# Game Over: An Application of Customer Churn Prediction using Survival Analysis Modelling in Automobile Insurance

**Joshua James A. Bravante**
School of Information Technology
Mapua University
Intramuros, Manila City, Philippines
jjabravante@mymail.mapua.edu.ph

**Rex Aurelius C. Robielos**
School of Industrial Engineering and Engineering Management
Mapua University
Intramuros, Manila, Philippines
rarobielos@mapua.edu.ph

## Abstract

Customer churn poses a big challenge in financial institutions including the field of insurance. Hence, retention analyses are designed to predict time to churn and identify which among the customers are most likely to leave the company so that retention campaigns and strategies may be done to address the issue. By using 6-year data of individual insurance policyholders from an automobile insurance company in Asia, the study utilized Kaplan-Meier, Log-rank tests and Cox-Proportional Hazard model to understand the risk of churning and to identify significant factors affecting the said event of interest. Demographic variables, including age, marital status, region, as well as policy and claim details, including mortgagee and claim indicator, were found to be significant in the hazard model. Furthermore, results of K-Prototypes, a clustering algorithm used for handling mixed data types, show that customers can be grouped into three (3) clusters and that strategies must be geared towards retaining the young, single and working adults' segment.

## Keywords
Customer churn, Customer relationship management, Survival analysis, Customer segmentation and motor insurance

## 1. Introduction

Customer churn, or the propensity at which clients cease their business relationship with a company, poses a big challenge in various industries not exempting motor insurance (Gunter et al. 2014). Losing customers leads to financial loss because of reduced sales and further leads to an increasing need for attracting new customers (Liguangar 2020). As confirmed by related studies, reducing attrition rate and focusing on existing policyholders are deemed as the more efficient and cost-effective marketing approach which maximizes shareholder's value. Long term customers would be more beneficial and, if satisfied, may provide new referrals. (Shirazi 2018, Ekinci et al. 2012, Vafeiadis et al. 2015, Wadikar 2020). Hence, customer churn retention analyses are designed to predict which customers are about to churn and facilitate an accurate segmentation of the market which allows organizations to target the customers who are most likely to churn with a retention campaign (Zaqueu 2019). With effective churn management, losses due to churn are minimized through prediction and profits are maximized by retaining valuable customers (Kasioglu 2011).

Aside from the entry of digital innovations and the threat of direct competitors, retaining customers within insurance companies is much more challenging as policies are generally renewed every year (Azzopardi and Spiteri 2018). Policyholders can easily switch providers once their expectations are not met (Accenture 2013). Customer experience is greatly affected by several factors such as premium pricing, quality of service, flexibility, and convenience which are the most common considerations which influence customers' decisions to stay or not (Azzopardi and Spiteri 2018, Wuchao 2017).

Due to the unpredictable nature of customers, it is quite a daunting task to predict whether a customer will quit the company or not. For financial institutions including insurance, it is even more complex to identify customer churn due to the sparsity of data as compared to other domains. This requires longer investigation periods for churn prediction (Kaya et al. 2018, Wadikar 2020). Hence, instead of predicting the binary response variable and classifying customers of a specific annual period into churners and non-churners, this research aims to observe the behavior of policyholders to define time before the said crucial event of interest. By doing this, companies would be able to detect the point at which the customer began thinking about leaving and come up with timely recommendations and strategies to prevent this (Havrylovych and Kuznietsova 2019).

Furthermore, customers are not all equally important. Some have higher expectations due to larger expenditure and monetary contribution to the company; hence, they should be given more attention and be considered as the most valuable segment. On the other hand, companies need not allocate more of their resources to those that may be tagged as “least priority” as their contributions are not that greatly significant to the portfolio and production, hardly affecting churn (Singh et al 2016). Consequently, this research also aims to further analyze the subject company’s customers by segmenting them by their characteristics, behavior, and time before churn. This model shall serve various purposes -- from improving targeted actions, refining pricing granularity, embarking on a more tailored customer relationship management journey, to providing a new perspective to look at customers (Zaqueu 2019).

