A Time-To-Event Predictive Model to Identify, Understand, and Strategically Retain Policyholders

Cheng Cheng, Lingtong Guo, Wenjun Wei, Ruoyun Yang, Matthew A. Lanham
Purdue University Krannert School of Management
cheng436@purdue.edu; wei179@purdue.edu; guo565@purdue.edu; yang1833@purdue.edu; lanhamm@purdue.edu

Abstract

Although many insurers are looking at ways to retain policyholders, the problem with most analytics solutions developed is such solutions are not timely nor strategic. Our time-to-event model provides insurers better decision-support by estimating when a customer is likely to leave and why rather than providing just a propensity to churn (e.g. classical churn prediction). We examine features that affect the policy status (e.g. active, cancelled and expired). We also demonstrate the strategic advantage the insurer can gain at retaining their customers by implementing our survival model approach to entice customers to stay rather compared to the baseline propensity model. The audience that will appreciate this work are those looking to extend their binary classification churn models to models that incorporate time or time-to-churn.

Introduction

There's a huge gap between selling to existing customers and searching for new opportunities.

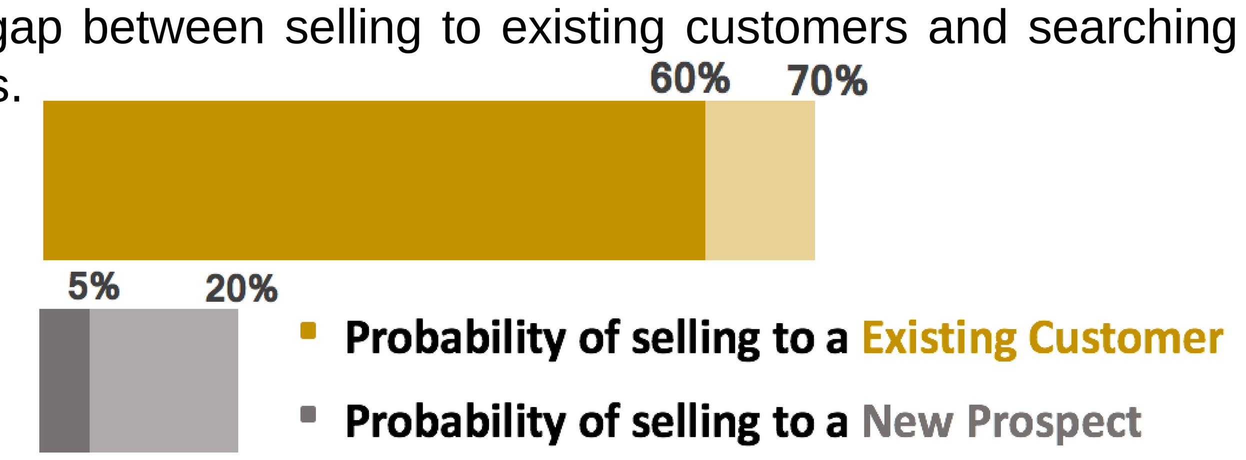

Figure 1. Probability of Selling

In the insurance industry where companies face fierce competition, it is therefore crucial to understand customer retention and churn behavior to sustain revenue growth. This study aims to leverage analytics to answer the following **research questions**:

Who is going to churn? » **Why** going to churn? » **When** to intervene? » **How** to intervene? » **What** are the cost and benefit?

Literature Review

Study	Logit	Survival	Cost-Benefit	Policy Level	Individual level
(Bolancé, 2016)	✓			✓	
(Wang, 2018)			✓		✓
(Haugen, 2016)		✓		✓	
Our Study	✓	✓	✓	✓	✓

Table 1. Literature review summary by method used

Researches on customer churn behavior typically focus on either policy or individual level to examine the causations from data. Our study is novel because we incorporate multiple models and create a reproducible framework.

Data

The dataset has three levels of information: individual (contains basic demographic and household information), policy and claim.

Methodology

Figure 2 shows our stepwise raw data manipulation.

