

Subscription Survival Analysis

SF Data Mining Meetup
San Francisco
June 2012

Jim Porzak – Senior Director, Business Intelligence

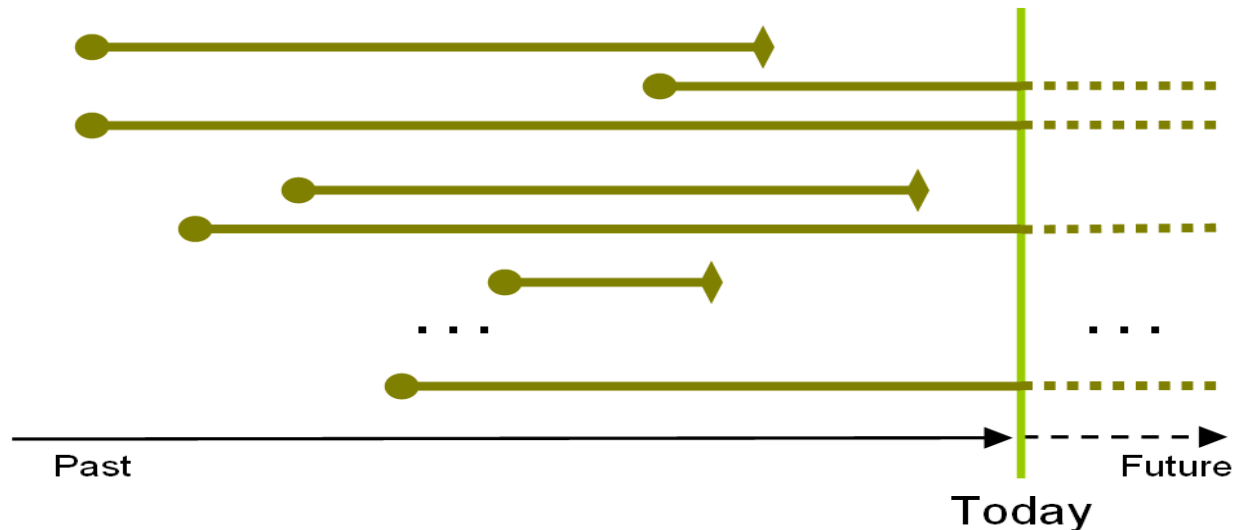
What We'll Cover...

- Quick introduction to survival analysis.
- Rational for marketing & business.
- Subscription businesses past & present.
- The example data set. How real is it?
- Calculating survival, average tenure & LTV.
- Applications of Survival & Hazard curves.
- Discussion.
- Appendix (links & details).

Traditional Survival Analysis

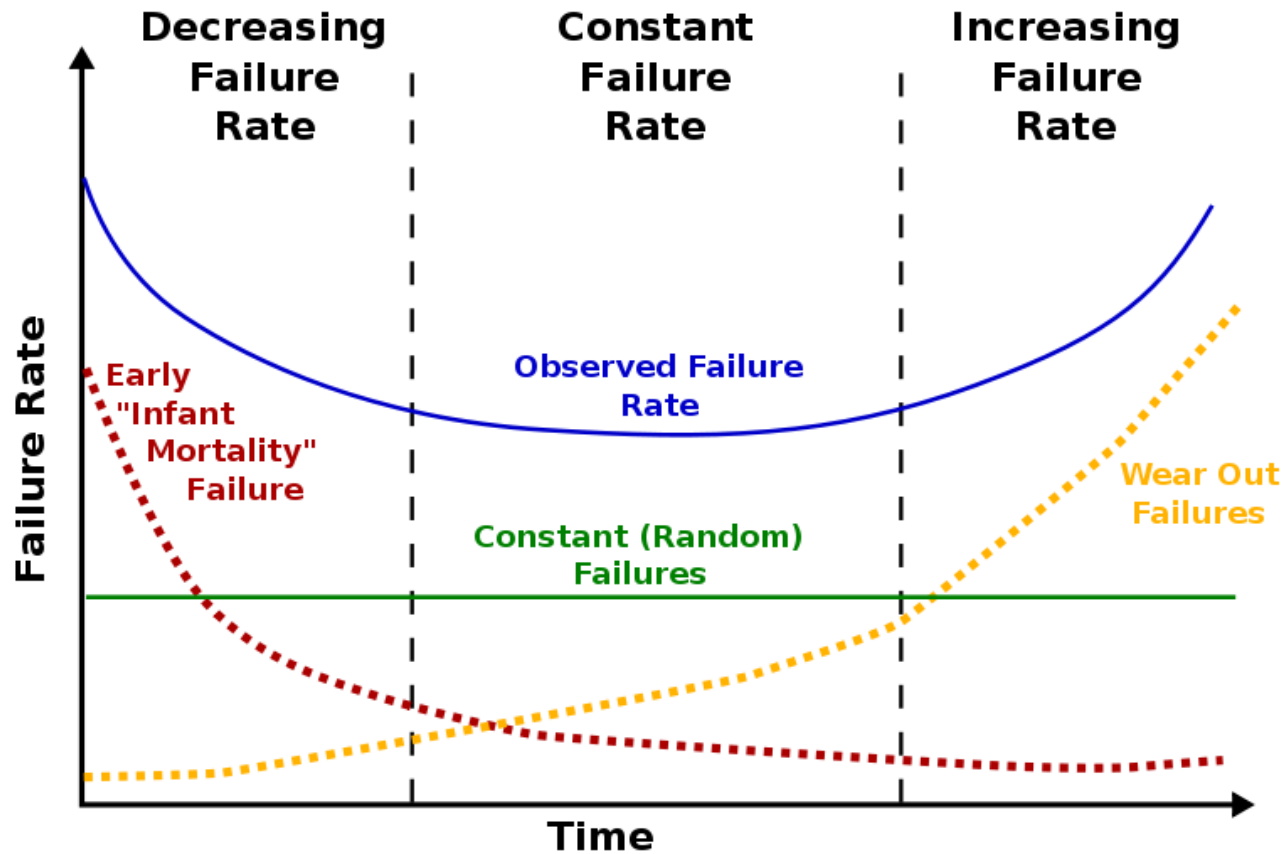
- Modeling of “time to event” data
 - Death in biological organisms (medicine)
 - Failure of machines (reliability engineering)
 - Political or sociological change (duration analysis)
- These kinds of data share common gotcha:
 - If an “individual” has not yet died, failed, or changed, what can we say about the expected time to the event?

