Estimate Attrition Using Survival Analysis

Luyang Fu, Ph.D., FCAS
Hongyuan Wang, Ph.D.

CAS Spring Meeting, May 2013, Vancouver, Canada

Agenda

- Introduction
- Survival Analysis
- Cox Proportional Hazard Model
- A Case Study
- Q&A

Introduction

Attrition/retention is important to insurance companies

Growth (Top Line)

- *"We were not successful in raising customer renewal rates, so that the new business success did not result in overall growth this quarter"*

  All State 2010Q4 Earnings call

- Higher retention, less pressure of attracting new business

  To grow a book with 10,000 policies by 10%,

  - if retention is 90%, need to attract 2000 new accounts: 1000 to make up attrition, 1000 for the growth
  - if retention is 70%, need to write 4000 new accounts: 3000 to make up attrition, 1000 for the growth

Introduction

Attrition/retention is important to insurance companies

Profitability (Bottom Line)

- *"Given our strong retentions as well as the new business and account growth we've achieved over the last few years, we have significant positive leverage to an improving environment."*

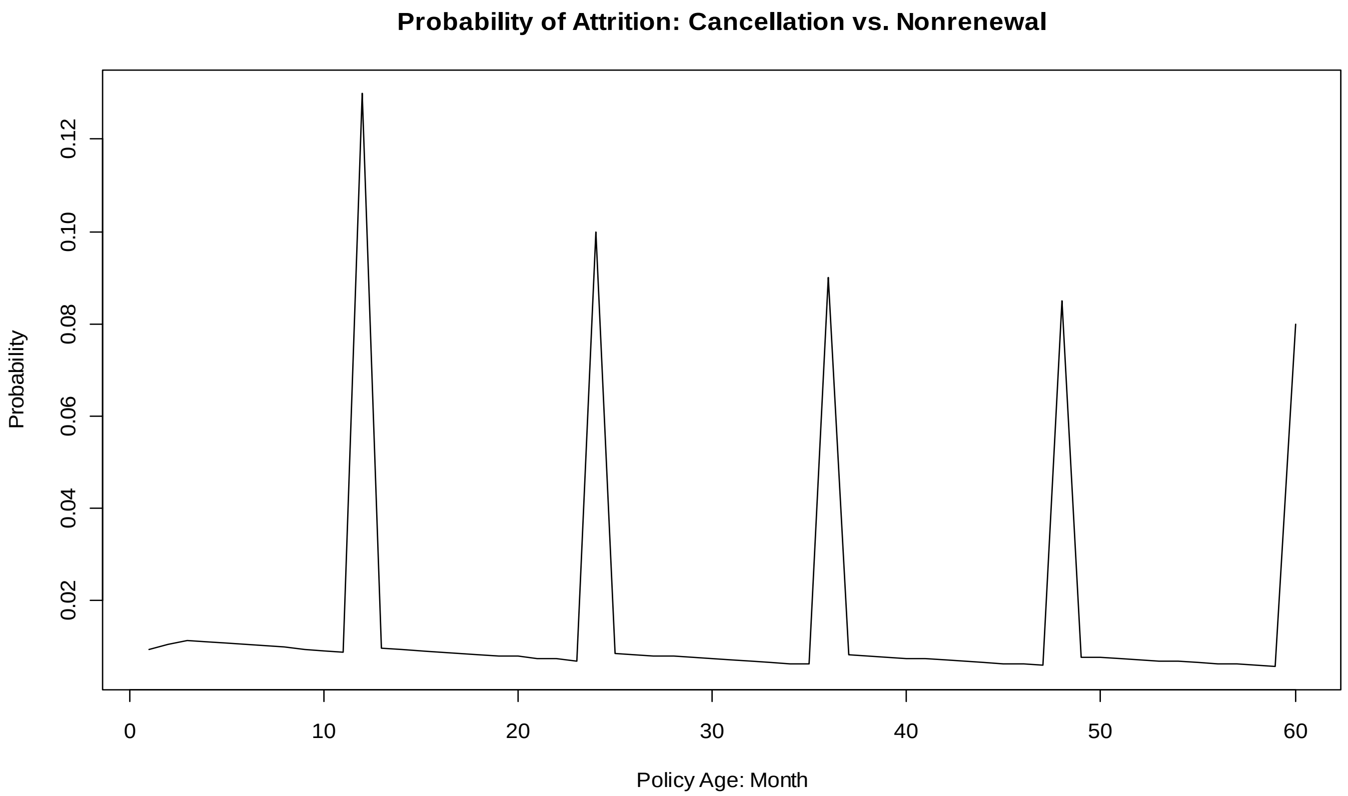
  Travelers 2010Q4 Earning call

- Ageing Phenomenon
  - D'Arcy and Doherty (1989; 1990): loss ratio improves with policy age
  - Wu and Lin (2009): renewal book on average has a loss ratio 13% better than new business by examining 8 lines of business, 25 books, $29 billion premium
- Price Optimization
  - Retention, conversion, price elasticity
  - Life-time value

Introduction

Two types of Attritions

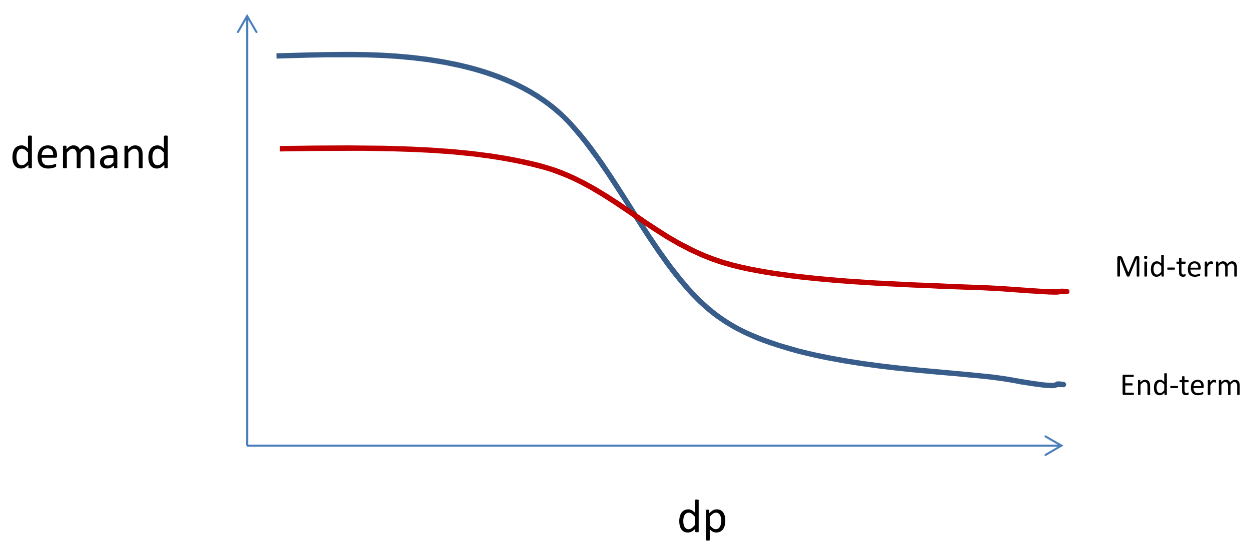
- Mid-term cancellation
- End-term nonrenewal



Introduction

Two types of attritions behave differently

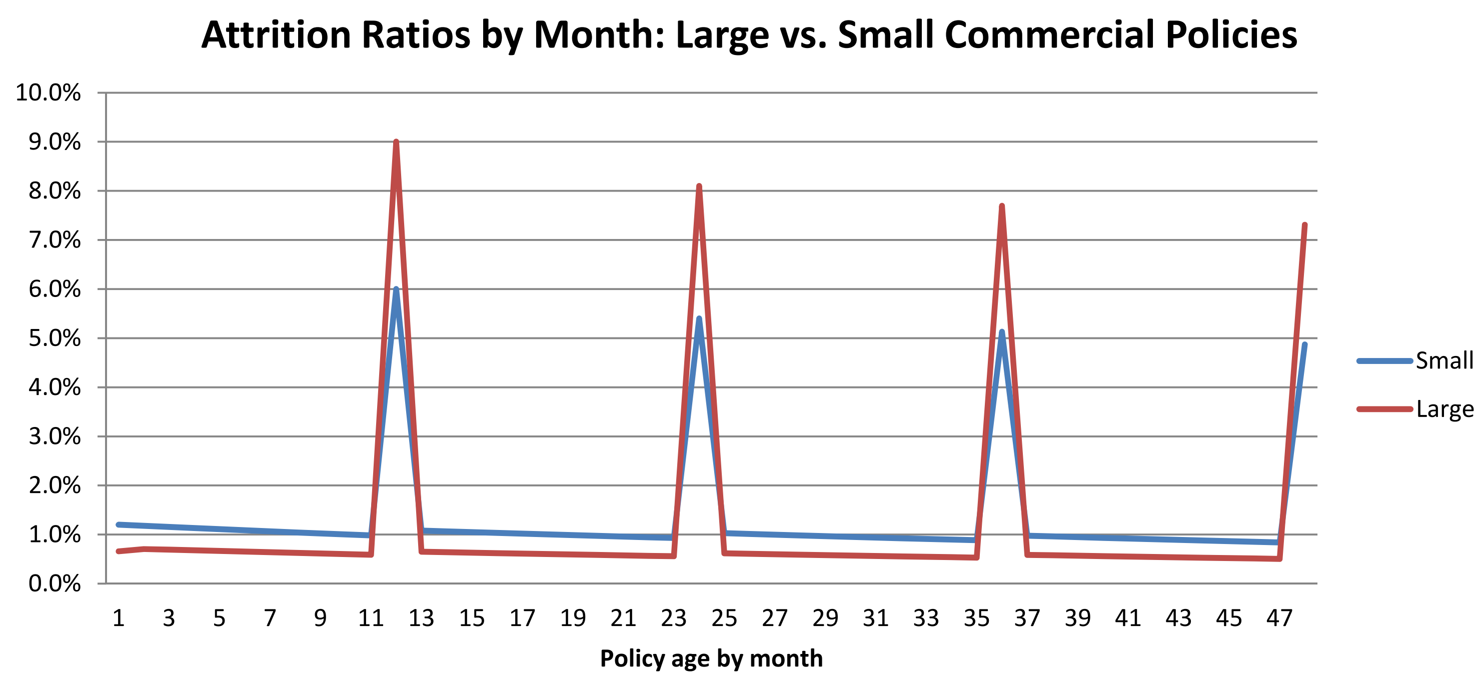
- Example 1: price elasticity
  - End-term nonrenewal is more sensitive to price change
  - Mid-term cancellation may be from non-pricing reasons



