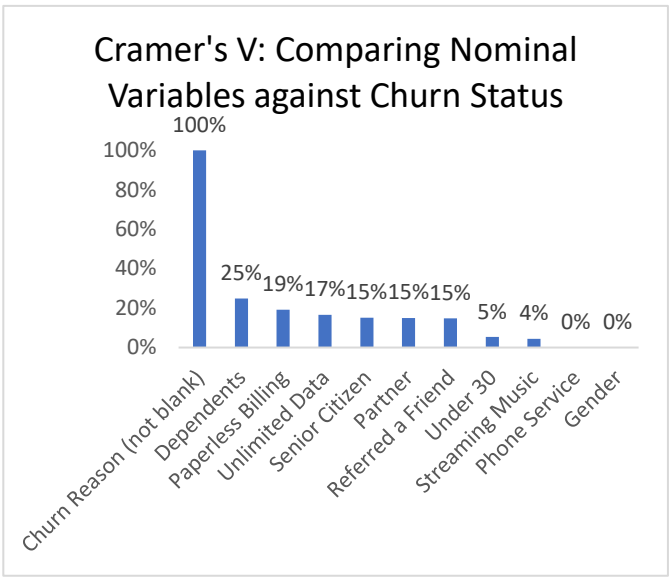

Customer Value
Customer's Total Charges (actual; includes current quarter)
count
Total_Charges_Present
Customer Lifetime Value (predicted; higher the value, more valuable the customer)
count
CLTV

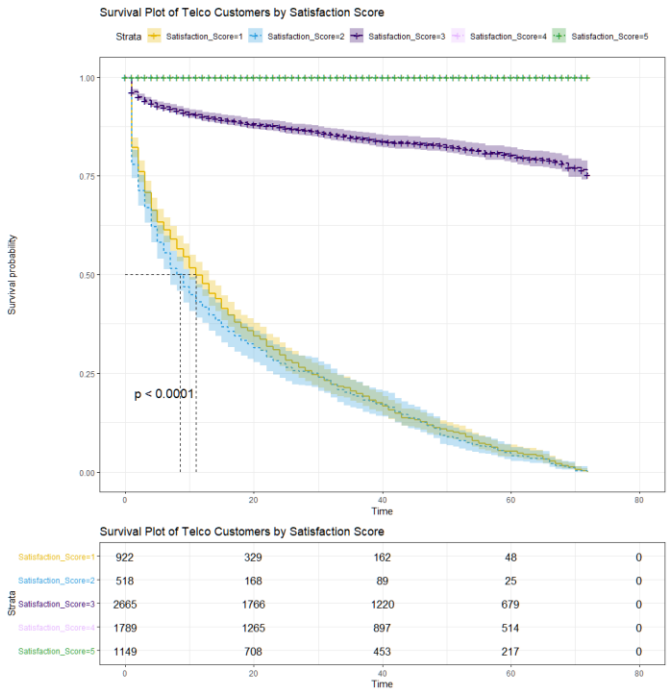

Number_of_Dependents
Total_Charges_Present
Total_Revenue
Tenure_Months
Total_Long_Distance_Charges
CLTV
Total_Refunds
Population
Avg_Monthly_Long_Distance_Charges
Avg_Monthly_GB_Download
Monthly_Charges
Total_Extra_Data_Charges
Churn_Score
Churn_Value
Corr