Trick: Use what we know – everything!



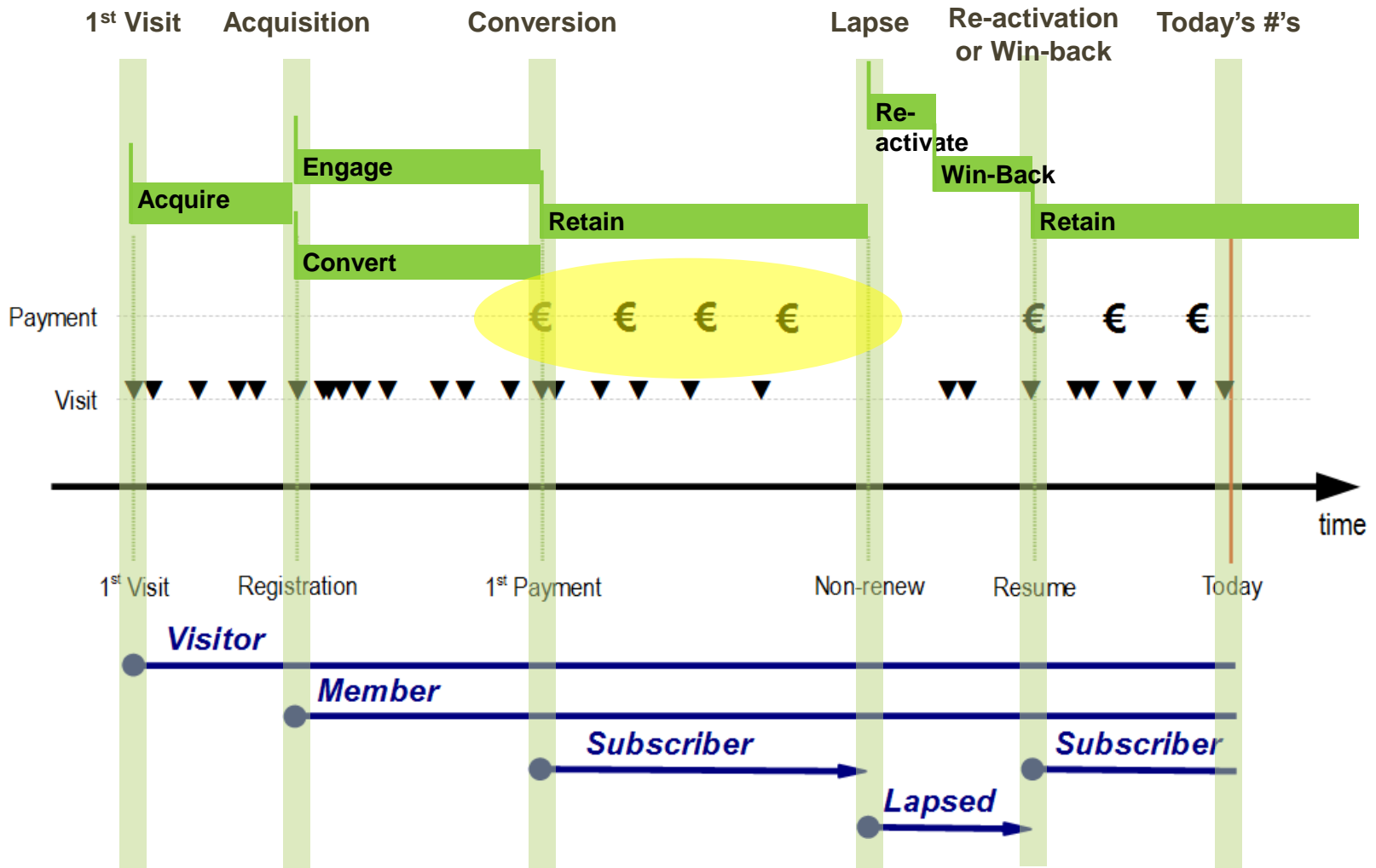
- All have started, some have stopped, & some continue on.
 - “Right censored”
- From this kind of data we can calculate probability of stopping.
 - “Hazard Ratio” = $f(\text{tenure}, \dots)$
- Probability of survival at tenure, T , is just $(1 - \text{cumulative hazard}(T))$