Introduction

Two types of attritions behave differently

- Example 2: policy size in commercial lines
  - Small policies may have a higher mid-term cancellation ratio than large policies
  - Large policies may have a higher end-term nonrenewal ratio



Introduction

Traditional Retention Analysis

- Renewal ratio at expiration month
  - If 1,000 policies expire at May 2013, 920 of them are still with the company at 05/31/2013. The renewal ratio is 92%.
  - The evaluation lag may vary.
  - It ignores the attrition from mid-tern cancellation
  - It does not give an annual view of retention or attrition

Introduction

Traditional Retention Analysis

- Annual Retention: Snapshot comparison
  - If there were 10,000 inforced policies at 12/31/2011, 8,500 of them were still effective at 12/31/2012, the annual retention ratio is 85%.
  - Does not analyze the sources of attritions. 15% is the sum of mid-term cancellation and end-term nonrenewal

Introduction

Traditional Retention Analysis

- Logistics models
  - Data: snap-shot data
  - Variable of interest: yes or no
  - Do not model cancellation and nonrenewal separately (can be extended to model two ways of attritions independently).
  - Static view

Introduction

Why survival analysis?

- Estimate mid-term cancellation and end-term nonrenewal sequentially and simultaneously
  - Survival Analysis:
    - Reflect two ways of attritions through the seasonality within survival curve
    - Recognize the aging sequence of the same policy (panel data approach)
  - Logistics Regression:
    - Snap-shot data cannot separate mid-term and end-term attritions
    - Treat each record within the same policy panel independently

Introduction

Why survival analysis?

- Better estimation of life time value: not just whether a policy will leave, but when it will leave
  - Survival Analysis:
    - Target variable of interest: t (time to attrition)
    - If 10,000 policies are inforce at 12/31/2009, 8,500 of them were still effective after a year. Among 1,500 attritions, how many of them left by cancelation and non-renewal, and when they left?
  - Logistics Regression:
    - Target variable of interest: yes or no
    - Ignore the time of attrition
    - Do not predict the attrition for non-integer multiples of the evaluation horizon

Introduction

Why survival analysis?

- Better utilization of time-varying macroeconomic variables
  - Survival Analysis:
    - Dynamic view of treasury yield, GDP change, and stock market return, etc.
    - Reflect interest rate, inflation, consumer confidence at the time of attrition
  - Logistics Regression:
    - Static view of those variables
    - If "yes" or "no" is constructed by comparing 2011 with 2012 year-end book, one summarized "unemployment rate" is used for all the records
    - Flinn and Heckman (1982): reliance on ad hoc procedures to cope with time-trended variables in logistic regression can produce very pathological estimates

Introduction

The disadvantages of survival analysis

- Model implementation is not as straightforward as binary model
  - Logistic
    - Probability of attrition is the direct output of model
  - Survival analysis
    - Develop baseline survival function
    - Derive hazard function for individual policies
    - Calculate the probability of attrition

Introduction

The disadvantages of survival analysis

- Time-varying macroeconomic variables are more difficult to predict than retention
  - How to capitalize the relationship between retention and time-varying macroeconomic variables
  - The models on interest rates and stock indexes are much more complex then retention models
  - Macroeconomic variables are more volatile than retention, and may introduce additional volatility into retention projection.

Introduction

Literatures on Marketing and Banking

- Helsen K. and D. C. Schmittlein, "Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models", *Marketing Science*, 1993, Vol. 12, No. 4, 395-414.
- Stepanova M. and L. C. Thomas, "Survival Analysis for Personal Loan Data", *Operations Research*, 2002, Vol. 50, 277-289.
- Van den Poel D., and B. Lariviere, "Customer Attrition Analysis for Financial Services using Proportional Hazard Models", *European Journal of Operational Research*, 2004, vol. 157, No 1, 196-217
- Graves S, D. Kletter, W. B. Hetzel, R. N. Bolton, "A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction", *Marketing Science*, 1998, Vol. 17, No. 1, pp. 45-65.
- Andreeva G., "European Generic Scoring Models Using Survival Analysis", *Journal of the Operational Research Society*, 2006, Vol. 57, No. 10, pp. 1180-1187.
- Bellotti T. and J. Crook, "Credit Scoring With Macroeconomic Variables Using Survival Analysis", *Journal of the Operational Research Society*, 2009, Vol. 60, pp. 1699–1707.
- Tang L, L. C. Thomas, S. Thomas, J. F. Bozzetto, "It's the Economy Stupid: Modeling Financial Product Purchases", *International Journal of Bank Marketing*, 2007, Vol.25, No 1, 22-38.

Survival Analysis

- Another name for *time to event* analysis
- Statistical methods for analyzing survival data.
- Primarily developed in the medical and biological sciences (death or failure time analysis)
- Widely used in the social and economic sciences, as well as in Insurance (longevity, time to claim analysis).

Survival Analysis

Survival Time

- *t* measures the time from a particular starting time (e.g., time initiated the treatment) to a particular endpoint of interest (e.g., patient died).
- Examples:
  - *Insurance Policy*: Started at Jan2008, terminated at Aug2012.
  - *Products*: Bought at Dec2006, failed at Feb2009.
  - *Marketing*: coupon mailed at Jan2013, redeemed at March 2013.

Survival Analysis

Censoring

- Occurs when the value of a measurement or observation is only partially known.
- Left Censoring:

  Example: Subject's lifetime is known to be less than a certain duration.
- Right Censoring:

  Example: Subjects still active when they are lost to follow-up or when the study ends.

Survival Analysis

Functions in Survival Analysis

- Survival Function *S(t)* :
  S(t) = Prob{T $\geq$ t}, here t $\geq$ 0 ;
- Lifetime Distribution Function *F(t)* :
  F(t) = 1-S(t) ;
- Event Density Function *f(t)* :
  Prob{t$\leq$ T $\leq$ t+δt} = f(t)δt, $\frac{dF(t)}{dt} = f(t)$
- Hazard Function *h(t)* :
  h(t) = f(t)/S(t)
  or h(t)δt = Prob{t $\leq$ T$\leq$ t+δt |T$\geq$t};

Survival Analysis

All those functions are connected.

- Density function is the negative of the derivative of the survival function;
- Hazard function is the negative of the derivative of the log of the survival function.

$$f(t) = F'(t) = -S'(t)$$

$$h(t) = -\, d(\ln S(t)) \big/ dt$$

$$S(t) = \exp\left\{ -\int_0^t h(s)\,ds \right\}$$

$$f(t) = h(t) \exp\left\{ -\int_0^t h(s)\,ds \right\}$$

Survival Analysis

The most popular distribution assumptions are exponential, Weibull, etc.

- Exponential: $S(t) = exp(-\lambda t)$ $\lambda > 0$;

  $f(t)= \lambda exp(-\lambda t)$;

  $h(t) = \lambda$; (so no ageing)
- Weibull; $S(t) = exp(-\beta t^{\alpha})$ $\alpha, \beta > 0$;

  $f(t) = \alpha\beta t^{\alpha-1} (exp(-\beta t^{\alpha}))$;

  $h(t) = \alpha\beta t^{\alpha-1}$;