### 1.1. Objectives

Accordingly, the research was set out to answer the following key questions:

- On the average, how long do customers stay with the company?
- What are the significant risk factors affecting a customer's decision to churn?
- Are there differences in churn rate between groups (i.e. sex, age, marital status, type of vehicle, claimant or non-claimant, etc.)?
- Which among the resulting customer groups are considered as the high-valued customers? Which should be tagged as ‘low priority’?
- What are the possible customer retention strategies for each customer segment?

## 2. Methods

### 2.1. Survival Analysis

Survival Analysis is an assemblage of statistical procedures for data analysis (Bogonko et al. 2020). It measures and predict time before a certain event, in this case, churn, or the event in which the policyholder ceases its relationship with TIC.

Individual opportunities to survive for time x is expressed by $S(x) = P(X > x)$. Let $X$ be the continuous random variables, then the survival function is the complement of the cumulative distribution function $S(x) = 1 - F(X)$ where $F(X) = p(X \leq x)$. The survival function is the integral of the probability density function $f(x)$:

$$S(x) = P(X > x) = \int_x^\infty f(t)dt \qquad (1)$$

$$f(x) = -\frac{dS(x)}{dx} \qquad (2)$$

Then if $X$ is the discrete random variables, and can be obtained $x_j$, $j = 1,2,3,\ldots$ with the probability mass function $(p.m.f)\ p(x_{j)}) = P(X = x_j),\ j = 1,2,3,\ldots$ where $x_1 < x_2 < x_3 \ldots$ then the survival function for the discrete variables $X$ is given by:

$$S(x) = P(X > x) = \sum_{x_j > x} p(x_j) \qquad (3)$$

The hazard function of the hold time $X$ is denoted by $h(x)$ and defined as individual probability fails in the time interval $(x, x + \Delta x)$that the individual has lived for time, the hazard function is expressed as:

$$h(x) = \lim_{\Delta x \to 0} [\frac{P(x < X < x + \Delta x | X > x)}{\Delta x}] \qquad (4)$$

The relationship between the hazard function and the survival function is expressed by:

$$h(x) = \frac{f(x)}{S(x)} \qquad (5)$$

On the other hand, the relationship between the hazard rate and the covariate set can be expressed using the Cox's Proportional Hazard model as follows:

$$\ln[h(t)] = \ln[h_0(t)] + \sum_{i=1}^{n} x_i\beta_i \quad (6)$$

$$or \;\; h(t) = h_0(t)e^{\sum_{i=1}^{n} x_i\beta_i} \qquad (7)$$

Where $x_1, x_2, \dots, x_n$ are covariates. $\beta_1, \beta_2, \dots \beta_n$ are the regression coefficients to be estimated. $t$ is time and $h_0(t)$ is the baseline hazard rate when all covariates are zero.

The Kaplan Meier survival function is expressed by:

$$\hat{S}(x_{(j)}) = \hat{S}(x_{(j-1)})\hat{P}[x > x_{(j)} | X \geq x_{(j)}] \;\; (8)$$

Furthermore, a log rank test is used to compare Kaplan Meier's survival curves formed by the following hypothesis:

$H_0$: There is no difference between the survival curves
$H_1$: At least one difference between the survival curves

$$Log\ Rank\ Statistic = \frac{(O_i - E_i)^2}{Var(O_i - E_i)} \qquad (9)$$

with

$$O_i - E_i = \sum_{j=1}^{n}(m_{ij} - e_{ij}) \qquad (10)$$

$m_{ij}$ denotes the number of individuals who experience the event at time $x_j$, and $e_{ij}$ is the value of hope. The null hypothesis will be rejected if log rank statistics ≥ $X^2_{a,df}$ with degrees of freedom $d(f) = 1$ or $p - value < \alpha$.

## 2.2. Customer Segmentation

Insurance policyholders of TIC were grouped into customer clusters according to their significant features and characteristics derived from the results of the stepwise variable selection of the Cox-Proportional model and the time of churn.