Classical “Bathtub” Hazard Curve



Source: http://en.wikipedia.org/wiki/Bathtub_curve

The Subscription Business Model

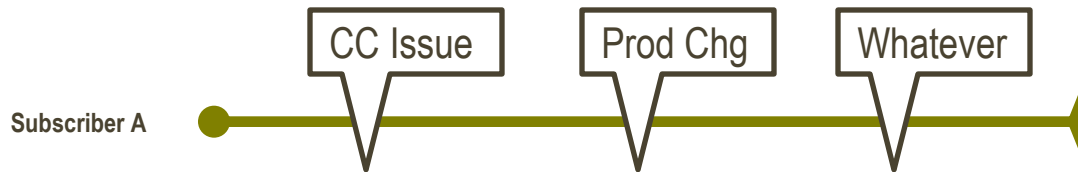


Aside: Subscription Stints – Rational & Method

- For each member we need to know subscription start &, if terminated, stop dates for each member:



- But data often has the accounting view with gaps in the subscription:

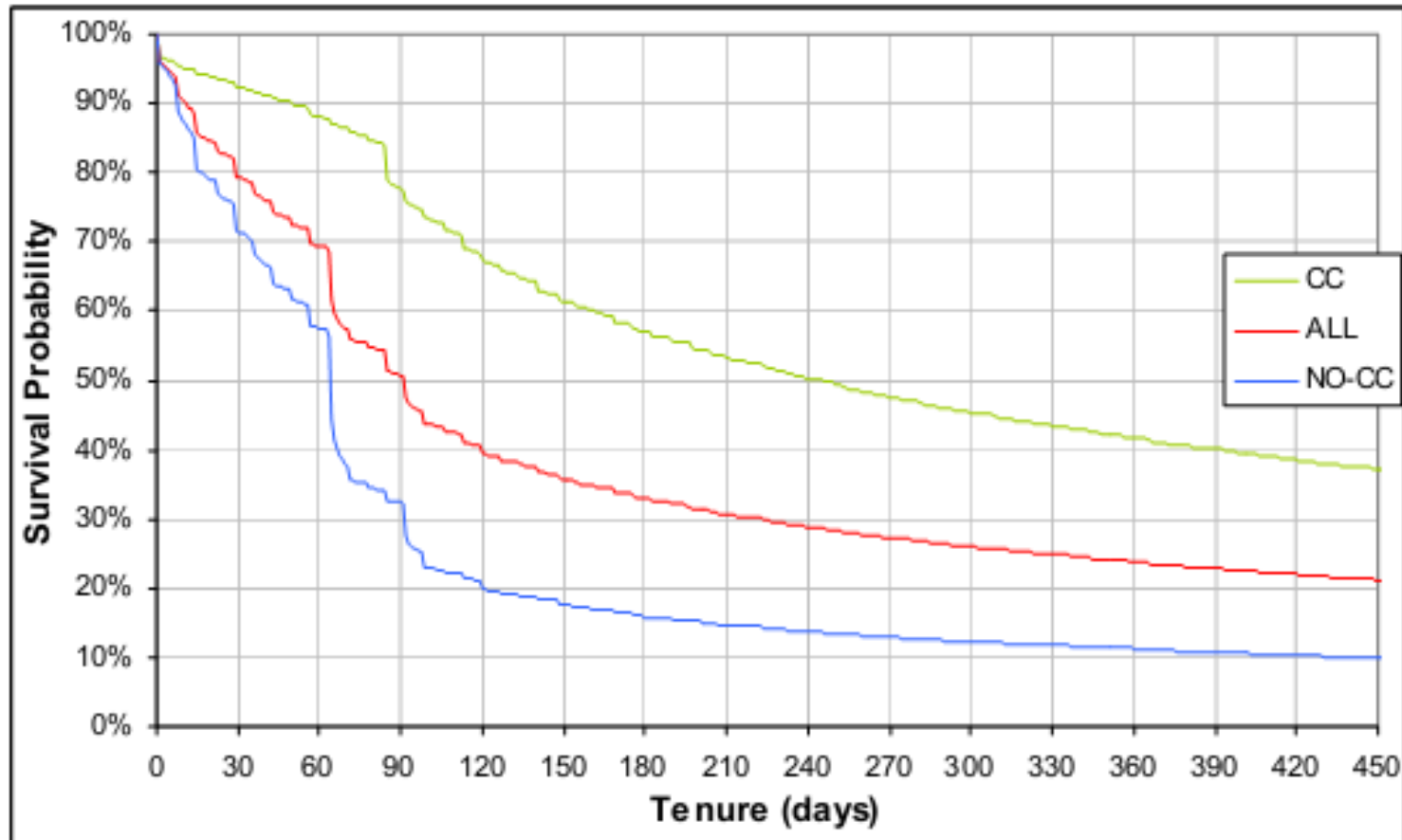


- From the customer's perspective, he had one continuous subscription. It's important to take that view since, in general, the longer a subscriber is active the lower the risk of churn.
- A subscription stint algorithm typically ignores any gaps up to, say, 30 days – which is also taken as the boundary between a lapsed subscriber in the “re-activate” vrs “win-back” pools.

Why Subscription Survival?

- Expected churn probability of new subscribers
 - Project future subscriber count
 - Especially after a successful campaign
 - Define tenure segment boundaries
- Expected tenure of a new subscriber & LTV
 - As function of duration, product, channel, offer
 - Evaluate marketing & site tests by value not rates
- Projected value of retention efforts
 - How improved renewal rates increase value
- A new way of thinking about subscribers
