Table of content

Business problem : Churn

Saas company care about customer churn for their subscription services

To predict customer churn at a given time, logistic regression or random forests

However, from a business strategy perspective, we really want a prediction on a longer horizon

Data of this Project

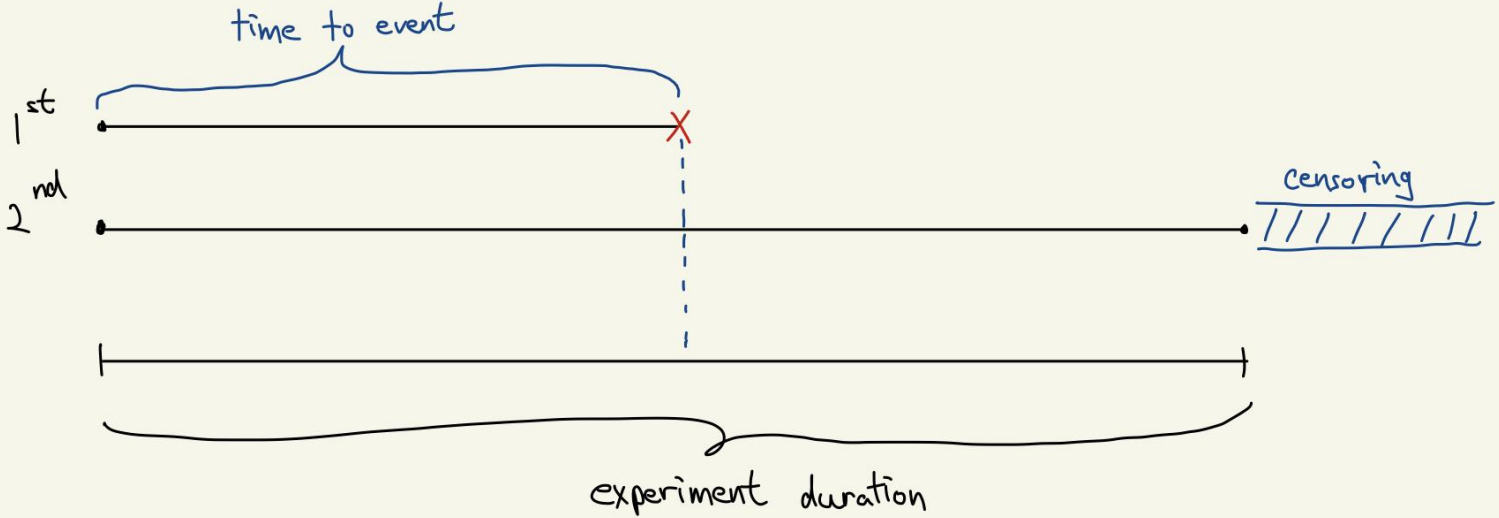
Question: at any given time how likely is customer X going to churn?

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes

DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes

Illustration of our data on a timeline:

Another complication: Censored data



Key characteristics of survival analysis

1) You want to answer: “ what is the time until an event of interest occur”
2) Censored data: lose track of an individual for various reasons.

2 crucial aspect of the data in order to apply survival analysis:

- time to event
- whether the event of interest has occurred or not.

Some other relevant business problem

- Inventory stock out is a censoring event for true "demand" of a good.
- A/B tests to determine how long it takes different groups to perform an action.

