Telecommunications Churn Analysis Using Cox Regression

Introduction

As part of its efforts to increase customer loyalty and reduce churn, a telecommunications company is interested in modeling the "time to churn" in order to determine the factors that are associated with customers who are quick to switch to another service. To this end, a random sample of customers is selected and their time spent as customers, whether they are still active customers, and various demographic fields are pulled from the database for use in a **Cox Regression** loyalty analysis.

Analysis

Now let's run the Cox Regression churn model, and see what we can find out about patterns and causes of churn. The dependent or criterion variable in the model (the variable we are trying to predict) is called the status variable. The status variable identifies whether the event (churn) has occurred for a given case. If the event has not occurred, the case is said to be censored. Censored cases are not used in the computation of the regression coefficients, but are used to compute the baseline hazard. The case-processing summary shows that 726 cases are censored. These are customers who have not churned.

Case Processing Summary

		N	Percent
Cases available in analysis	Event(a)	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a Dependent Variable: Months with service

We will be examining the potential influences on churn of several key candidate predictors: age; marital status; education; employment status (retired vs. still working); gender; length of time at current address; length of time with current employer; and customer category. Some candidate predictors that we will test in the churn model are quantitative variables such as age or length of time at current address. Other possible predictors (e.g., marital status) are categorical variables, because they cannot be measured on a quantitative scale. The following categorical variable codings are a useful reference for interpreting the regression coefficients for categorical covariates, particularly dichotomous variables:

Cox Regression Model Categorical Variable Codings(c,d,e,f,g)

		Frequency	(1)(a)	(2)	(3)	(4)
marital(b)	0=Unmarried	505	1			
	1=Married	495	0			
ed(b)	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire(b)	.00=No	953	1			
	1.00=Yes	47	0			
gender(b)	0=Male	483	1			
	1=Female	517	0			
custcat(b)	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	
	3=Plus service	281	0	0	1	
	4=Total service	236	0	0	0	

a The (0,1) variable has been recoded, so its coefficients will not be the same as for indicator (0,1) coding.
b Indicator Parameter Coding
c Category variable: marital (Marital status)
d Category variable: ed (Level of education)
e Category variable: retire (Retired)
f Category variable: gender (Gender)
g Category variable: custcat (Customer category)

In this particular analysis, by default, the reference category is the last category of a categorical covariate. Thus, for example, even though Married customers have variable values of 1 in the data file, they are coded as 0 for the purposes of the regression.

The Cox Regression model-building process takes place in two blocks. In the first, a forward stepwise algorithm is employed. The omnibus tests are measures of how well the model performs. (The chi-square change from previous step is the difference between the -2 log-likelihood of the model at the previous step and the current step.) Here is the summary table of output from the model-generation process, followed by an explanation and discussion:

Omnibus Tests of Model Coefficients(f,g)

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1(a)	3383.793	132.522	1	.000	142.571	1	.000	142.571	1	.000
2(b)	3331.588	161.504	2	.000	52.205	1	.000	194.776	2	.000
3(c)	3295.644	178.903	3	.000	35.943	1	.000	230.720	3	.000
4(d)	3295.688	174.203	2	.000	.044	1	.834	230.676	2	.000
5(e)	3282.533	186.817	3	.000	13.155	1	.000	243.831	3	.000

a Variable(s) Entered at Step Number 1: age
b Variable(s) Entered at Step Number 2: employ
c Variable(s) Entered at Step Number 3: address
d Variable Removed at Step Number 4: age
e Variable(s) Entered at Step Number 5: marital
f Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
g Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

If the step was to add a variable, the inclusion makes sense if the significance of the change is less than 0.05. If the step was to remove a variable, the exclusion makes sense if the significance of the change is greater than 0.10. In the first three steps, AGE, EMPLOY, and ADDRESS are added to the model.