 - Retention metrics, as used in marketing, are deceptive (see Appendix slide)

Typical Subscription Survival Curves



Linoff, Gordon (2004) Link in appendix.

Assumptions

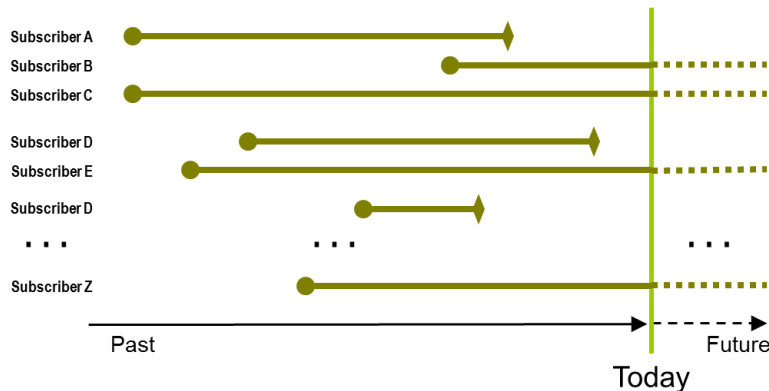
- Homogeneity of strata
 - Each includes only one segment of customers
- Stability over time
 - Hardest to satisfy?
 - Propensity to renew a function of:
 - General economic conditions
 - Offer(s) at start of subscription
 - Engagement efforts by marketing group
 - Perceived long term value

Example Data Set

- 100k records:
 - CustomerID
 - SubStartDate
 - SubEndDate (NULL if subscription active)
 - Duration = [A, Q, M] for Annual, Quarterly, or Monthly
 - Product = [B, P] for Basic or Professional
 - Channel = [R, A, S, B, O] for Referral, Affiliate, Search, Banner, or Other
- Randomly generated with day-of-week and week-in-year seasonality; 15% year over year growth for starts.
- Randomly generated number of renewals
 - With different churn rates for A, Q, & M
- Mimics general properties of subscription data from Playboy.com, LA Times, 24 Hour Fitness, Chicago Sun Times, Ancestry.com, & Viadeo.com
 - Additional parameters could be Offer, Promotion Cohort, ...
- Assumes “no-return” policy, i.e. no early refunds.