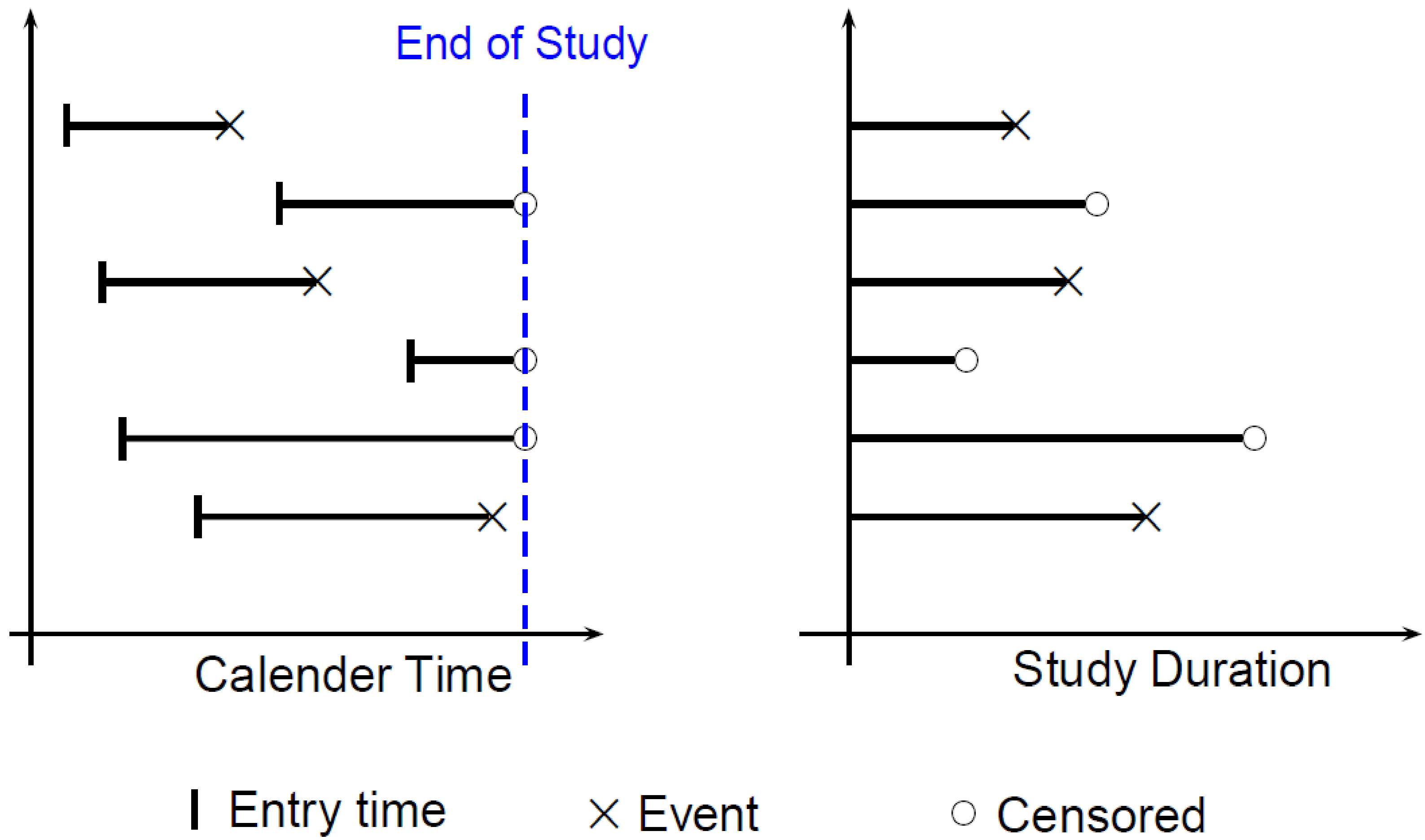
$\alpha > 1$ (increasing hazard) , $\alpha < 1$ (decreasing hazard)

Survival Analysis

Data

- Calendar time of whole study (Starting day, Ending day of the whole study period)
- Study Duration of each individual.
- Define the *censored observations.*
- Time measure units (Month, Year …)
- Define the dependent variable and independent.

Survival Analysis



Survival Analysis

Survival Analysis in Marketing

Duration Times of Interest in Marketing		
Subdiscipline	**Decision/Forecasting**	**Duration Time**
Pricing/Promotion	Timing of price changes or promotions; Measuring effect of promotion	Interpurchase duration; Timing of coupon redemption
Salesforce Management	Forecasting and managing salesforce turnover	Salesperson job duration
New Product Development	Forecasting trial, adoption, depth of repeat purchase	Duration time from new product introduction until initial trial; Interpurchase times
Marketing Research	Forecasting response rates; Forecasting size and composition of firm's customer base;	Time until survey response; Time until customer becomes inactive or disaffected; Time until cancellation of service contract;

Sources: Helsen K.. and D. C. Schmittlein, 1993, Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models; *Marketing Science*, Vol. 12, No. 4, page 396 .

Cox Proportional Hazard Model

Advantages

- The dependent variable of interest (survival/failure time) is most likely not normally distributed.
- Censoring(especially right censoring) of the Data.
- Baseline hazard function is unknown.
- Whether and when the customer will leave.
- Dynamics covariates and duration

Cox Proportional Hazard Model

Hazard equations

$h(t \mid x_t)$: hazard rate at time *t* for an individual have covariate value, x_t

$$h(t \mid x_t) = h_0(t) e^{\beta' x_t}$$

Here $x_t = (x_{1t}, x_{2t}, \ldots\ldots, x_{kt})$ $\beta = (\beta_1, \beta_2, \ldots\ldots, \beta_k)$

k is the total number of the covariates, x_j

β_j *is the constant Proportional effect of*

The term $h_0(t)$ is called the *baseline hazard*; it is the hazard for the respective individual when there is no covariate impacts.

Cox Proportional Hazard Model

Hazard Equations

We can linearize this model by dividing both sides of the equation by $h_0(t)$ and then taking the natural logarithm of both sides:

$$\ln\{ h(t \mid x_t) / h_0(t)\} = \beta' x_t$$

Taking partial derivative we have

$$\partial \ln h(t \mid x_t, \beta) / \partial x_{jt} = \beta_j$$

Cox Proportional Hazard Model

Partial Likelihood Estimation of β

$$L(i \mid t, j_1, j_2, \ldots\ldots, j_{n(t)}) = \frac{h_i(t)}{\sum_{k=1}^{n(t)} h_{j_k}(t)} \tag{1}$$

$$L(i \mid t, j_1, j_2, \ldots\ldots, j_{n(t)}) = \frac{h_0(t) e^{\beta' x_{it}}}{\sum_{k=1}^{n(t)} h_0(t) e^{\beta' x_{j_k t}}} \tag{2}$$

$$L(i \mid t, j_1, j_2, \ldots\ldots, j_{n(t)}) = \frac{e^{\beta' x_{it}}}{\sum_{k=1}^{n(t)} e^{\beta' x_{j_k t}}} \tag{3}$$

Estimation of $\boldsymbol{\beta}$ is obtained by Maximizing the Product of Expression (3) over all observed duration times.

Cox Proportional Hazard Model

Data Examples: 4 policies (A, B, C, D) from 01/01/00 to 12/31/03

Origination Date	Study Entry Date	Effective date	Term end date	Policy Age	Start Month T1	End Month T2	Right Censor	Attrition	Other Variables
01/01/2001	01/01/2001	01/01/2001	12/31/2001	0	0	12	0	0	$x_{A,12}$
01/01/2001	01/01/2001	01/01/2002	12/31/2002	1	12	24	0	1	$x_{A,24}$
07/01/2001	07/01/2001	07/01/2001	06/30/2002	0	0	12	1	0	$x_{B,12}$
07/01/2001	07/01/2001	07/01/2002	06/30/2003	1	12	24	1	0	$x_{B,24}$
07/01/2001	07/01/2001	07/01/2003	12/31/2003	2	24	30	1	0	$x_{B,30}$
03/01/1998	03/01/2000	03/01/2000	02/28/2001	2	0	12	0	0	$x_{C,12}$
03/01/1998	03/01/2000	03/01/2001	11/30/2001	3	12	21	0	1	$x_{C,21}$
01/01/1997	01/01/2000	01/01/2000	12/31/2000	3	0	12	1	0	$x_{D,12}$
01/01/1997	01/01/2000	01/01/2001	12/31/2001	4	12	24	1	0	$x_{D,24}$
01/01/1997	01/01/2000	01/01/2002	12/31/2002	5	24	36	1	0	$x_{D,36}$
01/01/1997	01/01/2000	01/01/2003	12/31/2003	6	36	48	1	0	$x_{D,48}$

Cox Proportional Hazard Model

Data examples: 4 policies (A, B, C, D) from 01/01/00 to 12/31/03)

$$L(A \mid 24, R_{24}) = \frac{\exp(\beta' x_{A,24})}{\exp(\beta' x_{A,24}) + \exp(\beta' x_{B,24}) + \exp(\beta' x_{D,24})}$$

$$L(B \mid 30, R_{30}) = \frac{\exp(\beta' x_{B,30})}{\exp(\beta' x_{B,30}) + \exp(\beta' x_{D,30})}$$

Cox Proportional Hazard Model

Literatures on Survival Analysis Theory

- Cox, D. R., "Regression Models and Life Tables (with discussion)," *Journal of the Royal Statistical Society Series B*, 1972, Vol. 34, 187-220.
- Efron, B., “Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve”, Journal of American Statistical Association, 1988, Vol. 83, 414-425.
- Flinn, C. and J. Heckman, "New Methods for Analyzing Structural Models of Labor Force Dynamics", Journal of Econometrics, 1982, vol. 18(1), 115-168.
- Kaplan, E.L. & Meier, P. "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 1958, Vol 53, 457-481.

Case Study

Data

- Not Real: simulated commercial lines data.
- Dependent variable:

  Duration = the time until the policy leaves
- If a policy is still effective at the end of study, it is right censored (i.e. Censor = 1)
- External data (including macroeconomic data) are joined into policy data.