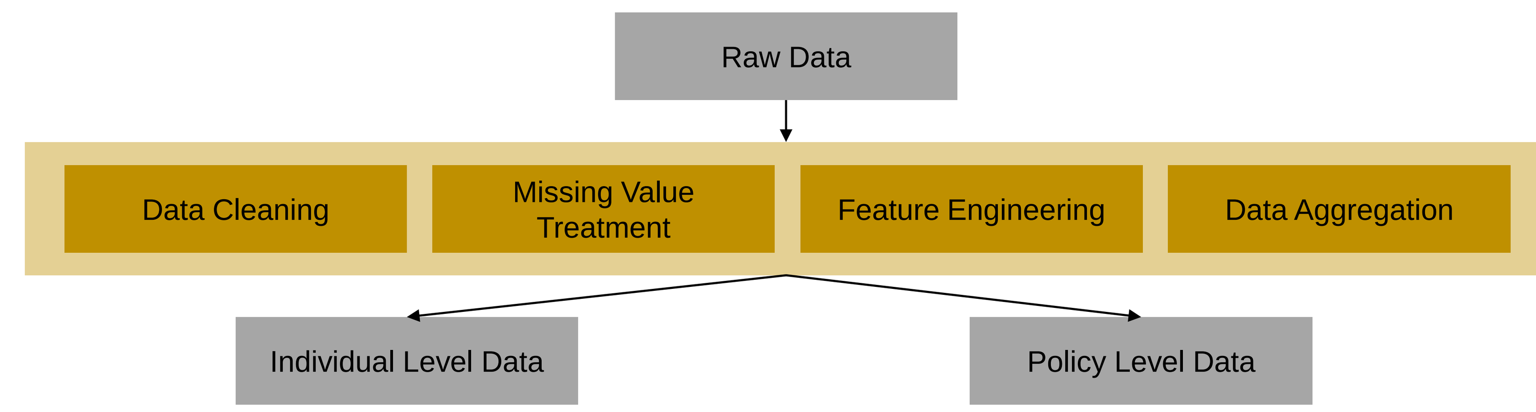

Figure 2: Raw Data Manipulation

Individual-Level Approach

Performing survival analysis to predict policyholder churn behavior and identify relative pattern and risk on an individual base in the insurance industry. Figure 3 outlines individual-level study design, and figure 4 illustrates our method of labelling target and selection of measurement window.