Algorithm – Part 1: Hazard Ratio

Subscribers on Calendar Time Axis



Subscribers on Tenure Time Axis

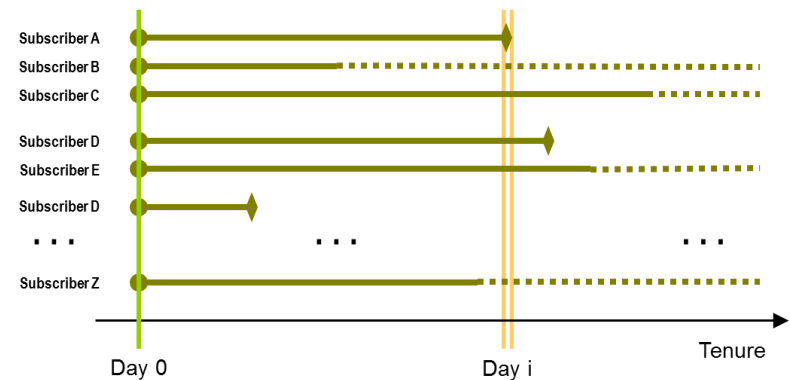


$$HR_i = \frac{(\# \text{ terminated during day } i)}{(\# \text{ active at beginning of day } i)}$$

$$= 1 / 4 = 0.25$$

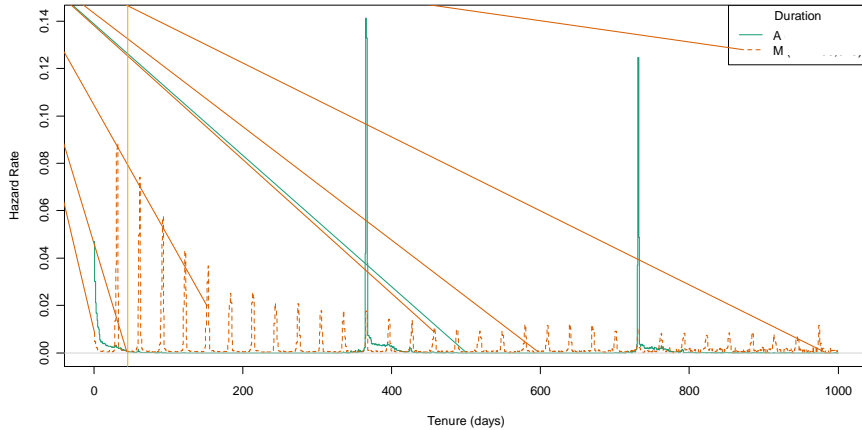
Remember N is typically huge in a subscription business so we tend to get smooth curves.