Case Study

Data

- Define rate changes, removing the impacts from
  - Exposure changes (add a building; cut a class)
  - Coverage changes (reduce limits; increase deductible)
  - Risk characteristics changes (have a violation/claim; add a youthful driver)
- Groupings/binnings can be arbitrary
  - Contractors vs. noncontractors
  - Size groups
  - Variable interactions:
    - Small, medium, large contractors
    - General, nongeneral with sub, artisan contractors

Case Study

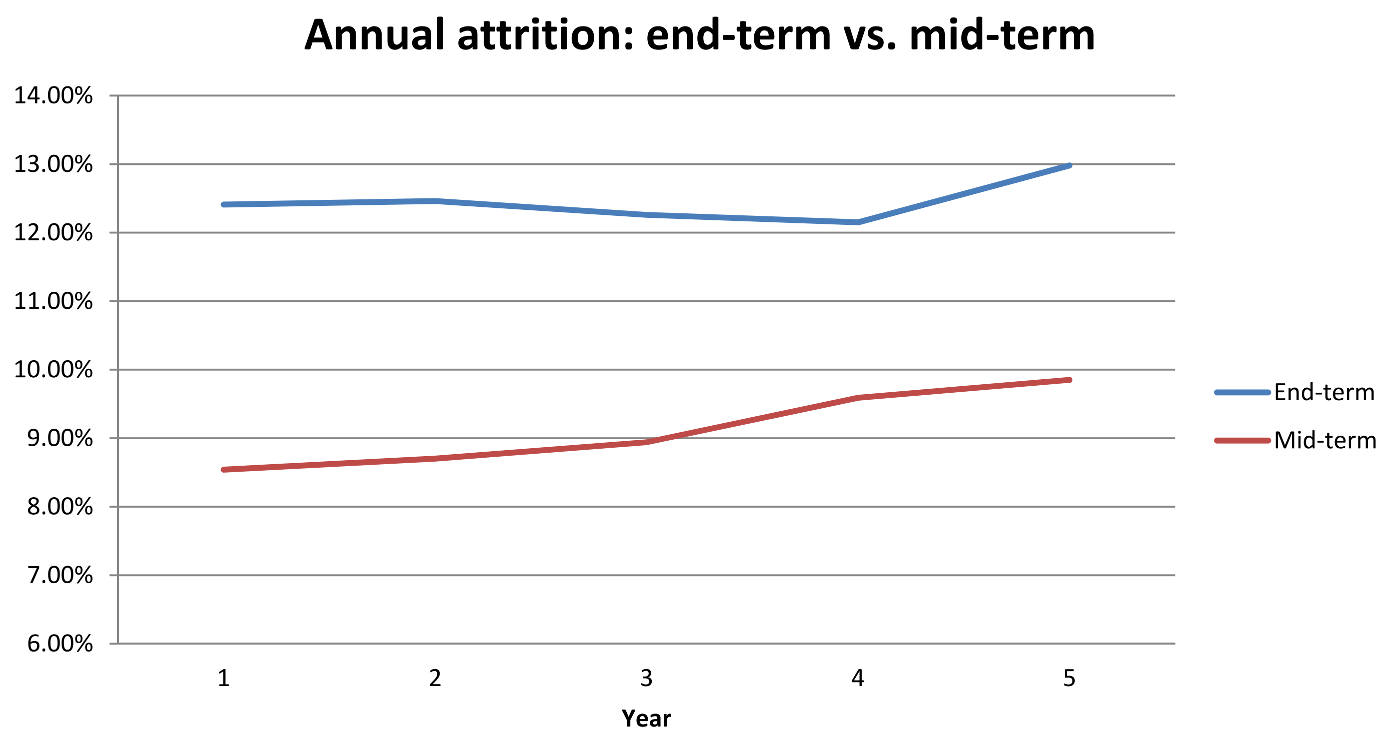
Annual Attrition Summary

Year	Total	Renewed	Non Renewed	Midterm Cancellation	Non Renewal %	Midterm Cancellation %	Retention %
1	197,954	156,477	24,570	16,907	12.41%	8.54%	79.05%
2	201,424	158,794	25,101	17,529	12.46%	8.70%	78.84%
3	201,893	159,080	24,756	18,057	12.26%	8.94%	78.79%
4	205,335	160,688	24,950	19,697	12.15%	9.59%	78.26%
5	211,061	162,875	27,398	20,788	12.98%	9.85%	77.17%

* The data is for illustration purpose.