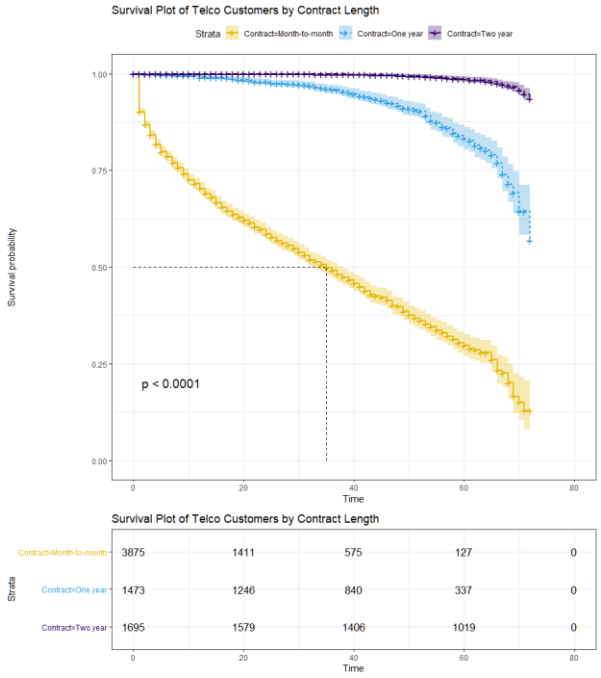

Churn_Label
No
Yes
Satisfaction_Score
100
100
83.9
16.1
100
100

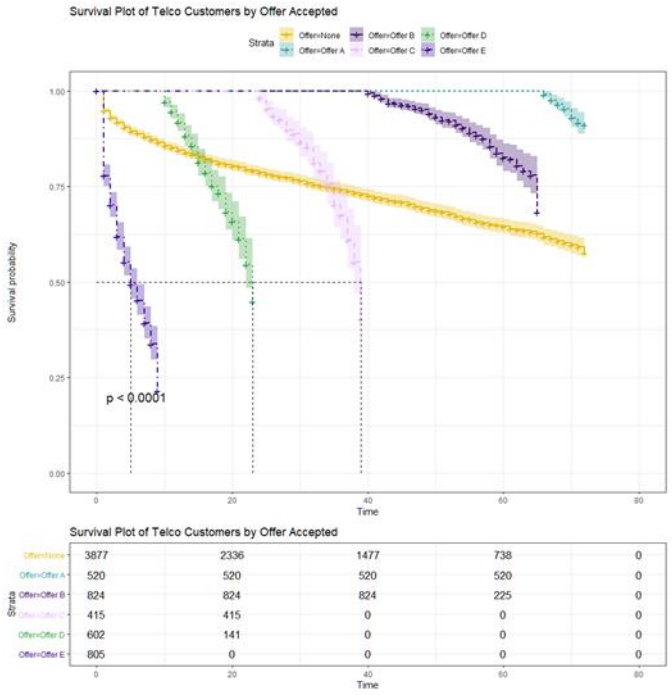

Churn_Label
No
Yes
Contract
Month-to-month
57.29
42.71
One year
88.73
11.27
Two year
97.17


Churn_Label
No
Yes
Offer
None
72.89
27.11
Offer A
93.27
6.73
Offer B
87.74
12.26
Offer C
77.11
22.89
Offer D
73.26
26.74
Offer E
47.08
52.92

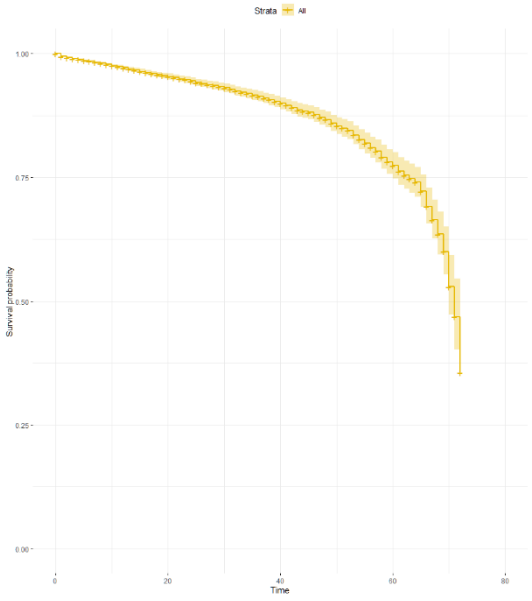

Cramer's V: Comparing Nominal Variables against Churn Status
100%
100%
80%
60%
40%
20%
0%
25%
19%
17%
15%
15%
15%
5%
4%
0%
0%
Churn Reason (not blank)
Dependents
Paperless Billing
Unlimited Data
Senior Citizen
Partner
Referred a Friend
Under 30
Streaming Music
Phone Service
Gender