Empirical Hazard Ratio Calculation



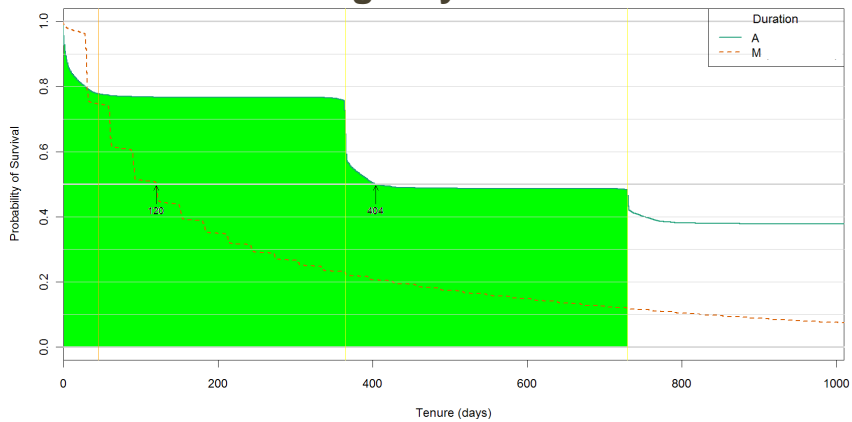
Algorithm – Part 2 Survival Curve

Hazard Ratio

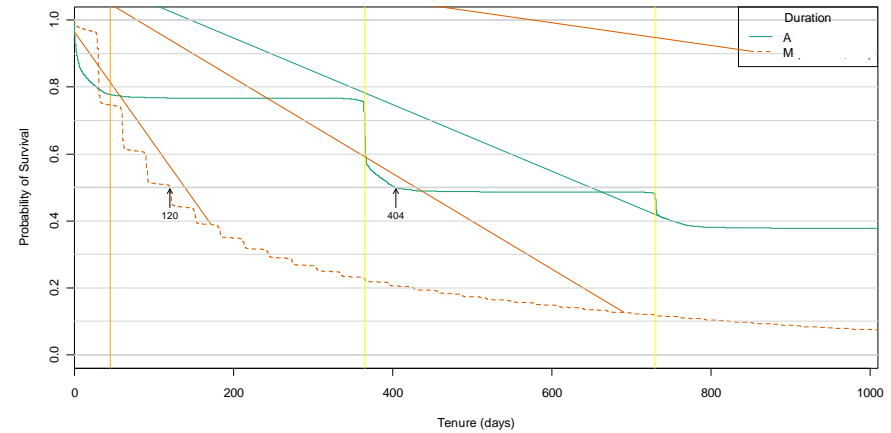


$$P_{\text{survival}}(t_i) = \prod_{j=0}^i (1 - \text{HR}(t_j))$$

Average 2-year Tenure

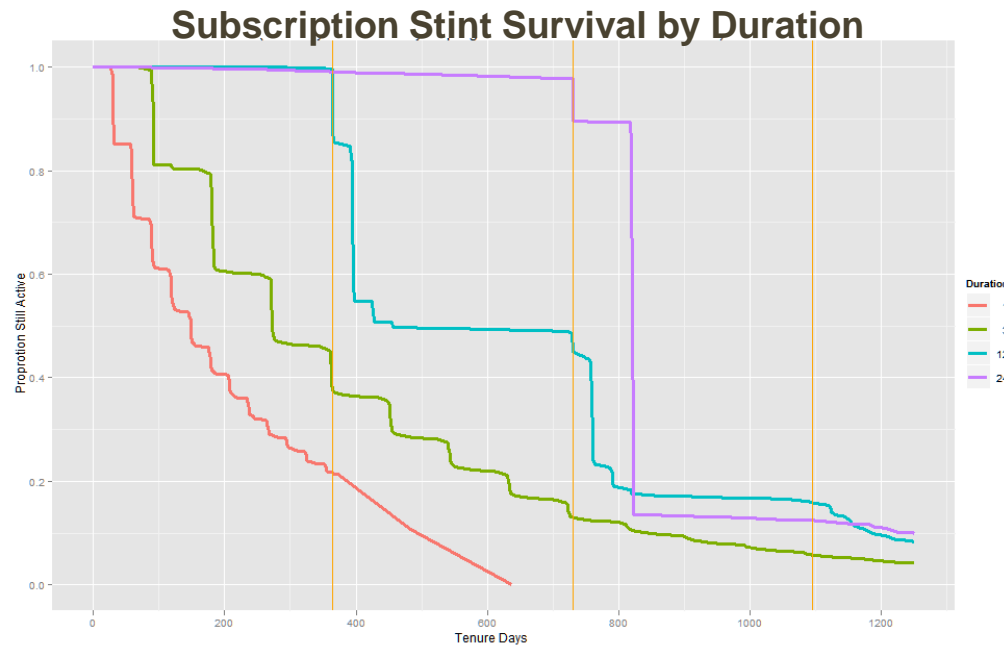


Survival Probability



LTR = (Daily Sales Price) * (Average Tenure)

Application: Average Tenure of Renewal Durations



Duration (Months)	Half Life (days)	1-Year Average Tenure	1-year Probability of Survival	2-Year Average Tenure	2-year Probability of Survival	3-Year Average Tenure	3-year Probability of Survival
1	149	180.9	0.217	NA	NA	NA	NA
3	273	260.9	0.380	354.4	0.131	387.7	0.058
12	456	364.7	0.880	557.8	0.452	629.9	0.159
24	822	363.8	0.991	723.1	0.978	840.3	0.125