K-Prototypes, developed by Huang, is a clustering algorithm used for handling mixed data types. This method is a combination of K-Means and K-Modes. K-Means is a partitioning clustering algorithm, where each cluster is connected with a centroid or central point (mean of points). During training, each object assigned to a cluster with the closest centroid, usually Euclidean distance is used. Number of clusters K should be defined at the beginning and initial centroids are defined randomly (Zaki and Wagner 2014, Qadadeh et al. 2018). A brief summary of the K-Means algorithm is presented as follows:

```
K-MEANS $(D, k, \epsilon)$
$t = 0$
Randomly initialize $k$ centroids: $\mu_1^t, \mu_{2,\dots,}^t \mu_1^t \in R^d$
Repeat $t \leftarrow t + 1$
$C_j \leftarrow \emptyset for\ all\ j = 1, \dots, k$
//Cluster Assignment Step
For each $x_j \in D$ do
$j * \leftarrow arg\ \ min_i\{\|x_j - \mu_i^t\|^2\}$ // Assign $x_j$ to closest centroid
$C_j \leftarrow\ C_{j^*} \cup \{x_j\}$

//Centroid Update Step
For each $i = 1\ to\ k$ do
```

$$\mu_i^t \leftarrow \frac{1}{|C_i|} \sum x_{j \in C_i} x_j$$

$$\text{Until } \sum_{i=1}^{k} \|\mu_i^t - \mu_i^{t-1}\|^2 \leq \ \epsilon$$

K-Modes, on the other hand, is the partitioning-based clustering algorithm which uses simple matching dissimilarity function instead of using Euclidean distance. Modes are used to represent centroids and a frequency-based method is used to find the centroids in each iteration of the algorithm (Jiang et al. 2016, Goyal et al. 2017). Likewise, summary of the algorithm is shown below.

INPUT: Number of desired clusters K, Data objects $D = \{d_1, d_2, \dots, d_n\}$

OUTPUT: A set of $K$ clusters

1. Generate $K$ clusters arbitrarily by selecting the data objects and choose $K$ initial cluster centre, one for every of the cluster.
2. Assign data object to the cluster whose cluster center is near toward it according to Equations (11) and (12)
$$d(X,Y) = \sum_{k-1}^{m} \delta(x_i, y_i) \qquad (11)$$
$$\delta(x_1, yi) = \{\begin{smallmatrix} 0, & x_i = y_i \\ 1, & otherwise \end{smallmatrix} \qquad (12)$$
3. Update the $K$ cluster base on allocation of data objects. Calculate $K$ latest modes of every one clusters.
4. Repeat step 2 to 3 awaiting no data object has changed cluster relationship otherwise some additional predefined criterion is fulfilled.

## 3. Data Collection

This research used administrative automobile insurance data of individual policyholders from 2014 to 2019 of The Insurance Company (TIC), an automobile insurance provider in Asia. Data was gathered from different databases of several key departments of the company and integrated into one coherent dataset. The formulation of variables used was based on initial research on features commonly used in analysis of churn in the industry which includes Azzopardi and Spiteri's (2018), and Lariviere and den Poel's (2004). In particular, these include information regarding customer demographics, vehicle description, policy coverage and characteristics, and claims history. Table 1 summarizes the variables used in the analysis.

Afterwards, data processing and clean-up were done to reduce redundancies and irregularities to assure quality data information. Dummy variables were also created for each categorical variable for modelling. Furthermore, observations with missing values for one or more variables were deleted and excluded in the study. That is, out of 522,826 individual motor insurance policyholders in 2014 to 2019, 425,130 were dropped and 97,696 were left for analysis.

In this study, churn is defined as the occurrence of either one of the following:

i. The event in which the customer decides not to renew his/her insurance policy with the company upon expiry, or
ii. The event in which the customer decides to cancel his/her insurance policy prior to expiry.

To compute for the time T, each observation was followed from origin of time until the last observation date, or the time of churn, whichever comes first. The origin of time is defined as lifetime equal to zero or the time when the customer's insurance policy is first enforced. Lifetime of the individual policyholders were observed until the last observation date, December 31, 2019, which is also considered as the termination time. Time in months were recorded for the survival modelling. Likewise, the status of each observation was observed. If a customer failed to renew or cancelled his/her policy on or before the last observation date, then, CHURN variable would be equal to 1. Otherwise, value of the variable would be 0, which refers to the "right censored" data, which means churn was not observed in the period of study.