In the fourth step, AGE is removed from the model, likely because the variation in time to churn that is explained by AGE is also explained by EMPLOY and ADDRESS; thus, when these variables are added to the model, AGE is no longer necessary. Finally, MARITAL is added in the fifth step. The final model for Block 1 includes MARITAL, ADDRESS, and EMPLOY.

Here is a table of predictive model coefficients, followed by an explanation and discussion:

Block 1: Method = Forward Stepwise (Likelihood Ratio)
Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 1	age	-.065	.006	124.361	1	.000	.937
Step 2	age	-.032	.007	22.806	1	.000	.969
	employ	-.075	.011	49.296	1	.000	.928
Step 3	age	-.002	.008	.044	1	.835	.998
	address	-.059	.010	35.184	1	.000	.942
	employ	-.080	.011	53.479	1	.000	.923
Step 4	address	-.060	.009	49.638	1	.000	.941
	employ	-.081	.010	71.408	1	.000	.922
Step 5	marital	.442	.122	13.117	1	.000	1.556
	address	-.061	.009	50.409	1	.000	.941
	employ	-.083	.010	73.287	1	.000	.920

The value of Exp(B) for MARITAL means that the churn hazard for an unmarried customer is 1.556 times that of a married customer. (Recall from the categorical variable codings that unmarried = 1 for the regression.) The value of Exp(B) for ADDRESS means that the churn hazard is reduced by 100%-(100%x0.941)=5.9% for each year (on a compounded basis) that a customer has lived at the same address. A more useful computational formula for calculating this involves raising the Exp(B) to a power equal to the number of years at current address. For example, the churn hazard for a customer who has lived at the same address for five years is reduced by 100%-(100%x(0.941^5))=26.2%. [Note that in this formula the ^ symbol represents raising a number to a power.]

Likewise, the value of Exp(B) for EMPLOY means that the churn hazard is reduced by 100%-(100%x0.920)=8.0% for each year (on a compounded basis) that a customer has worked for the same employer. Using the aforementioned alternative computational formula, the churn hazard for a customer who has worked for the same employer for three years is reduced by

100%-(100%x(0.920^3))=22.1%.

Now we move to the second phase of the Cox Regression model-building process ("Block 2"), where we add customer Category as a categorical predictor and then examine its influence on churn. Here is the next table of output, followed by explanation and discussion:

Omnibus Tests of Model Coefficients(a,b)

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
3253.012	213.600	6	.000	29.521	3	.000	29.521	3	.000

a Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
b Beginning Block Number 2. Method = Enter

The change from previous step and change from previous block both report the effect of adding customer category to the model selected in Block 1. Since the significance value of the change is less than 0.05, we can be confident that customer category contributes to the model.

Next comes the table of predictive model coefficients, followed by explanation and discussion:

Block 2: Method = Enter
Variables in the Equation

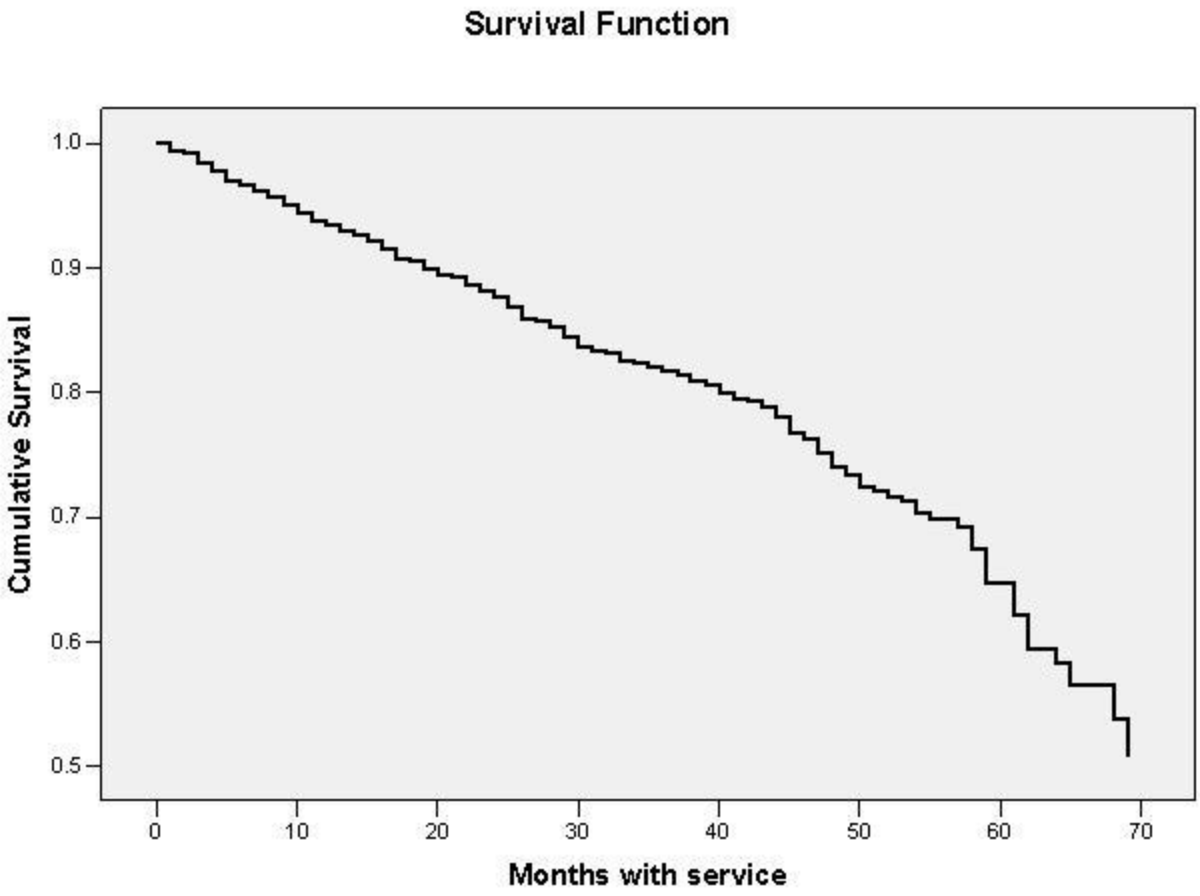
	B	SE	Wald	df	Sig.	Exp(B)
marital	.432	.123	12.358	1	.000	1.541
address	-.061	.009	49.768	1	.000	.940
employ	-.081	.010	67.141	1	.000	.922
custcat			28.506	3	.000	
custcat(1)	.121	.155	.612	1	.434	1.129
custcat(2)	-.574	.170	11.450	1	.001	.563
custcat(3)	-.658	.186	12.479	1	.000	.518

The Cox Regression coefficients for the first three levels of CUSTCAT are relative to the reference category, which corresponds to Total service customers. The regression coefficient for the first category, corresponding to Basic service customers, suggests that the hazard for Basic service customers is 1.129 times that of Total service customers. However, the significance value for this coefficient is greater than 0.10, so any observed difference between these customer