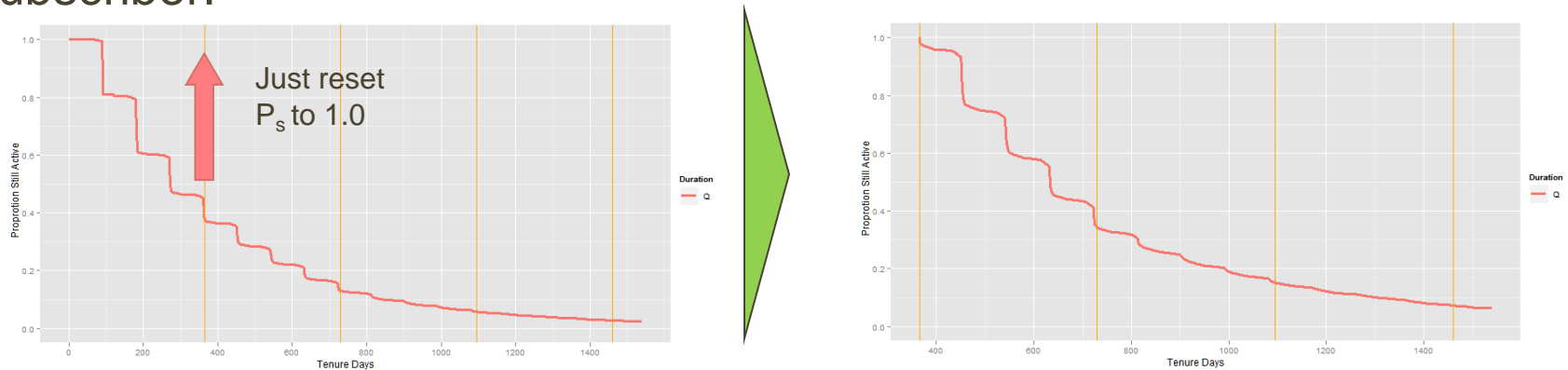
Application: Evaluating a Conversion Test

We test a great new design for the web site conversion page flow & get 1050 conversions for new flow vs. 1000 for old flow - a 5% lift!

				A: Control Page Flow		B: New Page Flow	
	2-Yr Tenure	Daily ASP	2-Yr Value	# Conv.	Projected 2-Yr Value	# Conv.	Projected 2-Yr Value
M	220	0.15	\$33.00	500	\$16,500	650	\$21,450
Q	350	0.12	\$42.00	300	\$12,600	275	\$11,550
A	560	0.09	\$50.40	200	\$10,080	125	\$6,300
Total				1000	\$39,180	1050	\$39,300
Revenue Per Unit (2-Year)					\$39.18		\$37.43

Application: Projected Value of Base

So far we have looked at survival from start of subscription. Now we need remaining survival for current subscribers. For example a 1-year subscriber:



Projected n-year value of your current subscriber base is:

For all subscribers, Sum their

(daily sales price) *

(Area under appropriate survival curve

from their current tenure over the next n years) /

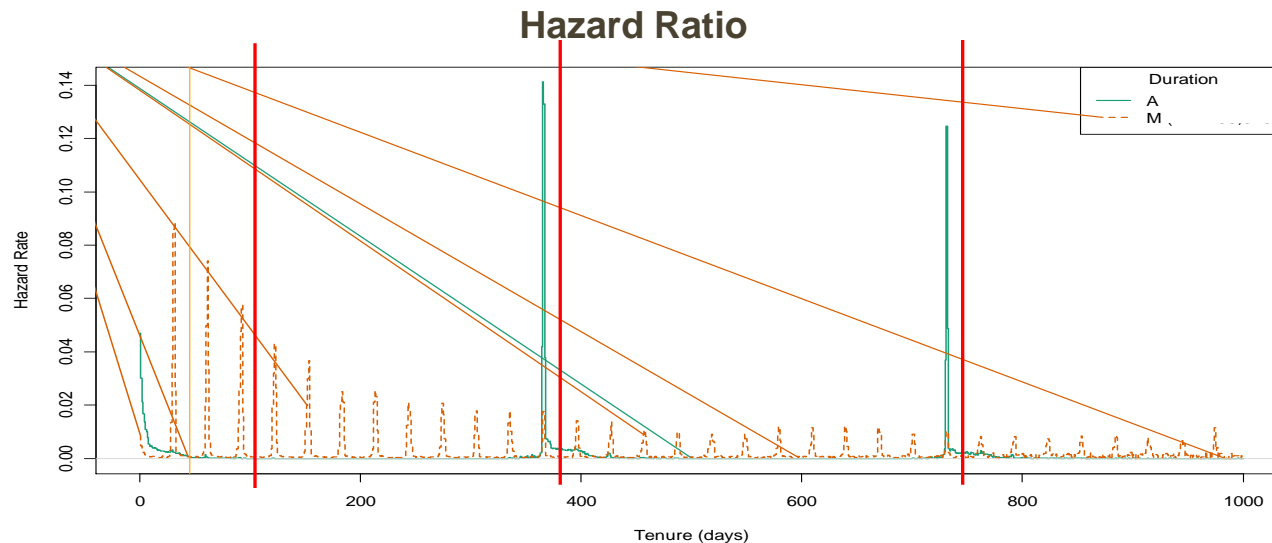
(P_s (current tenure))