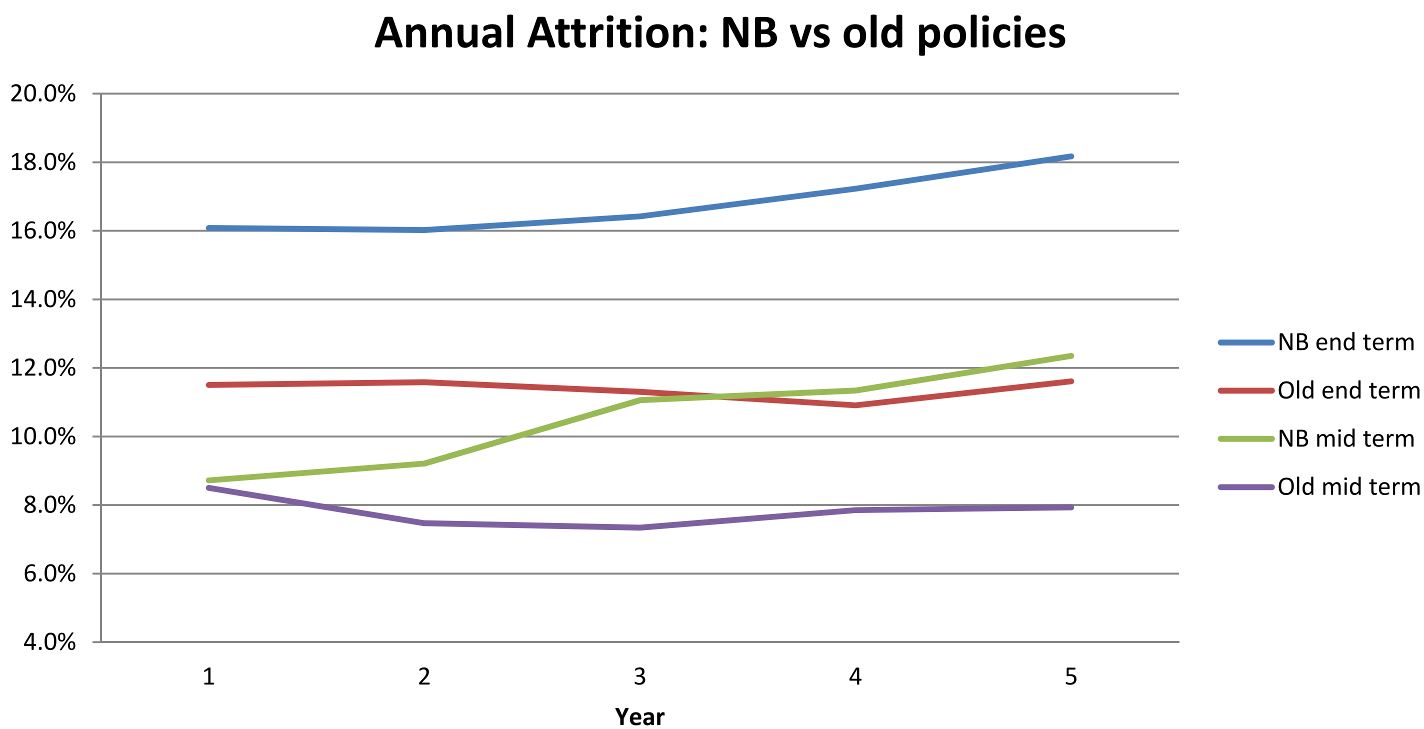
Case Study

Annual Attrition Summary



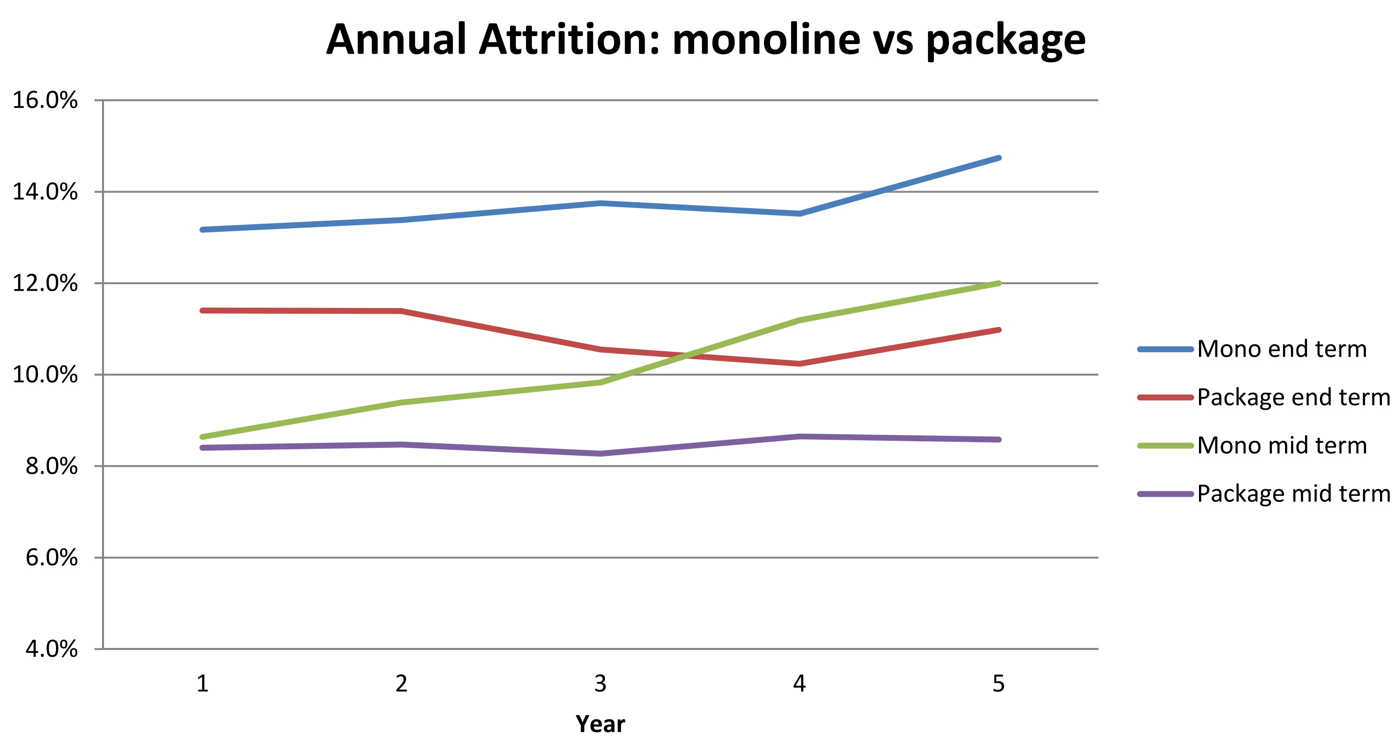
Case Study

Annual Attritions by Policy Age



Case Study

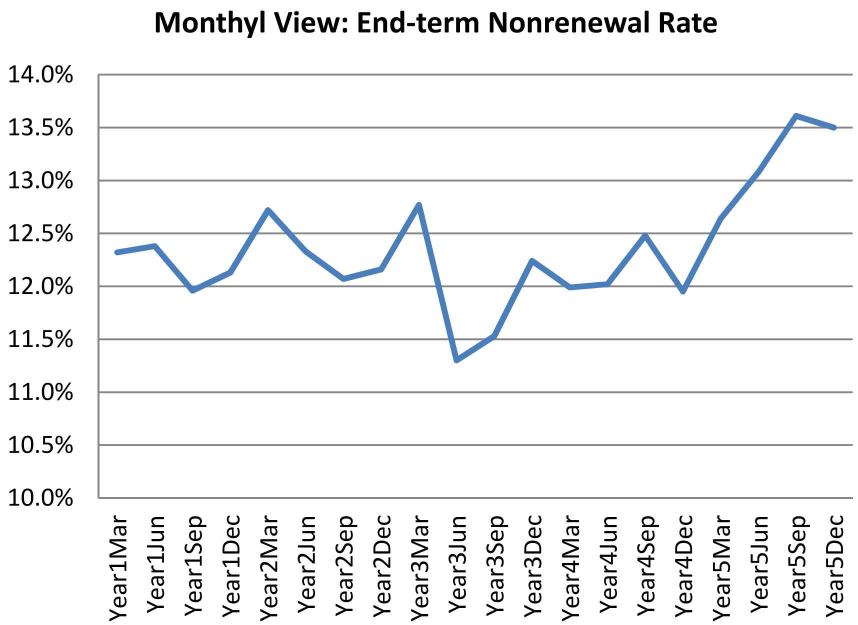
Annual Attritions by Policy Category

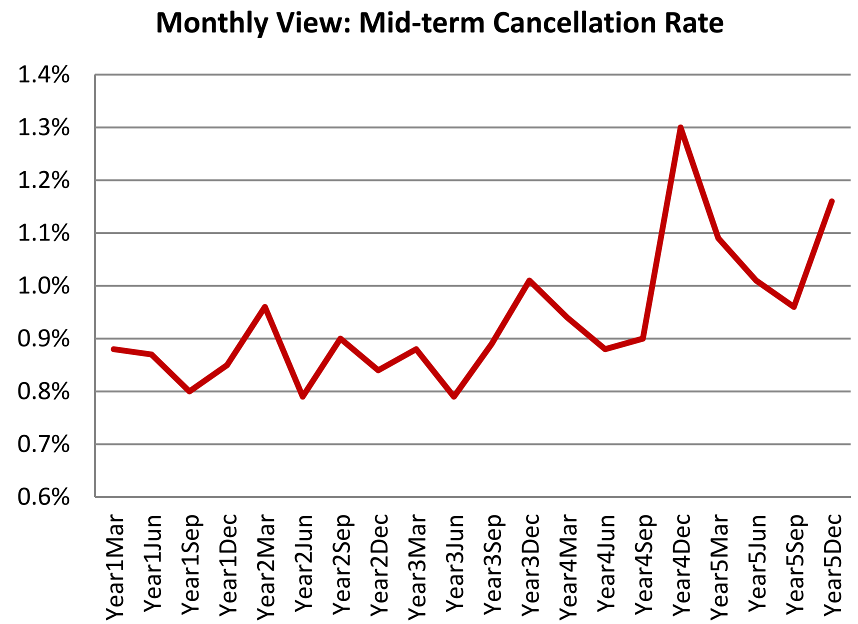


Case Study

Monthly View: March Year1

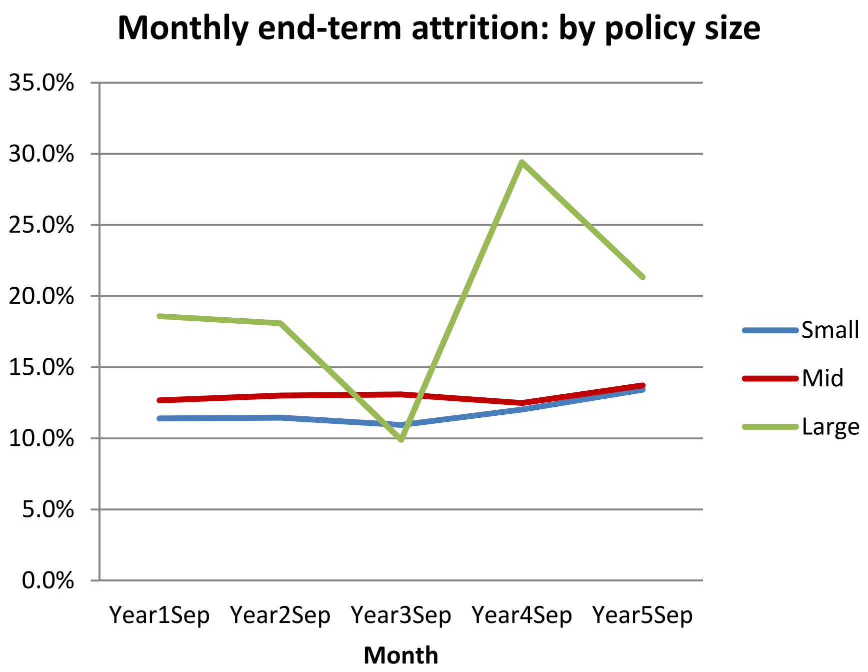
	Active	Attrition	Percent
End-term	16,939	2,086	12.32%
Others	182,160	1,609	0.88%
Total	199,099	3,695	1.86%