Table 1. List of Variables

| Group | Variable Name | Type | Description |
|---|---|---|---|
| Demographics | AGE | Numeric | Age of the policyholder upon insurance application |
| | SEX_CATEGORY | Categorical | Sex of the policyholder (Male or Female) |
| | MARITAL_STAT_CATEG | Categorical | Marital status of the policyholder upon insurance application (Single, Married, Widowed or Annulled) |
| | REGION_CATEGORY | Categorical | Address of policyholder (Region I, Region II, Region III, Region IV) |
| Vehicle Details | BRAND_CATEGORY | Categorical | Brand of the unit (Ford, Foton, Honda, Hyundai, Isuzu, MG, Mitsubishi, Nissan, Suzuki, Toyota, or Others) |
| | RISK_CATEGORY | Categorical | Body type and usage of the unit (Buses, Heavy Truck, Light Commercial Vehicle, Motorcycle, Private Car, PUV/PUJ/Minibus, Taxi/Tourist Car, or Tricycle) |
| | AGE_UPON_ENTRY | Numeric | Age of the vehicle upon insurance application |
| | ZIMVTT | Categorical | Vehicle transmission type (A/T, M/T, C, G, or T) |
| | MORTCODE_CATEGORY | Categorical | Mortgaged indicator; Is unit mortgaged or financed? (Yes, or No) |
| Policy Details | AGENTTYPE_CATEGORY | Categorical | Type of Agent (1, 2, or 3) |
| | RSKTYPE_CATEGORY | Categorical | Type of coverage (Compulsory Third Party Liability only, or Comprehensive Cover) |
| | CNTTYPE_CATEGORY | Categorical | Type of policy (VPC – Private Car, VMC – Motorcycle, VCM – Commercial Vehicle, VLT – Land Transportation Operators) |
| Claim Details | CLAIMCODE_CATEGORY | Categorical | Claim indicator; has unit incurred a claim within the study period? (Yes, or No) |

# 4. Results and Discussions

## 4.1. Exploratory Data Analysis

The final data consists of 97,696 TIC individual customers. 84,098, or 86%, of which have churned (CHURN = 1) whilst 13,598, or 14%, were identified as censored data or those who have not yet churned within the period of study (CHURN = 0). Table 2 shows the distribution of churners and non-churners per inception year.

Table 2. Churners and Non-Churners per Inception Year

| Year | Censored | Churn | Total |
|---|---:|---:|---:|
| 2014 | 808 | 20,043 | 20,851 |
| 2015 | 439 | 14,391 | 14,830 |
| 2016 | 275 | 17,780 | 18,055 |
| 2017 | 2,932 | 15,210 | 18,142 |
| 2018 | 4,110 | 9,372 | 13,482 |
| 2019 | 5,034 | 7,302 | 12,336 |
| Total | 13,598 | 84,098 | 97,696 |

Figure 1 summarizes the distribution of churners and non-churners per categorical variable. It can be seen that churn percentage does not significantly vary for variable SEX (Figure 1.a). On the contrary, graphs of other categorical variables suggest attrition rate varies across category groups (Figure 1.b to 1.h).

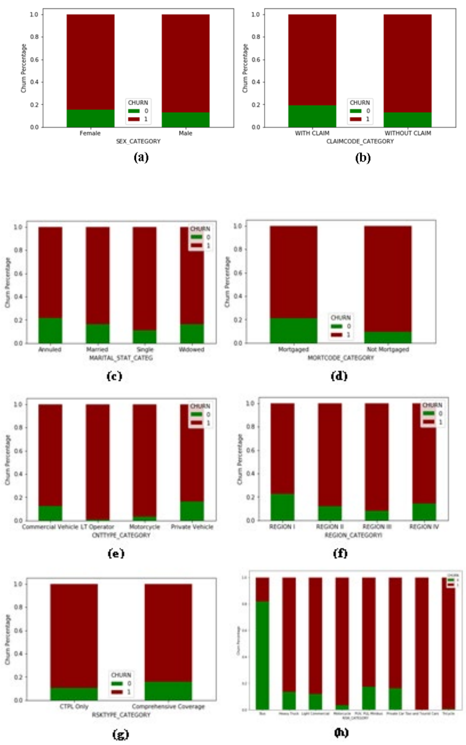


Figure 1. Distribution of Churn and Non-Churners per Categorical Variable

## 4.2. Survival Analysis

The Kaplan-Meier Curve in Figure 2 shows the overall survival rate of the 97,696 individual policyholders in the study. The graph shows a decreasing line suggesting that the longer a policyholder has been with the company, the higher the probability of churn. Furthermore, significant drops in survival probability occur every 12 months which is expected as motor insurance policies are generally renewed annually. In particular, after one year, the probability of survival is as low as approximately 50% due to attrition. This is consistent with customer retention related literature stating that policyholders have high switching probability in the first year after becoming a customer. It was also identified that the early years after becoming a customer is one of the most critical periods in customers' life cycle (Lariviere and den Poel, 2004). Based on the plot, chances of continuing relationship with the company are 40%, 20% and around 15% two, three, and four years after the inception date of the policy respectively.