Application: Thinking in Hazard Space

Picking Tenure Segment Boundaries

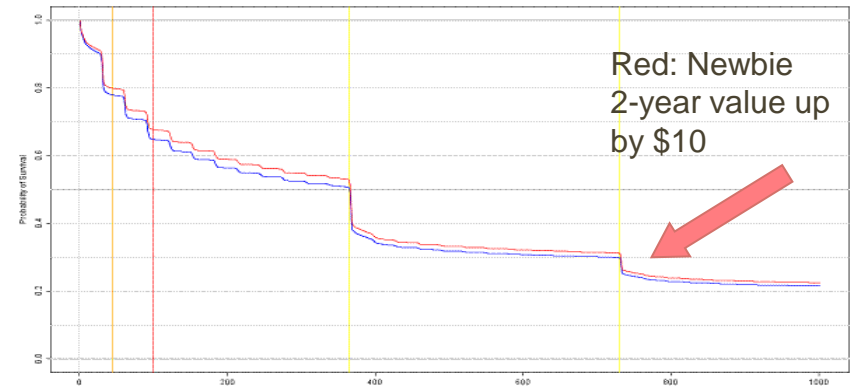
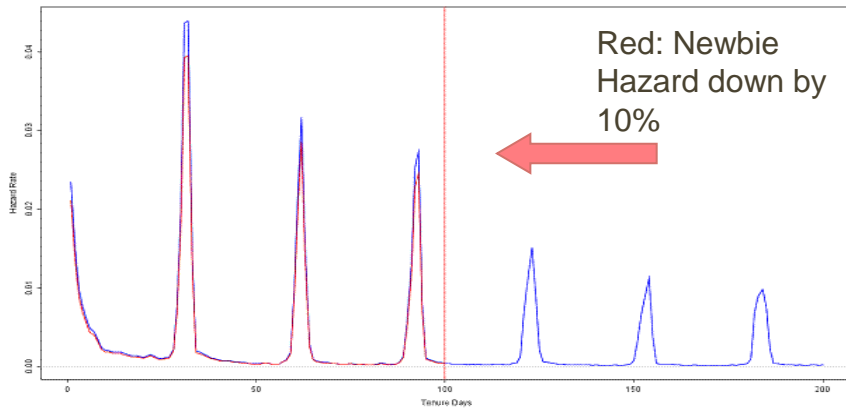
Simple minded customer segmentation. But very useful!

- Newbie: 0 - 100 days
- Builder: 101 - 380 days
- Established: 381 - 735 days
- Core: 736 days & up



Application: Modeling in Hazard Space

What-if Newbie Churn Could be Reduced by 10%?



Wrap-up. What we saw.

- Introduction to classical survival
- Subscribers & calculating their hazard & survival
- Applying survival & hazard to real-world questions

- See Appendix for links & some more details.

Questions?



And, ping me later with questions or
comments: JPorzak@gmail.com

Appendix

- Links to learn more
- Retention vs Survival Curves

Subscription Survival Background

- On subscription survival for customer intelligence:
 - Gordon Linoff's 2004 Article from Intelligent Enterprise (now InformationWeek) <http://www.data-miners.com/resources/Customer-Insight-Article.pdf>
 - Will Potts' technical white paper <http://www.data-miners.com/resources/Will%20Survival.pdf>
 - Berry & Linoff's white paper for SAS, emphasizes forecasting application http://www.data-miners.com/companion/sas/forecastingWP_001.pdf
 - Chapter 10 in <http://amzn.to/mRAVpp>
- Background on “classical” survival analysis:
 - http://en.wikipedia.org/wiki/Survival_analysis
- R packages used
 - survival by Terry Therneau
 - ggplot2, plyr, lubridate by Hadley Wickham, et al
 - zoo, xts by Achim Zeileis, Jeffrey Ryan, et al

Retention vs Survival Curves

- Retention Rate = 1 – Churn Rate
- Churn Rate (typically) defined as:
$$\frac{(\# \text{ Subscribers leaving})}{(\text{Average } \# \text{ subscribers})}$$

over some period.
- Which means:
 - Ignores information out of period
 - Not monotonically decreasing

Traditional vs. Subscription Survival

Property	Traditional	Subscription
N	Smallish ($10^1 - 10^3$)	Large to huge (10^5 & up)
Hazard	Assumed continuous	Usually spiky
Survival curves	Often Modeled	Empirical
Survival Curve(s)	Assumed continuous	Usually stair step