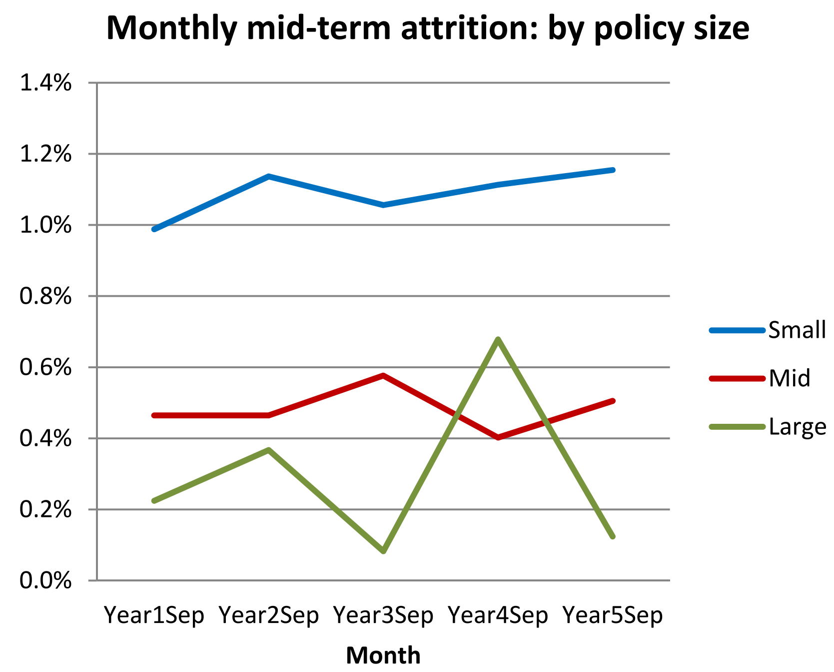




Case Study

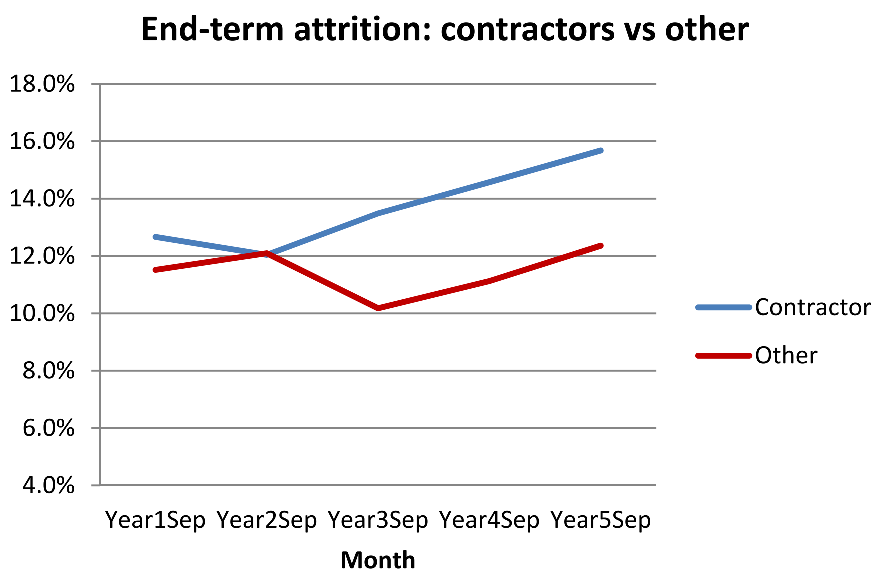
Monthly Attritions by Policy Size

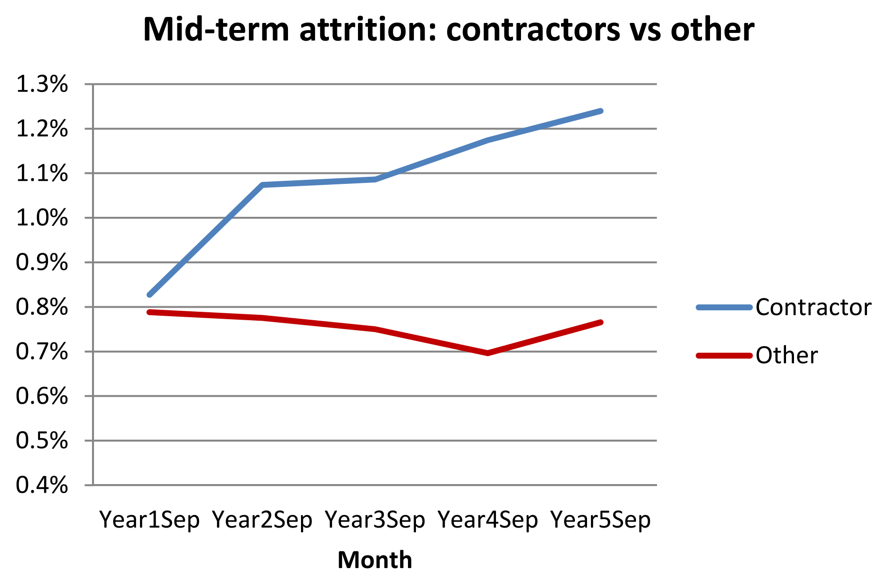




Case Study

Monthly Attritions: Contractors vs. noncontractors





Case Study

Parameter Estimates from Proportional Hazard Models

Analysis of Maximum Likelihood Estimates			
Variable Name	Parameter Estimate	Chi-Square	Pr > ChiSq
Package Indicator	-0.12365	51.77775	<.0001
Rating Change	0.4847	9361.2017	<.0001
Policy Age	-0.00778	1838.8259	<.0001
GDP	-0.02942	58.5243	<.0001

There are about 20 variables plus several interaction terms in the models. Only selected variables are reported.

Case Study

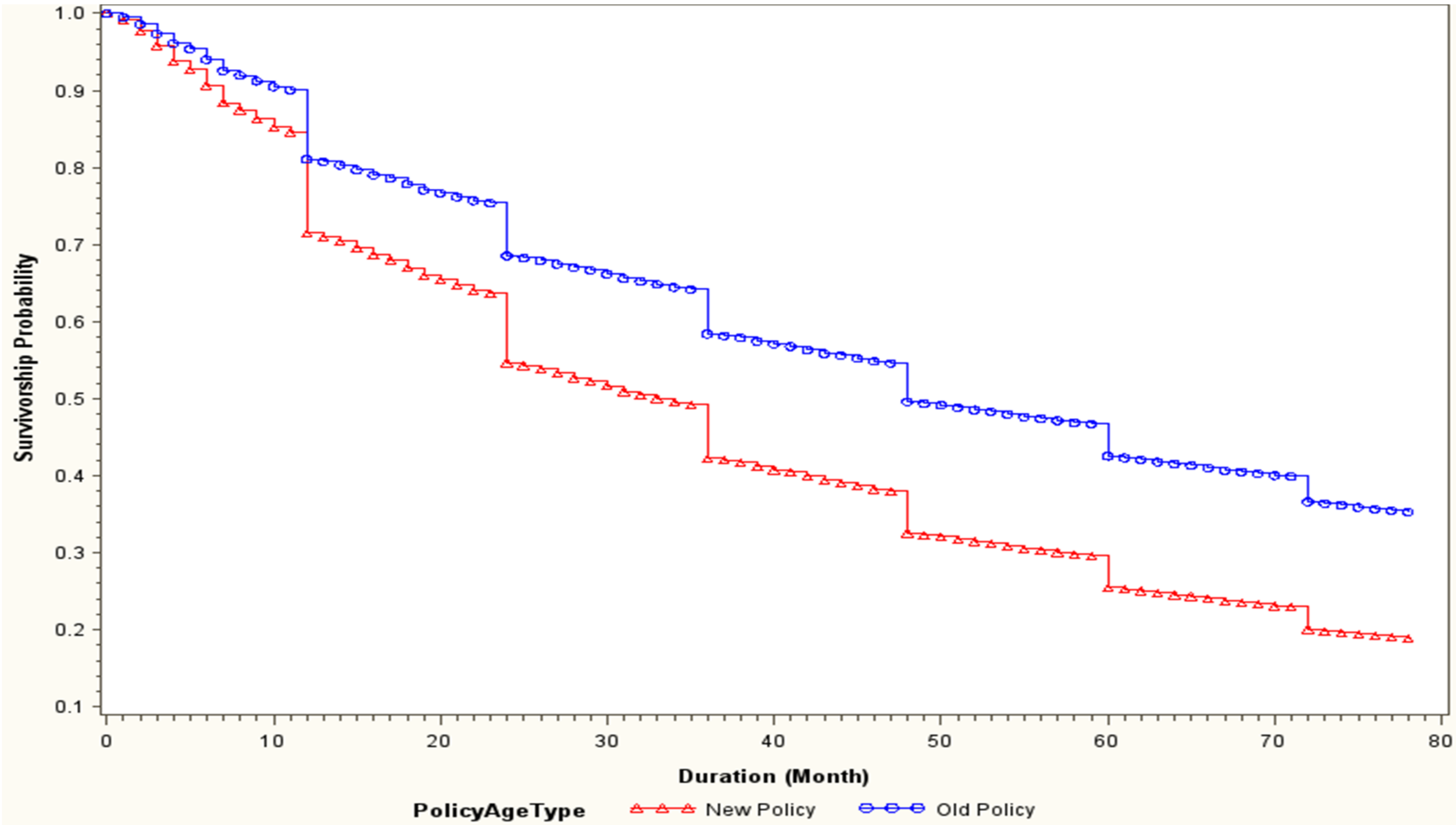
Parameter Estimates from Logistic Regression

Logit Analysis of Maximum Likelihood Estimates

Variable Name	Parameter Estimate	Chi-Square	Pr > ChiSq
Package Indicator	-0.1542	63.52335	<.0001
Rating Change	0.4167	899.4738	<.0001
Policy Age	-0.00691	3590.2861	<.0001
GDP	-0.0245	16.4331	<.0001