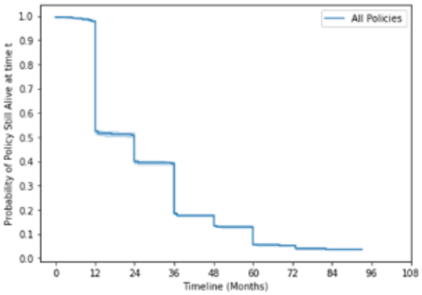
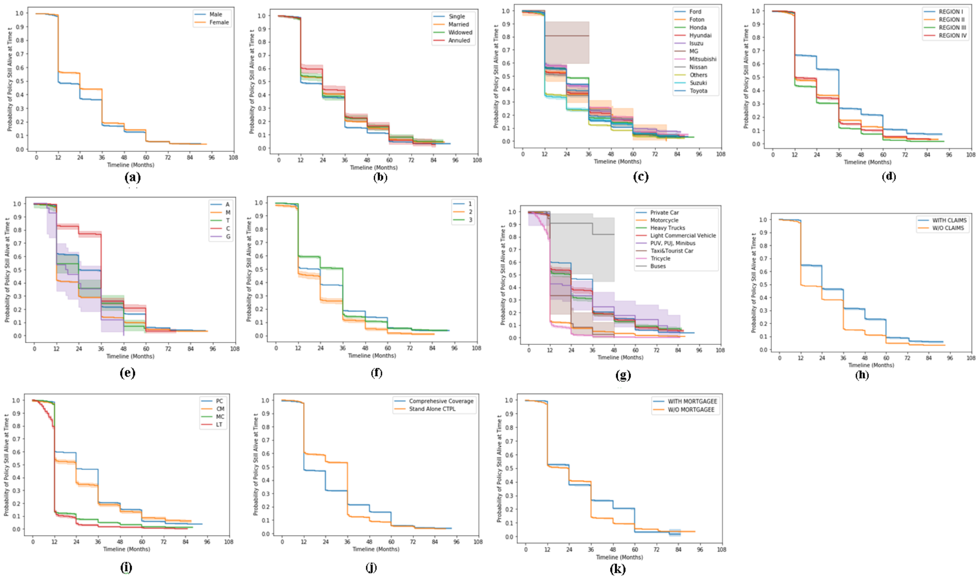


Figure 2. Overall Kaplan-Meier Fitter Curve

The Kaplan-Meier curves for each categorical variable are also presented in Figure 3. Despite the insignificant difference in the proportion of churners and non-churners per sex, it can be noted that across time, males have lower probability of staying with the company as compared to females (Figure 3.a). With regard to marital status, single individuals tend to leave the company earlier than married, widowed and annulled policyholders. In general, individuals who are annulled have the highest probability of survival across time (Figure 3.b). Per Figure 3.d, policyholders living in Region I have the highest probability of survival. On the contrary, those located in Region III have the lowest.

In terms of vehicle brand, in the first three years, brands Suzuki and others have the lowest chance of survival. However, after this period, Suzuki and Toyota units tend to have higher attrition rate (Figure 3.c). Survival curves of vehicle transmission, on the other hand, do not clearly indicate which among the categories have the lowest chance of survival (Figure 3.e). The same trend is observed in mortgaged and non-mortgaged units (Figure 3.k.). For body type and/or usage, in the first three years, private cars seem to have the highest probability of survival. Tricycles, buses and motorcycles, on the contrary, have the highest chances of churn (Figure 3.g.).

Per Figure 3.h, comparing policies with and without claim, the former has higher probability of survival. Furthermore, among the three groups, accounts handled by agent type 2 have the lowest chance of survival. With regard to policy type, motorcycles and public conveyors have higher attrition rates as compared to private cars and commercial vehicles (Figure 3.i). Lastly, in the first three years, compulsory third party liability policies have higher attrition rates as compared to comprehensive insurances whilst the opposite is true after this period (Figure 3.j).