Survival Plot of Telco Customers by Satisfaction Score
Strata
Satisfaction_Score=1
Satisfaction_Score=2
Satisfaction_Score=3
Satisfaction_Score=4
Satisfaction_Score=5
Survival probability
p < 0.0001
Time
Survival Plot of Telco Customers by Satisfaction Score
Satisfaction_Score=1 922 329 162 48 0
Satisfaction_Score=2 518 168 89 25 0
Satisfaction_Score=3 2665 1766 1220 679 0
Satisfaction_Score=4 1789 1265 897 514 0
Satisfaction_Score=5 1149 708 453 217 0
Strata
Time


Survival Plot of Telco Customers by Contract Length
Strata
Contract=Month-to-month
Contract=One year
Contract=Two year
Survival probability
p < 0.0001
Time
Survival Plot of Telco Customers by Contract Length
Contract=Month-to-month 3875 1411 575 127 0
Contract=One year 1473 1246 840 337 0
Contract=Two year 1695 1579 1406 1019 0
Strata
Time



Seven-variable Cox Model, without strata



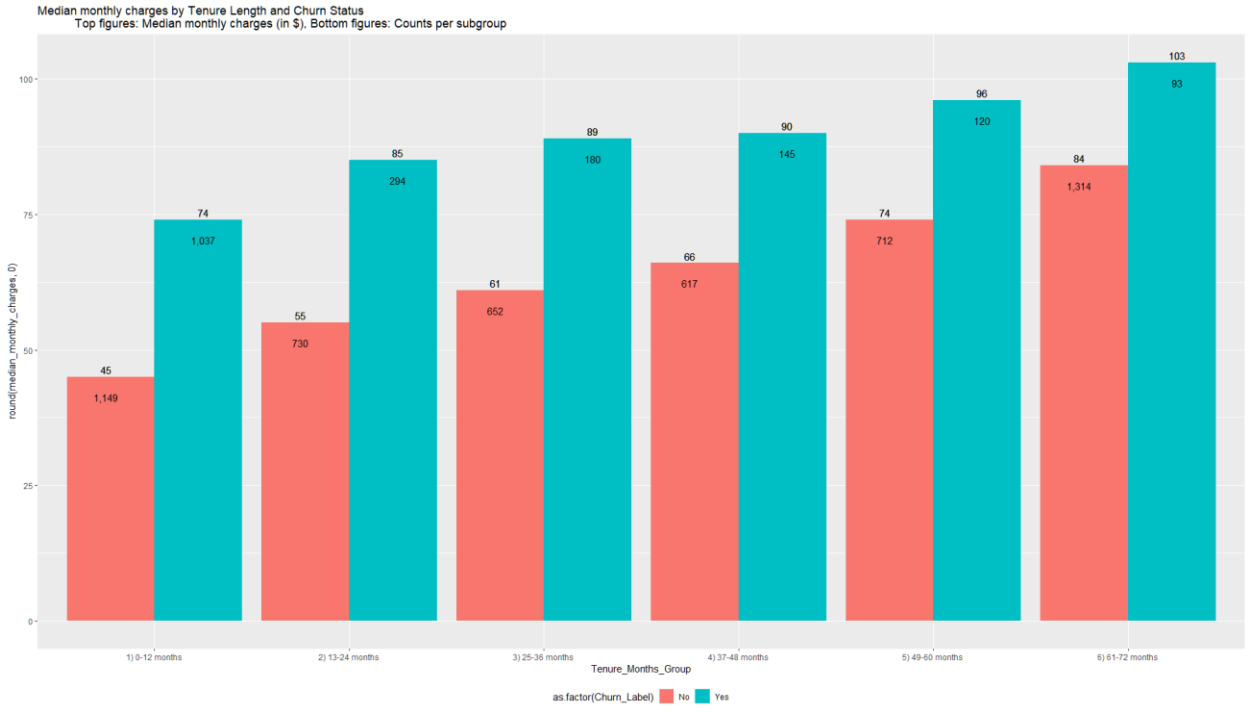
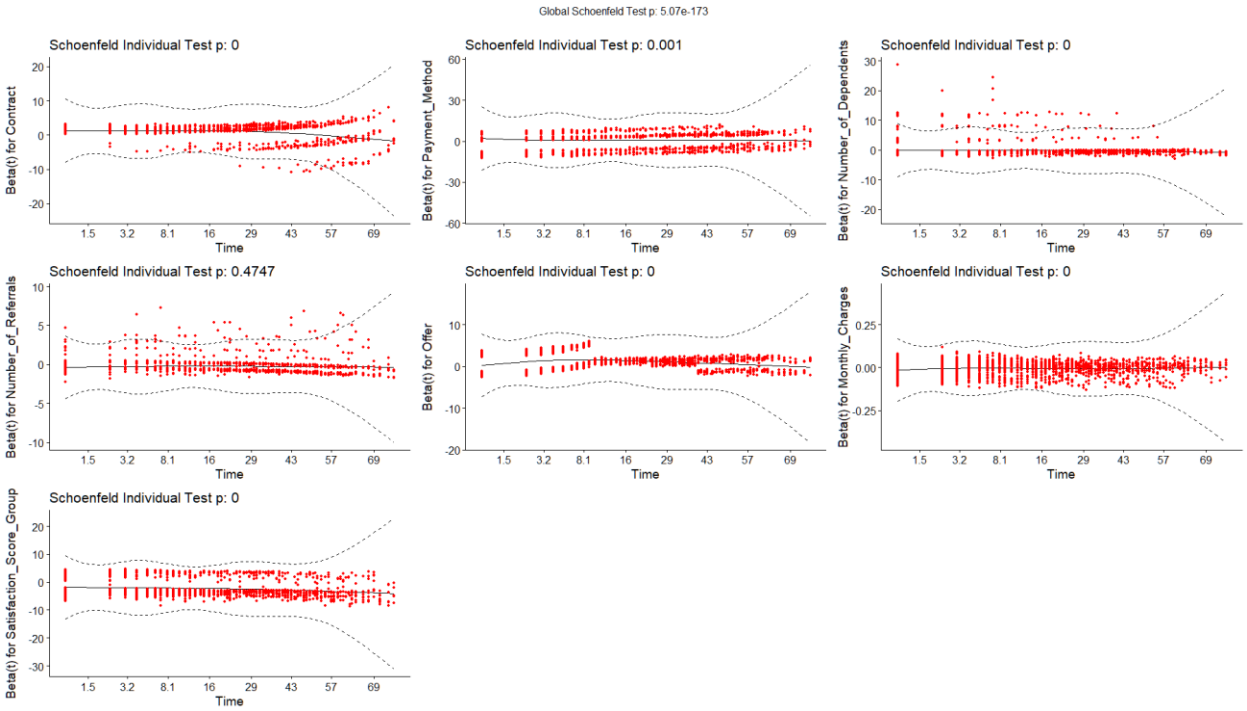