Terminology

The *survival curve*, or *survival function*, is defined as

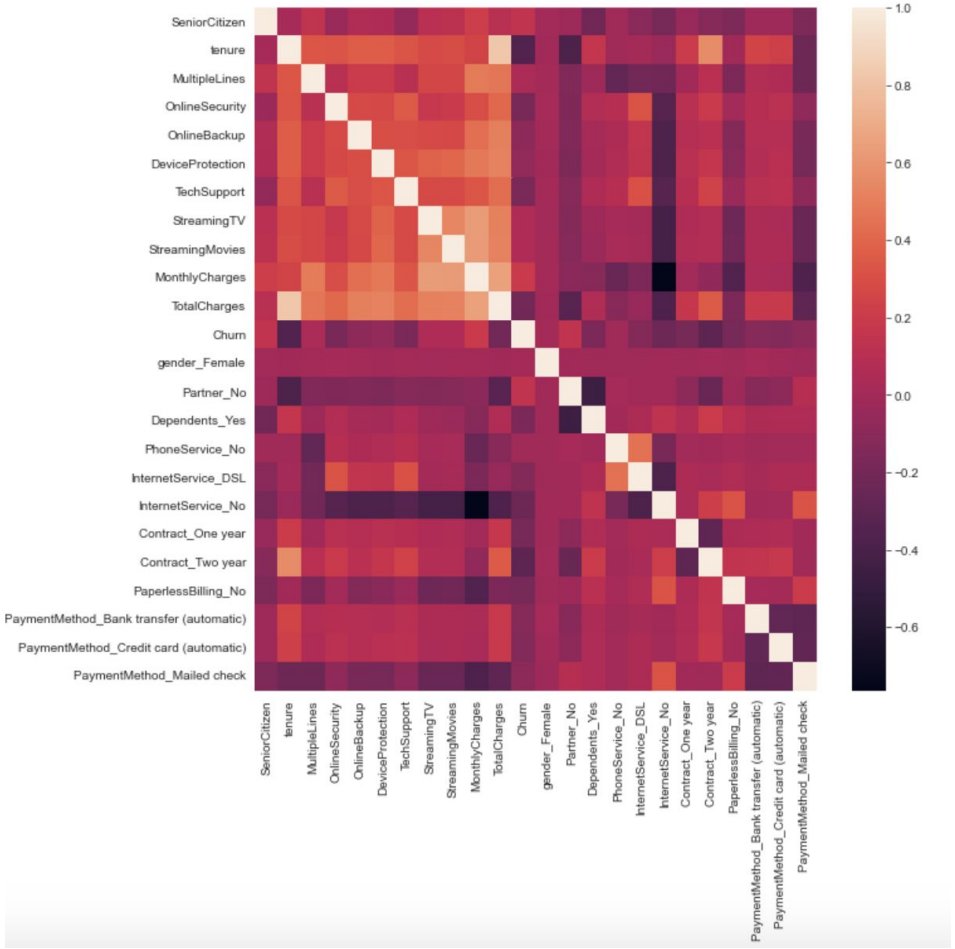
$$S(t) = \Pr(T > t).$$

The *hazard function* or *hazard rate*

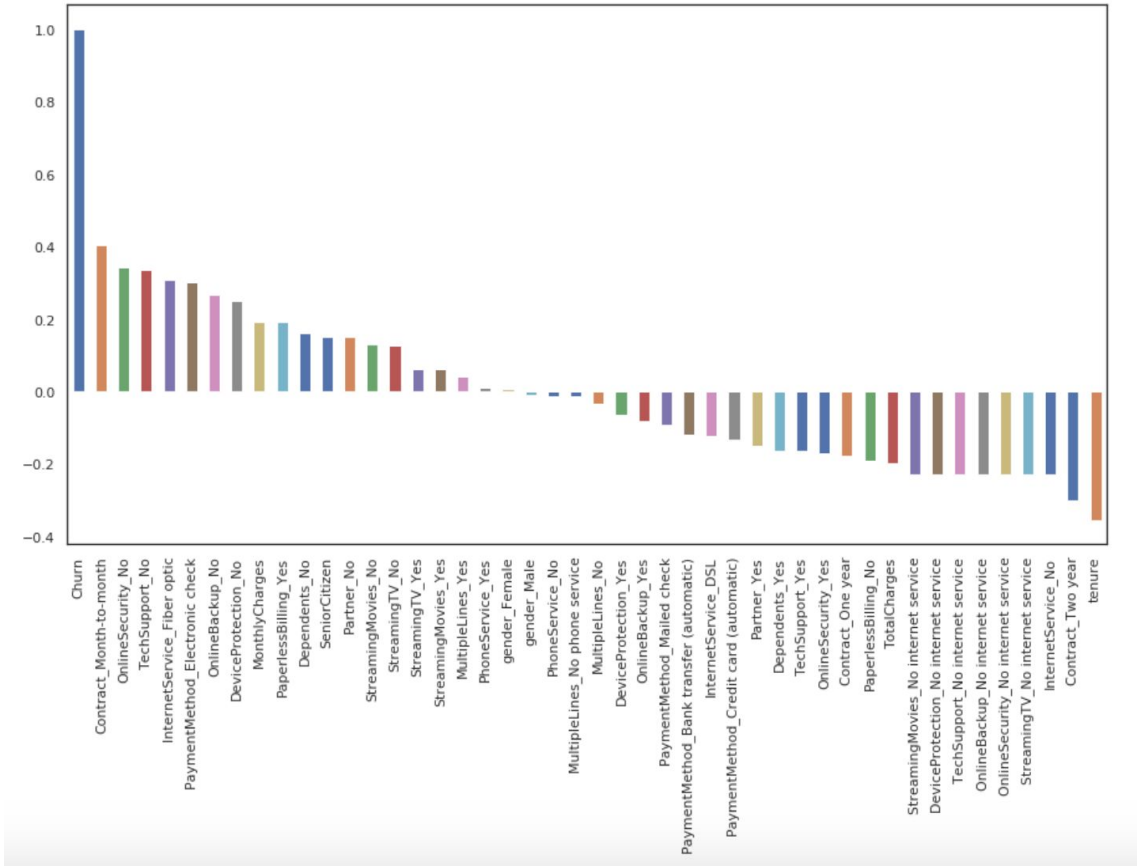
$$\lambda = \quad h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