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k)

Figure 3. Kaplan-Meier Fitter Curve Covariates

Log-rank tests were also implemented to determine whether there are significant differences between the survival curves across levels/groups in the categorical variables. This was also done to determine which categorical variables should be included in the proportional hazard model. As presented in Table 3, results of Log-rank tests for all categorical variables indicate that there is at least one significant difference between the survival curves of levels or groups per variable. However, due to parsimony and interpretability, vehicle transmission, body type, and agent type were no longer included in the model.

The study utilized the stepwise selection method to identify the final variables to be included in the model from the initial list and results of the Log-rank tests. It can be noted that the demographic variable SEX is not significant and was further removed in the model.

Despite not having any assumptions regarding the nature or shape of the underlying survival distribution function, the Cox-proportional hazard model requires that the ratio between the hazards of churning of an individual from one group and an individual from another group is proportional at any given time (Saefuddin et al. 2013). Collet (2003) as cited in Saefuddin et al. (2013), and Evrensel (2008) discussed that if this assumption holds, the survival curves would not cross one another. However, visual inspection of the graphs initially presented in Figure 3 suggest that variables marital status, mortgaged indicator, policy type and coverage type violate the said model assumption. Hence, interaction with time for the said variables were also tested for inclusion to address the problem. Through this, estimation of the beta coefficients and analysis of how hazard ratio changes over time would be possible (Borucka, 2013). In particular, MARITAL_STAT_SINGLE*T, MORTCODE_Y*T, CNTTYPE*T and CNTTYPE_ COMPREHENSIVE*T were taken into consideration. As the graphs of the survival functions suggest, interactions of the identified covariates with the time variable are significant in the model. Hence, indeed, the effects of a unit being mortgaged with a bank or financial institution, marital status, policy type and coverage type to survival time are not constant over time. The final Cox-Proportional Hazard model is presented in Table 4.

Table 3. Result of Log Rank Test

| Variable | df | Test Statistic | P-Value |
|---|---|---|---|
| AGTYPE | 2 | 285.72 | 0.005 |
| CNTTYPE | 3 | 12,065.80 | 0.005 |
| RISKTYPE | 1 | 289.98 | 0.005 |
| ZIVCLS | 7 | 12,401.86 | 0.005 |
| CLAIMCIDE | 1 | 1,395.76 | 0.005 |
| MORTCODE | 1 | 435.72 | 0.005 |
| BRAND | 10 | 2,183.85 | 0.005 |
| REGION | 3 | 3,792.30 | 0.005 |
| ZIMVTT | 4 | 3,363.18 | 0.005 |
| SEX | 1 | 351.17 | 0.005 |
| MARITAL_STAT | 3 | 236.44 | 0.005 |

The parameter estimates of the obtained model have varying signs. In particular, age, with claim indicator, single marital status indicator, region indicators Region II and Region III, contract type dummy variables, and comprehensive insurance indicator have positive coefficients. On the contrary, age of the vehicle upon entry, mortgaged indicator, and Region I indicator have negative coefficients.

Increase in age of the policyholder increases the risk of churning, holding all other constant. However, the effect is very small as the coefficient of age is close to zero. On the other hand, the risk of churning of single individuals is 1.28 times higher than married, widowed and annulled policyholders. By analyzing the coefficient of geographical locations, Region II and Region III have increased risk of churning as compared to the baselines category, Region IV. However, if a policyholder lives in Region I, time of stay with the company is longer. A plausible explanation is that the latter category is more urbanized, hence, use of vehicles is more rampant, increasing the need for insurance. Moreover, in this region, it is expected that branches and offices of the insurance provider are more proximate to policyholders as compared to provinces.