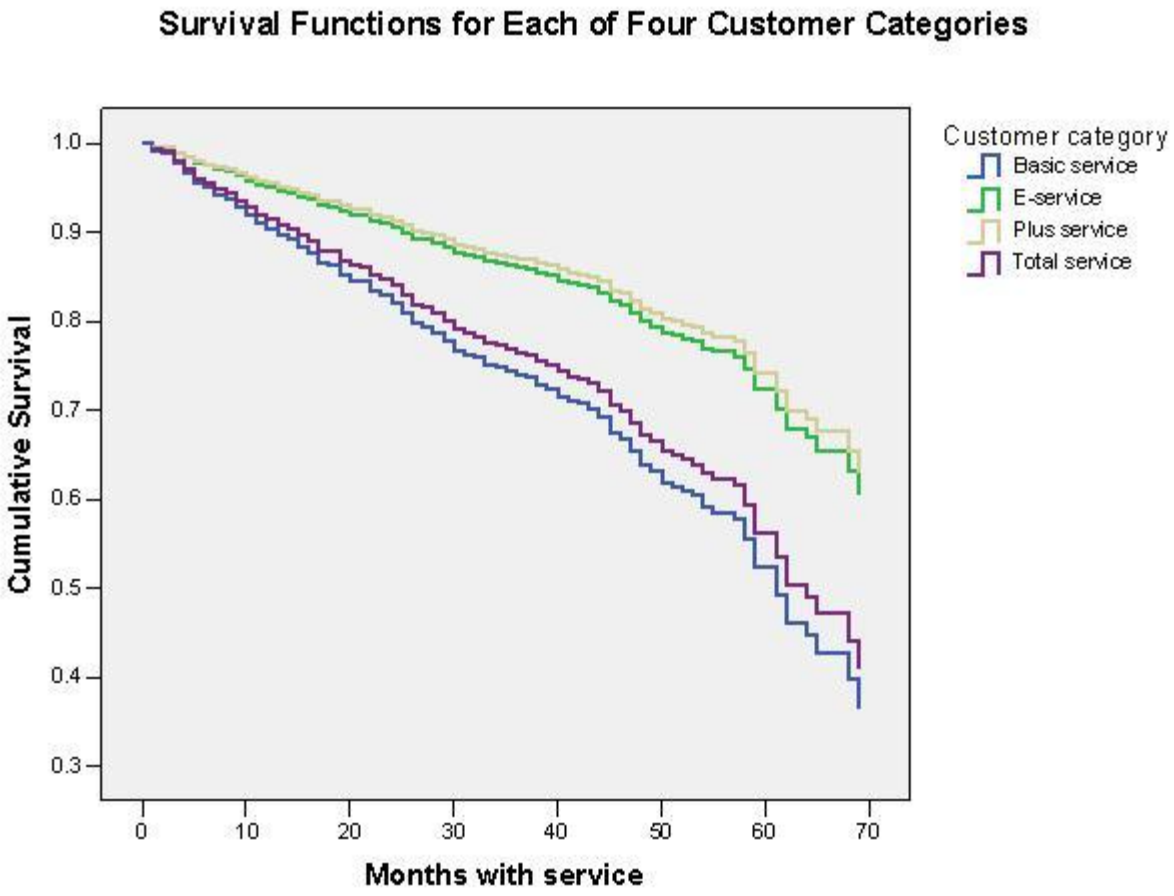
categories could be due to chance.

By contrast, the significance values for the second and third categories, corresponding to E-service and Plus service customers, are less than 0.05, which means they are statistically different from the Total service customers. The regression coefficients suggest that the hazard for E-service customers is 0.563 times that of Total service customers, and the hazard for Plus service customers is 0.518 times that of Total service customers.