Methodology

Correlation



Initial plot of how attributes correlate with Churn:

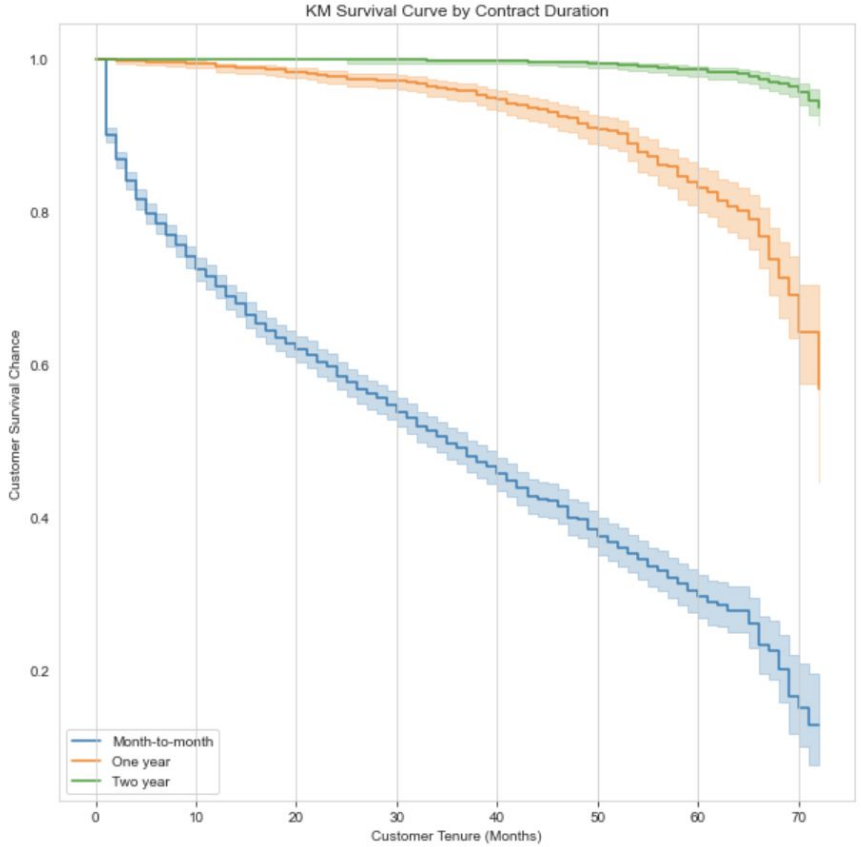


Hypothesis

We hypothesize that those who are on month to month contract are more likely to churn compared to those who are on 2 year contracts

Kaplan Meier Survival Curve

$S(t)$



sequential construction (toy example)

Time	Died (events)
2	1
3	0
6	1
7	1

Time	# at risk	# Died	Haz= (#died/ #at risk)	1-Haz	S(t)
0	4	0	0/4	1 - 0/4	100%
2	4	1	1/4	1- 1/4	75%
6	2	1	1/2	1- 1/2	75% *50%
7	1	1	1/1	0	0