Looking into the covariates involving policy and unit details, the higher the age of the vehicle upon entry, the lesser the chance of churn. This is because the level of competition in the market is higher in brand new and relatively younger units. Furthermore, insurance companies prefer the said segments as older units would tend to have lower fair market values and consequently lower premiums and higher chances of claims and mechanical breakdown due to wear and tear. On the other hand, the relative risk of churning of mortgaged units is 0.91 times lower than non-mortgaged units. This is because aside from providing financial protection to the policyholder, the insurer is also protecting the interests of the bank or financing party over the insured unit. In some cases, mortgaged units are automatically issued with insurance policies by banks and/or partner insurance providers. Comprehensive insurance policyholders' relative risk of attrition is 2.67 times higher than those with compulsory liability cover only. This is because there is greater competition on the market for the former and since premium for the latter is the same across all insurance providers. As compared to the baseline category, public conveyors and operators, all contract types namely private vehicles, commercial vehicles and motorcycles have higher risk of churn.

Lastly, the risk of churning of policies with claim is 1.03 times higher than those without claim within their stay, holding all other variables constant. This is consistent with insurers' risk selection practices as policies with losses contribute to lower profitability.

Table 4. Result of Cox Proportional Hazard Model

| Variable | Coefficient | Exp (Coefficient) | SE (Coefficient) | Z | P-value |
|---|---|---|---|---|---|
| AGE | 0.00 | 1.00 | 0.00 | 10.25 | <0.005 |
| AGE_UPON_ENTRY | 0.00 | 1.00 | 0.00 | -4.13 | <0.005 |
| CLAIMCODE_Y | 0.03 | 1.03 | 0.01 | 3.15 | <0.005 |
| MAR_STAT_SINGLE | 0.24 | 1.28 | 0.01 | 17.04 | <0.005 |
| MAR_STAT_SINGLE*T | -0.01 | 0.99 | 0.00 | -14.64 | <0.005 |
| MORTCODE_X_Y*T | -0.01 | 0.99 | 0.00 | -12.56 | <0.005 |
| MORTCODE_X_Y | -0.09 | 0.91 | 0.02 | -5.15 | <0.005 |
| REGION_RCDX_Region I | -0.10 | 0.90 | 0.01 | -8.44 | <0.005 |
| REGION_RCDX_Region II | 0.04 | 1.04 | 0.01 | 3.66 | <0.005 |
| REGION_RCDX_Region III | 0.09 | 1.10 | 0.01 | 7.56 | <0.005 |
| RSKTYPE_RCDX_ComprehensiveCover | 0.98 | 2.67 | 0.02 | 53.08 | <0.005 |
| RSKTYPE_RCDX_ComprehensiveCover*T | -0.03 | 0.97 | 0.00 | -47.20 | <0.005 |
| CNTTYPE_VCM*T | -0.18 | 0.83 | 0.00 | -116.26 | <0.005 |
| CNTTYPE_VCM | 6.02 | 413.54 | 0.06 | 93.69 | <0.005 |
| CNTTYPE_VMC*T | -0.20 | 0.82 | 0.00 | -125.20 | <0.005 |
| CNTTYPE_VMC | 6.40 | 600.79 | 0.06 | 111.26 | <0.005 |
| CNTTYPE_VPC*T | -0.18 | 0.84 | 0.00 | -179.16 | <0.005 |
| CNTTYPE_VPC | 5.82 | 338.47 | 0.06 | 102.38 | <0.005 |

## 4.3. Customer Segmentation

Since the dataset consists of mixed data types, K-Prototypes was implemented to group homogenous customers into heterogeneous clusters. By plotting the Huang Cost Function vis-a-vis the number of clusters, the Elbow method suggests that the observations be assigned into one of either three clusters. The centroids of the three clusters are summarized in Table 5.

Table 5. Centroids per Cluster

| Cluster | Age | Age_Upon_Entry | Time | Churn | Marital_Stat_Category |
|---|---|---|---|---|---|
| 1 | 65.2109 | 4.2752 | 19.4364 | 0.9156 | Married |
| 2 | 41.5113 | 2.2291 | 14.4850 | 0.9372 | Single |
| 3 | 46.6455 | 1.6176 | 43.1657 | 0.7074 | Married |

Cluster 1, or the "elderly churners", generally consists of married individuals whose average age is 65 years old. This segment owns and insures units which are relatively old. Furthermore, the probability of churn is high with average time of stay with the company of around one to two years. Cluster 2, on the other hand, are identified as the "young preferred churners". Customers in this segment mainly consist of single and young working adults insuring relatively new units at point of application. However, this segment tends to leave the company at an early stage, particularly after one to two years of insuring. Lastly, Cluster 3, or the "matured loyalists", are quite older than individuals in Cluster 2. These are married individuals, possibly with their own families, who insure relatively new vehicles at point of entry. Among the three clusters, policyholders in this segment have the lowest chance of churning, and highest chance of staying with the company within a longer period.