Chart 16: Schoenfeld Residuals of Seven-Variable Model



References

Data Set

1. Telco customer churn (11.1.3+) (2019). IBM Cognos Analytics sample data sets [Data set]. Kaggle. https://www.kaggle.com/ylchang/telco-customer-churn-1113

Other References (Listed Alphabetically)

2. "Cox Model Assumptions." *STHDA (Statistical tools for high-throughput data analysis)*, http://www.sthda.com/english/wiki/cox-model-assumptions
Accessed December 12, 2020

3. IBM Business Solutions. "Telco customer churn (11.1.3+)" *IBM Community*, https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113
Accessed December 12, 2020

4. Kleinbaum, David G. and Klein, Mitchel. *Survival Analysis: A Self-Learning Text, Third Edition*. Springer Science, 2012. https://www.springer.com/gp/book/9781441966452

5. LaMorte, Wayne W. "Cox Proportional Hazards Regression Analysis." *Boston University,* https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival6.html
Accessed December 12, 2020

6. Luna, Jenny. "Why Every Business Will Soon Be a Subscription Business." *Stanford Graduate School of Business*, https://www.gsb.stanford.edu/insights/why-every-business-will-_soon-be-subscription-business
Accessed December 12, 2020

7. Rizopoulos, Dimitris. "EP03: Survival Analysis in R Companion." https://www.drizopoulos.com/courses/emc/ep03_%20survival%20analysis%20in%20r%20companion
Accessed December 12, 2020