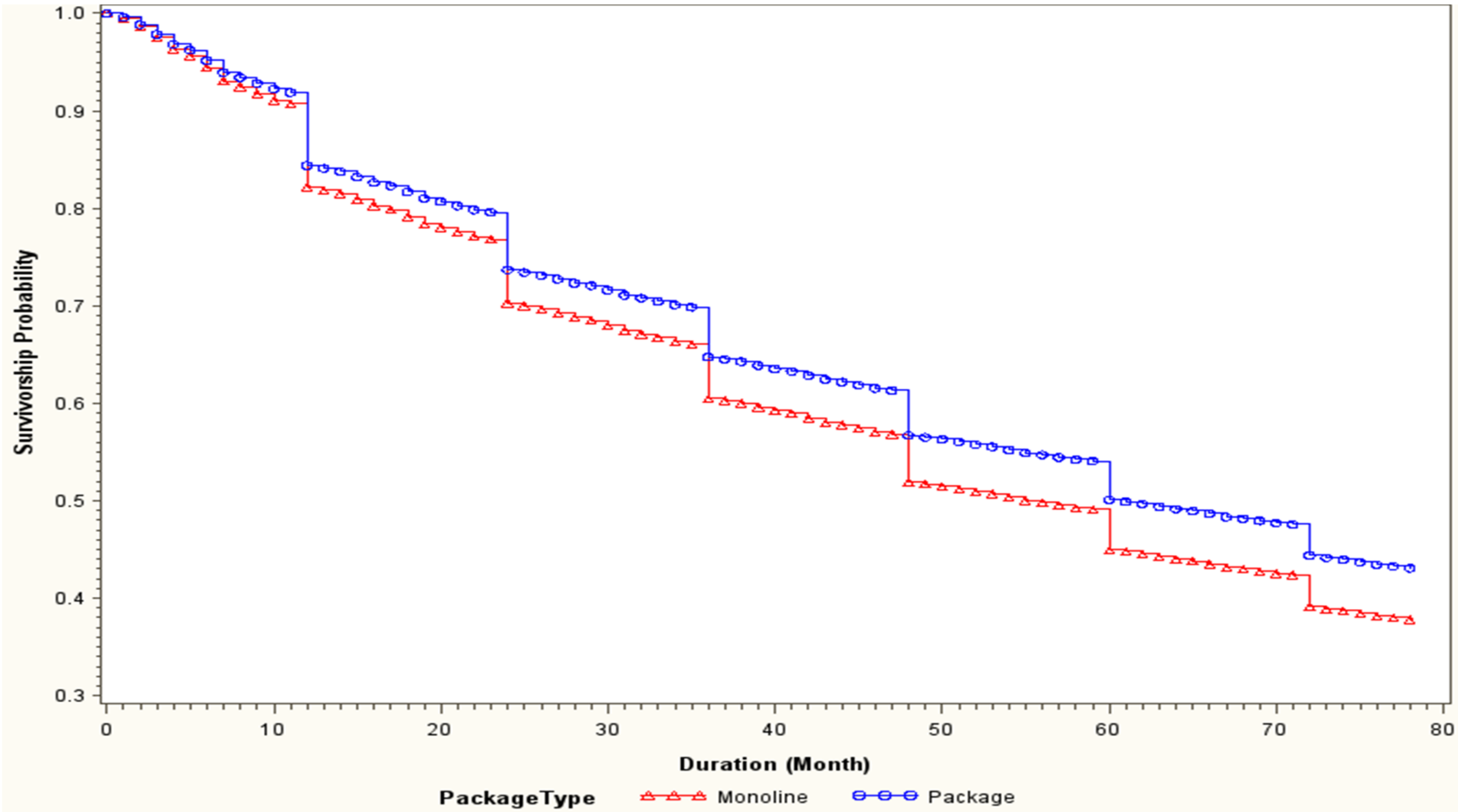
Case Study

Survival Curve for Policy Age



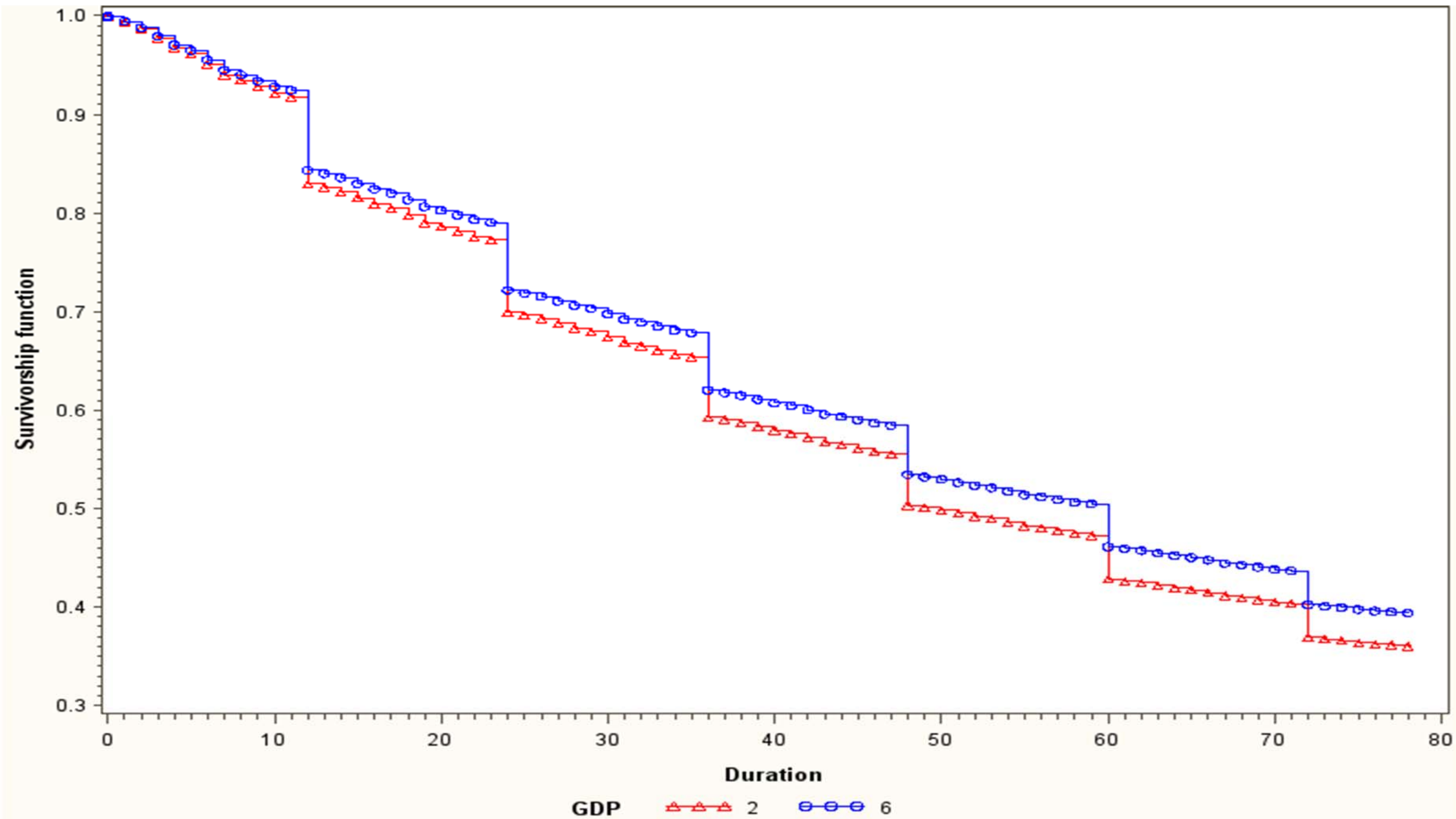
Case Study

Survival Curve for Policy Category



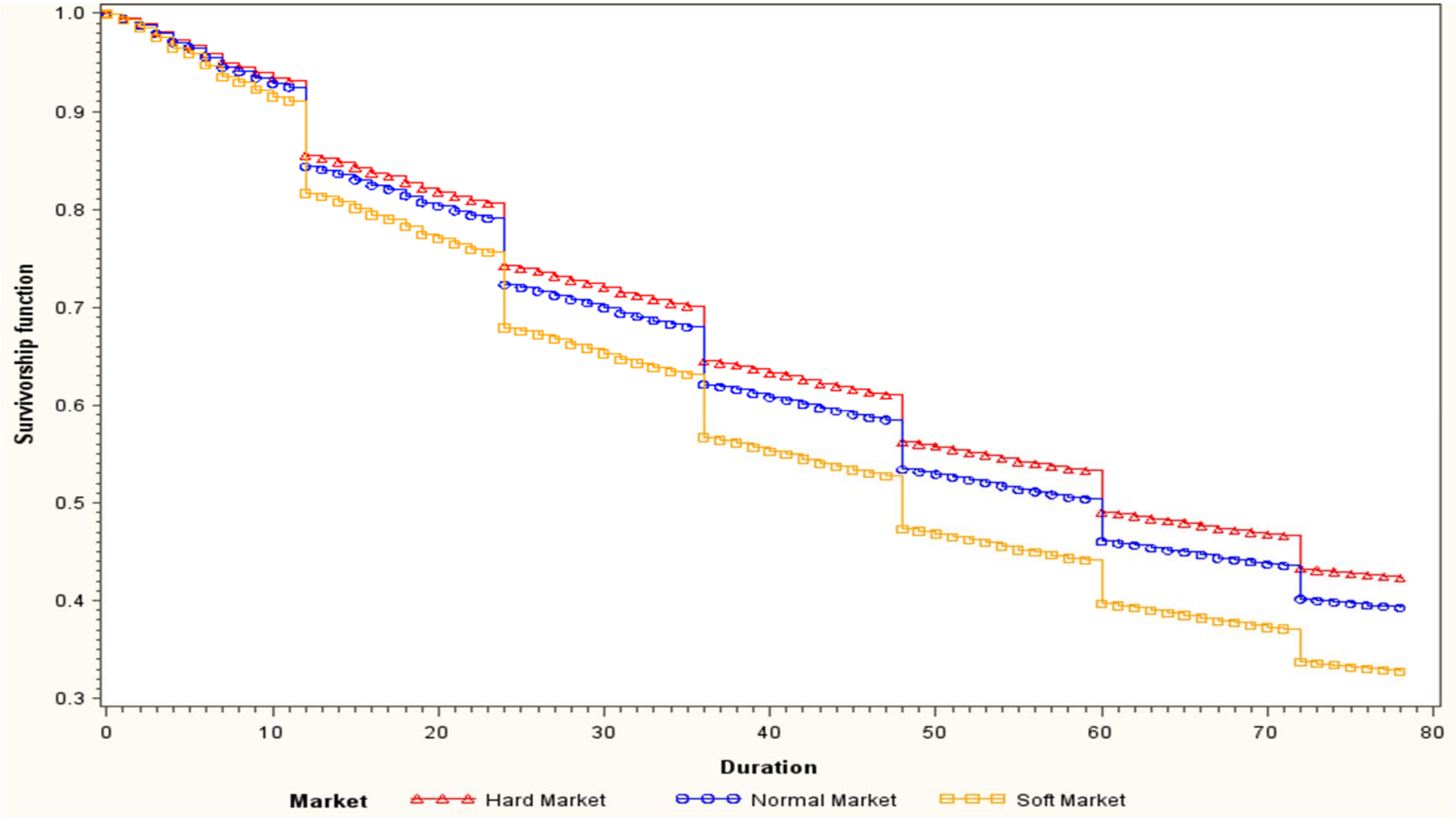
Case Study

Survival Curve for GDP Change (Percent)