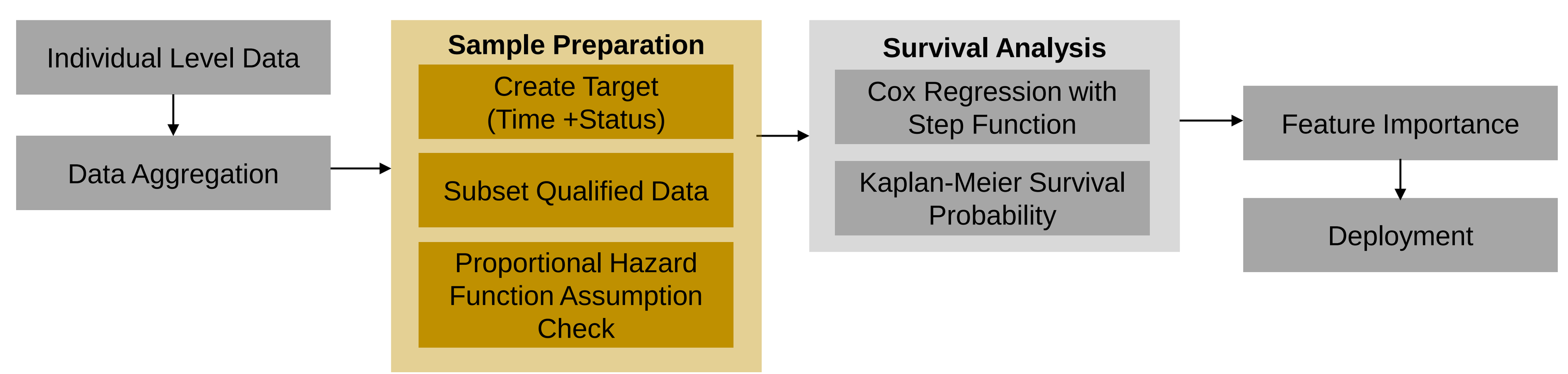

Figure 3: Individual-Level Study Design

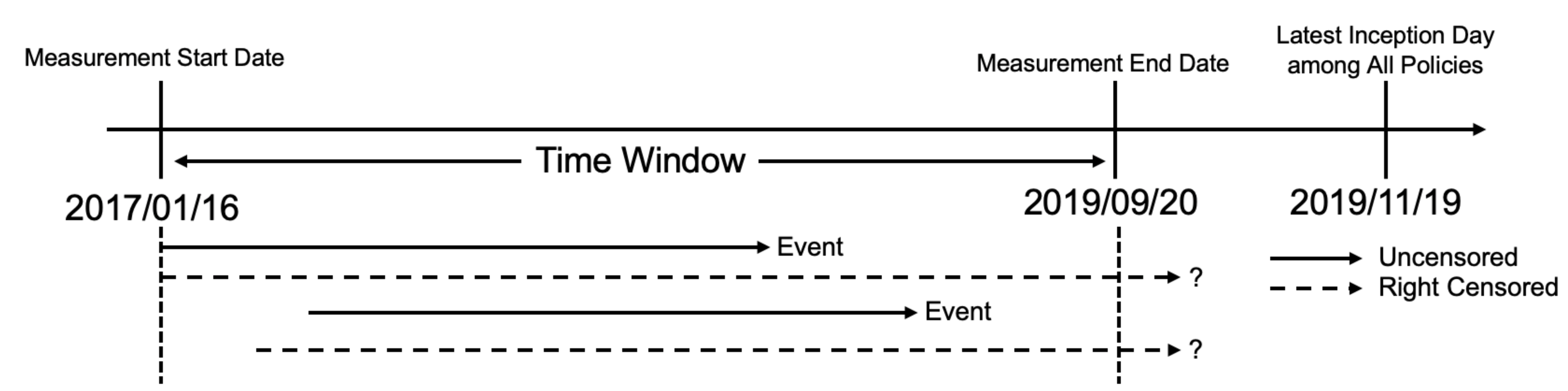

Figure 4: Measurement Window

Policy-Level Approach

Identify active, expired, and cancelled insurance policy, predict the future status of new policy, extract significant features for policy cancellation. Figure 5 outlines policy-level study design.