Formal test 2 year contract vs month-to-month : Log rank test

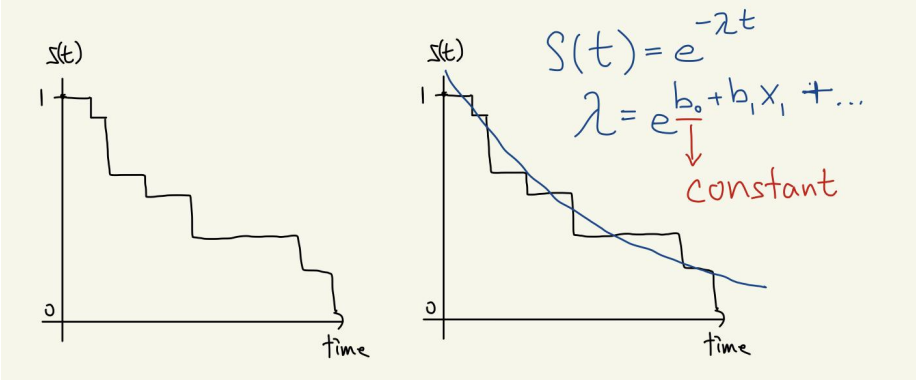
How can we carry out a formal test of equality between these 2 survival curves

- Modified version of a 2 sample t-test

test_name	logrank_test		
	test_statistic	**p**	**-log2(p)**
0	926.06	<0.005	673.27

Pros and Cons of different models

	Kaplan Meier	Exponential	Cox Proportional Hazard
Pro	**-C**an estimate S(t)	**-C**an estimate S(t) and Hazard Ratio	**-H**azard can fluctuate with time **-C**an estimate hazard ratio
Con	**-N**o functional form **-C**an't estimate hazard ratio **-E**ach line is 1 category (no other explanatory variables)	**-A**ssumes constant hazard rate	**-C**an't estimate S(t)



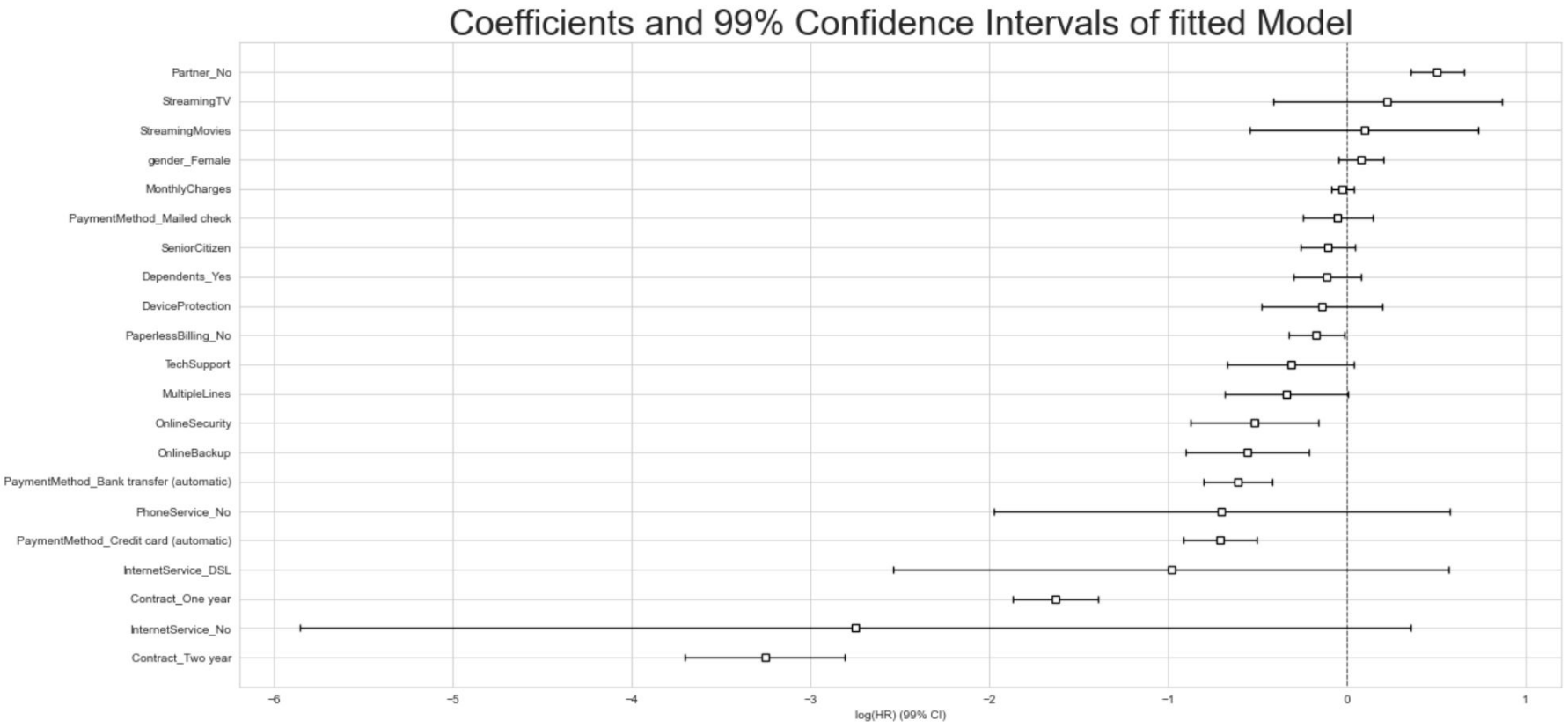
$$h(t|x_i) = h_0(t) \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$$