Case Study

Survival Curve for Market Condition



Case Study

Validation of the Models (Table)

Out-of-sample Performance of Survival Analysis on the 1 year attrition

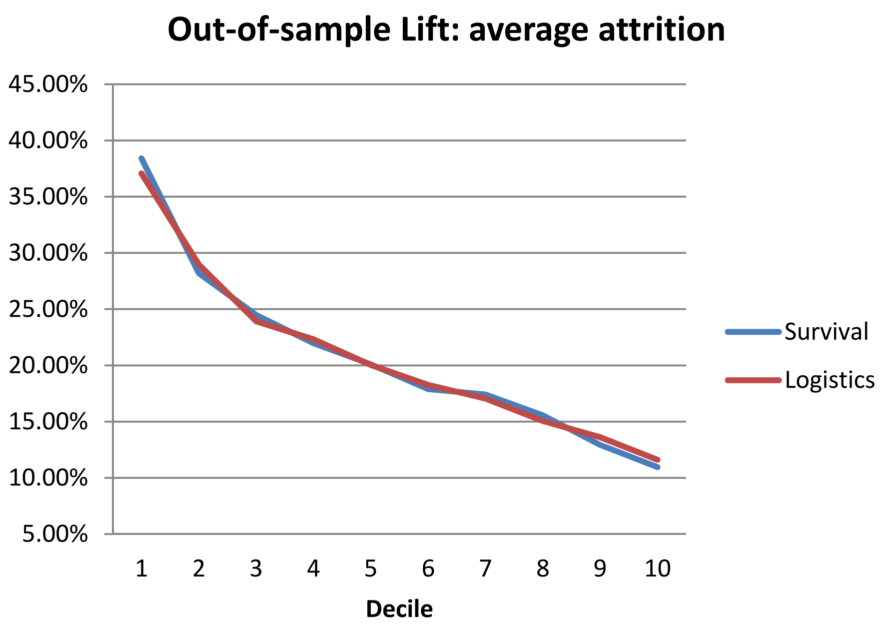
Model Decile	Available Obs	Attrition Obs	Attrition Rate	Cumulative Quantity
1	9,625	3,697	38.41%	9,625
2	9,627	2,714	28.19%	19,252
3	9,624	2,356	24.48%	28,876
4	9,628	2,116	21.98%	38,504
5	9,628	1,935	20.10%	48,132
6	9,626	1,722	17.89%	57,758
7	9,627	1,677	17.42%	67,385
8	9,625	1,498	15.56%	77,010
9	9,628	1,245	12.93%	86,638
10	9,626	1,054	10.95%	96,264
Total	96,264	20,014	20.79%	96,264

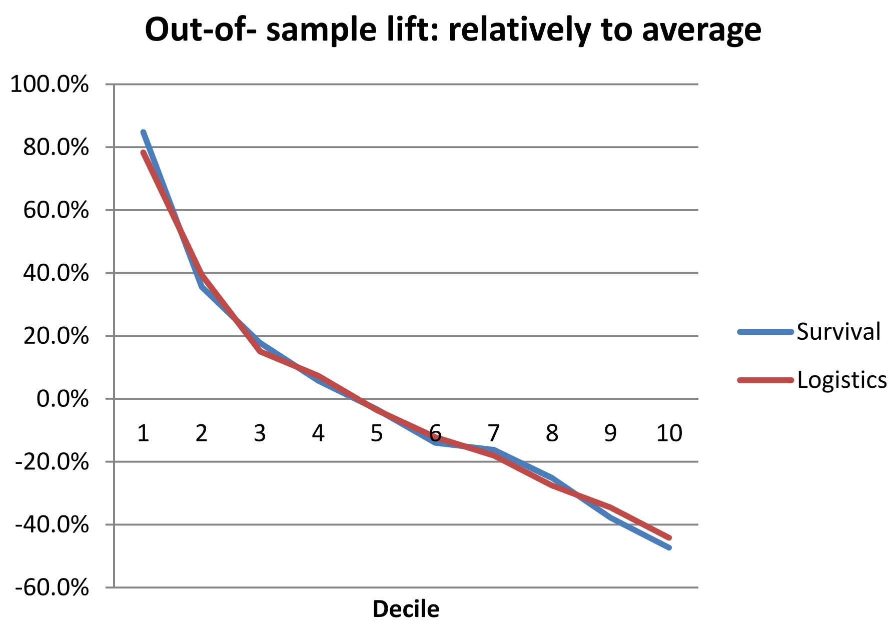
Out-of-sample Performance of Logistic Regression on the 1 year attrition

Model Decile	Available Obs	Attrition Obs	Attrition Rate	Cumulative Quantity
1	9,622	3,567	37.07%	9,622
2	9,630	2,790	28.97%	19,252
3	9,627	2,303	23.92%	28,879
4	9,626	2,148	22.31%	38,505
5	9,628	1,929	20.04%	48,133
6	9,626	1,758	18.26%	57,759
7	9,626	1,641	17.05%	67,385
8	9,626	1,450	15.06%	77,011
9	9,627	1,310	13.61%	86,638
10	9,626	1,118	11.61%	96,264
Total	96,264	20,014	20.79%	96,264

Case Study

Validation of the Models (Lift)

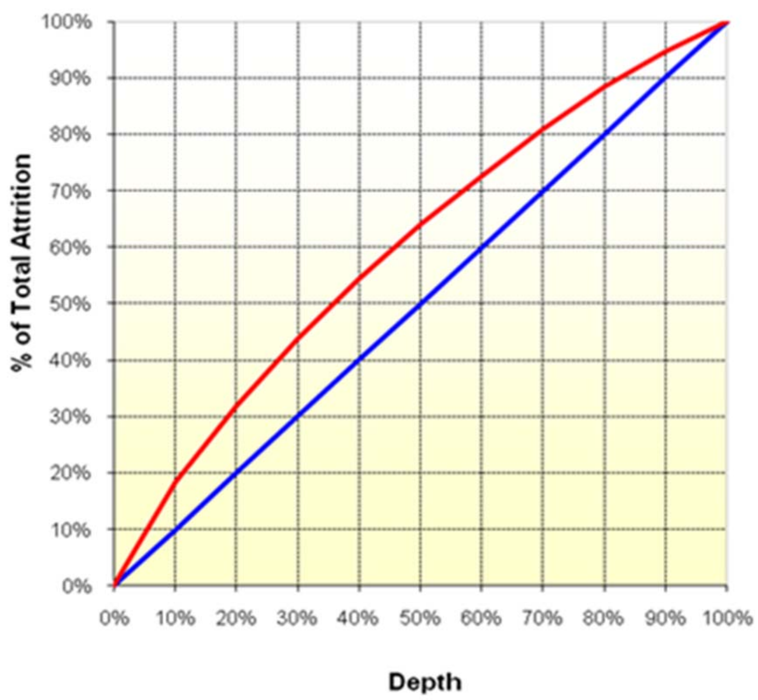




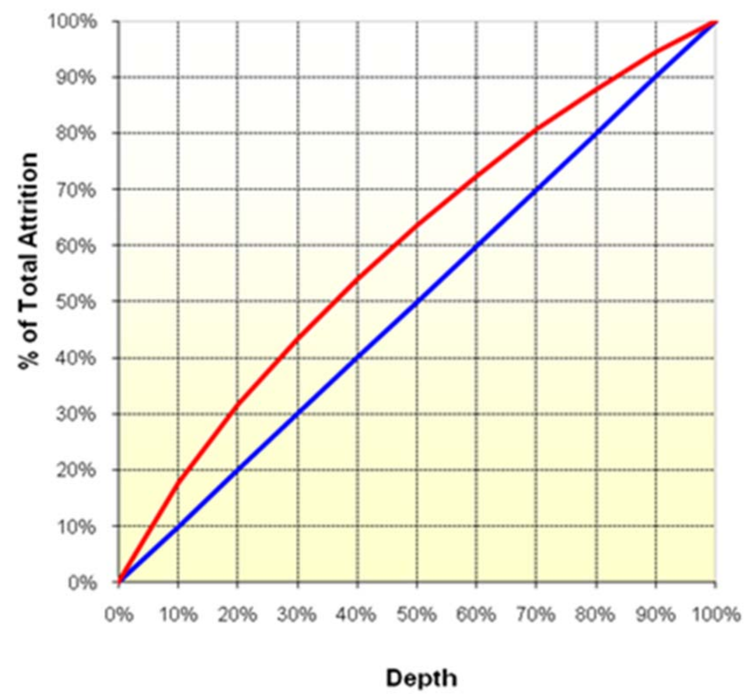
Case Study

Validation of the Models (Gini Chart)

Out –of-sample Performance of Survival Analysis on the 1 year attrition



Out –of-sample Performance of Logistic Regression on the 1 year attrition



Conclusions

- Survival analysis addresses not only whether a policy will leave, but also when it will leave.
- Provide a dynamic insight by utilizing panel data and improve the static view derived from snapshot data.
- Analyze mid-term cancellation and end-term nonrenewal sequentially and simultaneously.
- Able to measure the impacts of time-variant macroeconomic variables on attrition.
- Empirical study does not show significant lift improvement over logistics regression