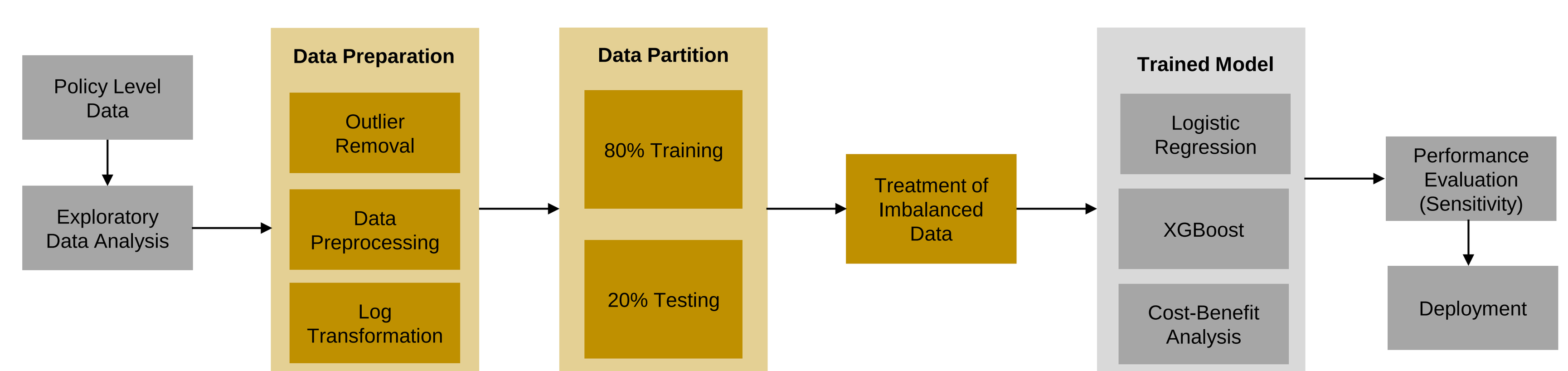

Figure 5: Policy-Level Study Design

Results

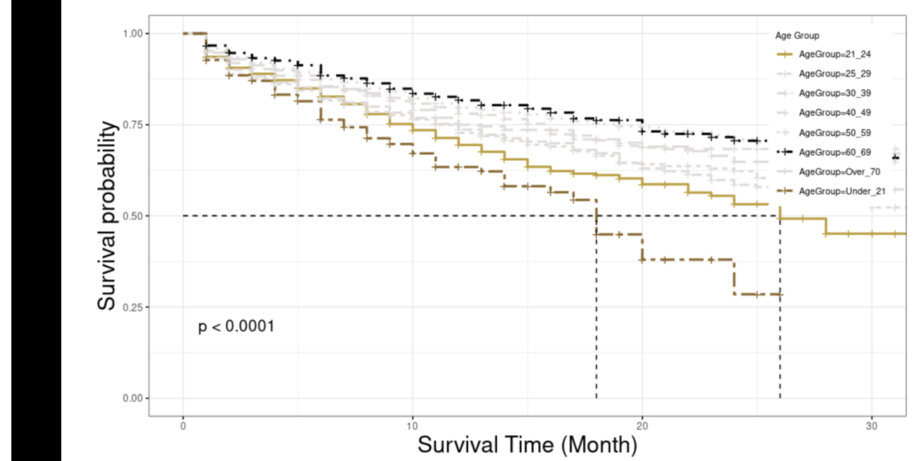

Figure 6: K-M Survival Curve

Survival curve for individuals by age group:

- Individual under 21 has higher probability of experiencing churn behavior within the measurement
- The median survival period for individual under 21 is 18 months

Time	Survival Time			
Strata	0-6 Months	6-12 Months	12-24 Months	>24 Months
Feature ↘	• Age Group (60-69) • Age Group (over 70) • **Auto Policy Count** • **Home Policy Count** • **Child Count**	• Age Group (60-69) • Age Group (over 70) • **Auto Policy Count** • **Home Policy Count** • **Child Count**	• Age Group (60-69) • Age Group (over 70) • **Auto Policy Count** • **Home Policy Count**	• Age Group (60-69) • Age Group (over 70)
Feature ↗	• Age Group (Under 21) • Other State Household • **Claim before first canceled policy**	• Age Group (Under 21) • Other State Household	• Age Group (Under 21) • Other State Household	• Age Group (Under 21) • Other State Household

Model Performance (ROC at 6 months): 81.3%

Table 2: Cox Regression Feature Selection

In the policy-level, we experiment different model with 5-fold cross validation to get the best accuracy:

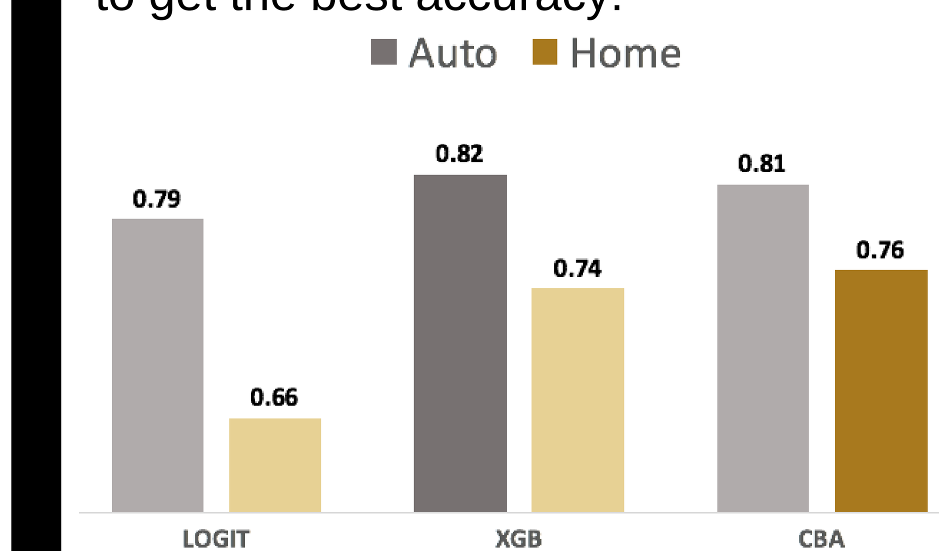

Figure 7: Model Performance

The best overall accuracy:

- XGBoost model on Auto data (0.82)
- Cost-Benefit Analysis (CBA) on Home data (0.76)

Auto Feature	Impact	Home Feature	Impact
Coverage Type Vehicle Year	↘ Less likely to be inactive	Policy Term Home Premium	↘ Less likely to be inactive
Claim Count Auto Premium	↗ More likely to be inactive	Age Group Relationship	↗ More likely to be inactive

Table 3: Selected Feature Impact

Conclusions

- People with age under 21, less auto policy, or less child are likely to churn;
- After understanding which customer group to focus on, companies should focus on features they can make changes on;
- Policy term, premium, coverage type and discount have strong impact on policy cancellation;
- Insurance companies should set a threshold on survival curve to determine the timepoint of intervening policyholders.

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project.