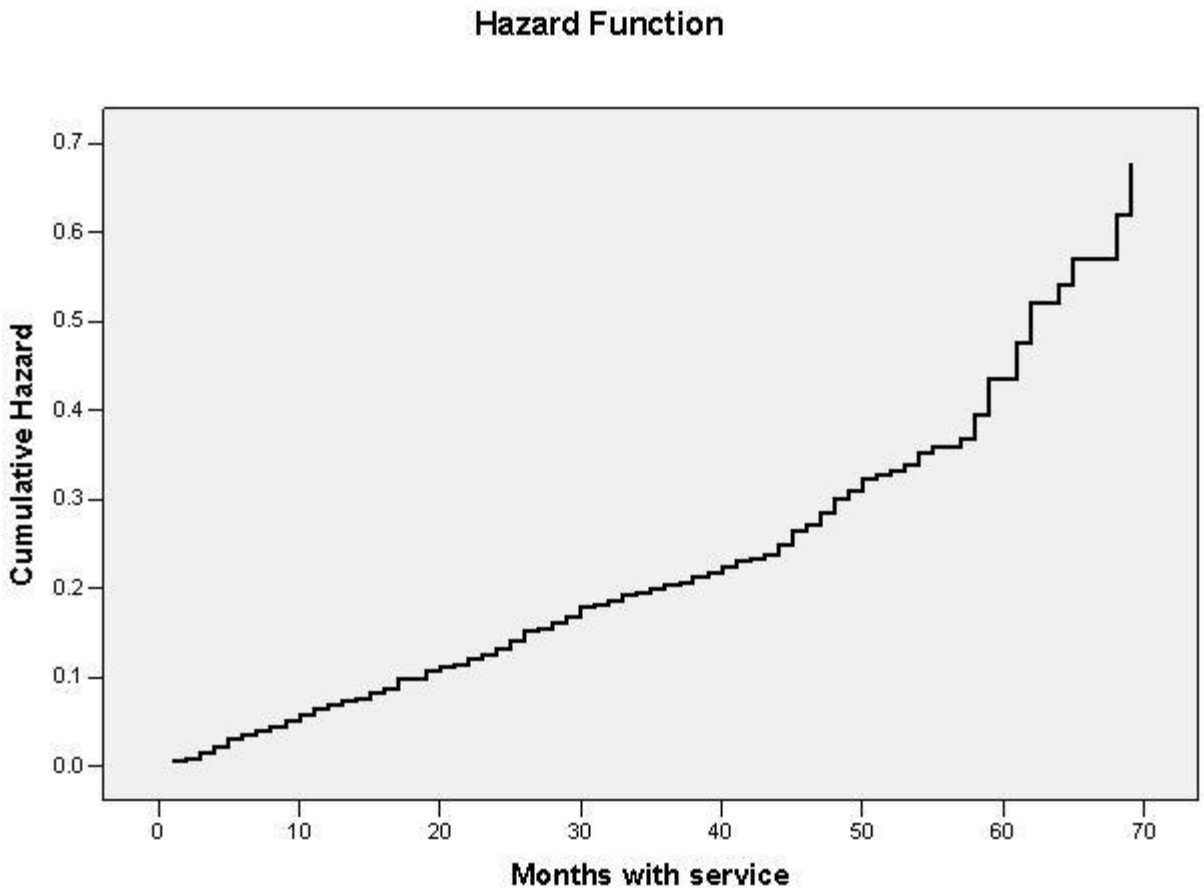
Below is a graphical representation of the “survival” or loyalty function generated from the model. The basic survival curve is a visual display of the model-predicted time to churn for the "average" customer. The horizontal axis shows the time to event. The vertical axis shows the probability of survival. Thus, any point on the survival curve shows the probability that the "average" customer will remain a customer past that time. Past 55 months the survival curve becomes less smooth. There are fewer customers who have been with the company for that long, so there is less information available, and thus the curve is blocky.



The plot of the survival curves gives a visual representation of the effect of customer category, which is shown in the graph below:

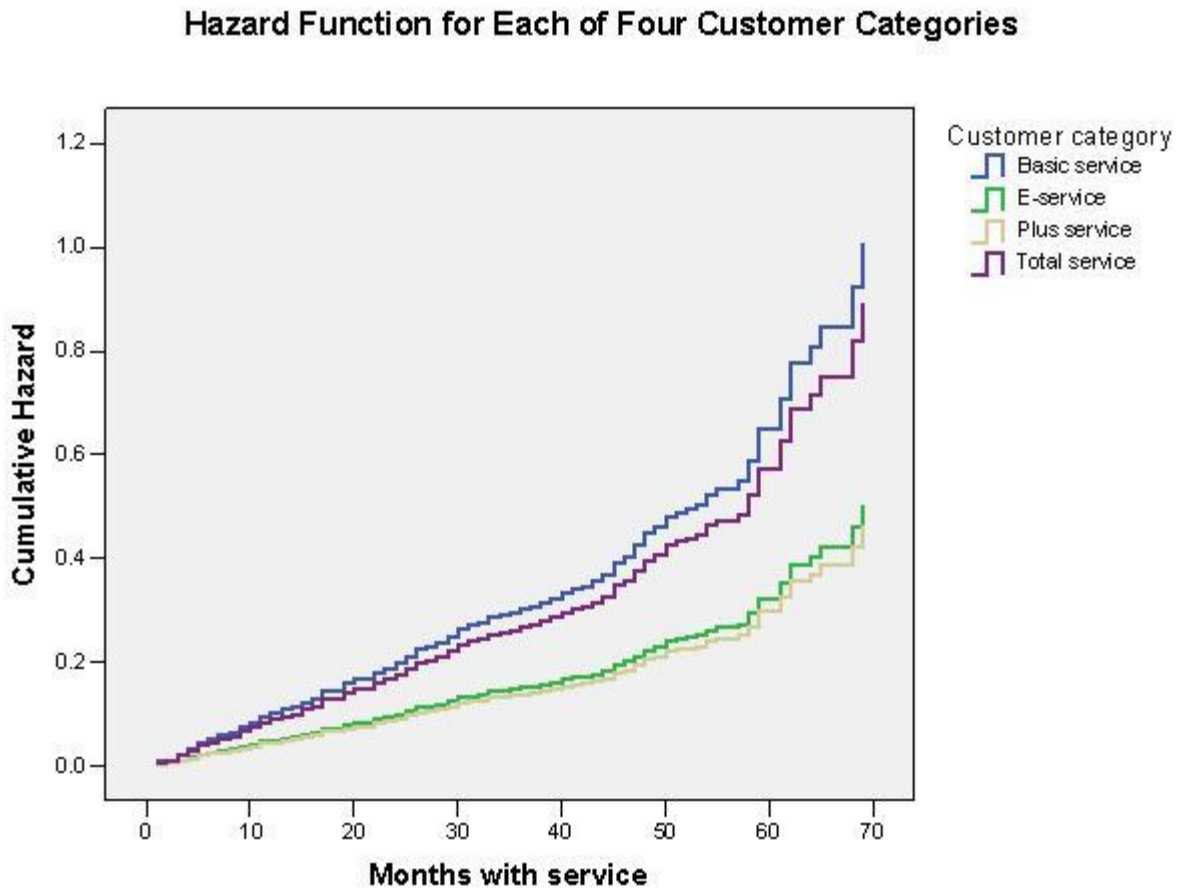


From the above graph we can see that Total service and Basic service customers have lower survival curves because, as we have learned from their regression coefficients, they are more likely to have shorter times to churn. The basic hazard curve, shown below, is a visual display of the cumulative model-predicted potential to churn for the "average" customer:



The horizontal axis shows the time to event. The vertical axis shows the cumulative hazard, equal to the negative log of the survival probability. Beyond 55 months, the hazard curve, like the survival curve, becomes less smooth, for the reason stated previously.

The plot of the hazard gives a visual representation of the effect of customer category:



Total service and Basic service customers have higher hazard curves because, as we have learned from their regression coefficients, they have a greater potential to churn.

Summary and Conclusions

We have found a suitable **Cox Regression model** for predicting time to customer churn. The use of separate blocks for fitting the model has allowed us to guarantee that customer category would be in the final model, while still taking advantage of the stepwise techniques for choosing the other variables in the model. To create this model, we included customer category in the second block. [Alternatively, the addition of customer category to the model could have taken place in the first block, and the stepwise methods to choose the other variables in the second block.]

We have discovered that marital status, length of time at current address, and length of time with current employer are all significant influences on time to churn, as is customer

category. By understanding these influences, we can identify customers who are most likely to defect at any given point in the customer relationship. This makes it possible for us to target these vulnerable customers with timely outreach efforts aimed at maintaining loyalty.

The foregoing case study is an edited version of one originally furnished by SPSS, and is used with their permission.