Cox proportional Hazard

	coef	exp(coef)	p
OnlineSecurity	-0.612	0.542	<0.0005
OnlineBackup	-0.613	0.542	<0.0005
Partner_No	0.518	1.679	<0.0005
Contract_One year	-1.615	0.199	<0.0005
Contract_Two year	-3.232	0.039	<0.0005
PaymentMethod_Bank transfer (automatic)	-0.586	0.557	<0.0005
PaymentMethod_Credit card (automatic)	-0.670	0.512	<0.0005

Concordance	0.866
Partial AIC	27811.201
log-likelihood ratio test	3536.878 on 21 df
-log2(p) of ll-ratio test	inf

Crossed validated and interpretation



Prediction:

And some questions I haven't figured out yet

- It appears that the prediction function in lifeline resets the tenure of all remaining customers to time 0.
- The prediction function is under cox proportional module, uses exponential in the background.

	0	1	3	6	7	9	10	11	12	14	15	16	17	19
1.0	9.865287e-01	9.821800e-01	0.980593	9.584022e-01	9.870164e-01	0.984968	0.973050	0.999902	0.995536	9.579725e-01	0.999763	0.936380	0.998634	0.876950
2.0	9.733905e-01	9.670089e-01	0.956018	8.981427e-01	9.684255e-01	0.962484	0.931183	0.999790	0.988936	9.147429e-01	0.999528	0.864369	0.998634	0.805678
3.0	9.602823e-01	9.402829e-01	0.942409	8.077974e-01	9.493901e-01	0.911304	0.895397	0.999686	0.982108	8.672213e-01	0.999073	0.830677	0.998634	0.707775
4.0	9.465329e-01	9.124417e-01	0.908845	7.594965e-01	9.326578e-01	0.858330	0.857477	0.999574	0.974109	8.032162e-01	0.999073	0.753046	0.998634	0.572844
5.0	9.362754e-01	8.944037e-01	0.881442	7.107693e-01	9.063390e-01	0.793681	0.815908	0.999493	0.968815	7.382433e-01	0.999073	0.702613	0.998634	0.506541
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
68.0	2.039518e-130	9.920619e-18	0.000098	1.202735e-227	1.049658e-216	0.238641	0.000000	0.177488	0.581734	1.097364e-158	0.999073	0.000023	0.998634	0.000000
69.0	3.765787e-181	9.920619e-18	0.000098	1.202735e-227	1.049658e-216	0.238641	0.000000	0.177488	0.581734	1.097364e-158	0.999073	0.000023	0.998634	0.000000
70.0	1.821399e-231	9.920619e-18	0.000098	1.202735e-227	1.049658e-216	0.238641	0.000000	0.177488	0.581734	1.097364e-158	0.999073	0.000023	0.998634	0.000000
71.0	0.000000e+00	9.920619e-18	0.000098	1.202735e-227	1.049658e-216	0.238641	0.000000	0.177488	0.581734	1.097364e-158	0.999073	0.000023	0.998634	0.000000
72.0	0.000000e+00	9.920619e-18	0.000098	1.202735e-227	1.049658e-216	0.238641	0.000000	0.177488	0.581734	1.097364e-158	0.999073	0.000023	0.998634	0.000000

Conclusions

- Try to tie customer with a 2 year contract
- Offer incentive to have automatic/ credit card payment
- Family bundles!

future

Lasso, to do variable selection

Not sure if people are going to game the system (credit loan approval)