Based on the aforementioned profiling, the company may fine-tune customer retention strategies, campaigns, and marketing approaches. As presented in Table 6, Cluster 3 comprises around one-third of the total population. TIC may maintain its existing strategies in this segment as individuals may possibly be not as sensitive as those in other segments. It is important, however, to take note that the company should be consistent in providing the level of service they are currently receiving to prevent the event of churn. Cluster 1, which makes up 23% of the customers, are early churners. Despite the higher age of individuals, relatively old-aged vehicles are still being insured with the company, possibly due to business and/or regulatory requirements. Further studies may be done to explore offering products for units as such, which are possibly not frequently used on the road. However, products should consider the profitability of the vehicles in this segment as these are generally valued lower and consequently, with lower premium yet higher chance of claims or breakdown due to their condition. Most importantly, customer marketing, campaigns, and product development strategies must be geared towards retaining the preferred segment, Cluster 2. In reference to the Kaplan-Meier curve in Figure 3, the significant drop in the probability of survival may be attributed to the early churning of individuals in this segment which comprises the majority of the company's customers. As this group is mainly composed of single and young working adults, customers have higher chances of exploring the market. They have numerous means of checking and comparing offers of the current provider and its competitors which does not exclude online and digital platforms. In addition to that, vehicles insured are mostly brand new up to three years old which are usually the target market of insurance companies in terms of vehicle segmentation. Therefore, review of insurance pricing, special offers and services and product differentiators are crucial which should be coupled with utilization of appropriate tools, platforms and marketing approach.

Table 6. Breakdown of Observations per Cluster

| Cluster | No of Individuals | % |
|---|---|---|
| 1 | 22,200 | 23% |
| 2 | 45,111 | 46% |
| 3 | 30,385 | 31% |
| Total | 97,696 | 100% |

## 5. Conclusion and Recommendation

Customer churn poses a big challenge in various industries including motor insurance (Gunter et al. 2014). Aside from the entry of digital innovations and threat of direct competitors, retaining customers within insurance companies is more challenging as policies are renewed every year (Azzopardi and Spiteri, 2018). Hence, customer retention analyses are being done to predict customers with higher chances of leaving the company and to facilitate market segmentation for development of campaigns and strategies (Zaqueu 2019).

By using 6-year data of motor insurance individual policyholders from an insurance company in Asia, the study utilized Cox's Proportional hazard model to analyze and to identify significant factors in understanding the risk of churning. To address violation of the proportionality assumption for specific covariates, interactions with time were included in the model. From the data, it was noted that almost half of the total customers have already left the company after insuring their vehicles for one year. Demographic variables age and single status were also found to be directly proportional to the individual's risk of churning. Other variables and policy details such as mortgaged indicator, claim code indicator, coverage type, and policy type also differently yet significantly affect the chance of churn. Furthermore, results of K-Prototypes, a combination of K-modes and K-means, show that individual customers of the company can be grouped into three clusters depending on age and churning characteristic. The largest and most significant cluster was identified to consist of young to working single adults who own preferred units yet whose churn rate is very high and survival time is momentary.

These results may be used by insurance providers in fine-tuning marketing approaches for customer relationship management and leverage on this knowledge for retention and acquisition campaigns (Zaqueu 2019).

For further studies, it is suggested to include more demographic variables in the study such as occupation, income, educational attainment for better profiling and analysis of customers. Furthermore, underwriting guidelines, profitability analysis and segmentation of insured risks of companies may also be considered. Future researchers may also opt to compare the survival analysis model to other machine learning techniques to predict time of churn. Insurance providers may also follow the methodology of this research to evaluate effectiveness of retention strategies developed and implemented, as if inclusion of observations in a specific campaign as a 'treatment' or covariate in the survival analysis model.

## References

Accenture, Available: https://insuranceblog.accenture.com/wp-content/uploads/2018/07/Accenture-2013-Consumer Driven-Innovation-Survey.pdf, Accessed on January 29, 2021.

Bogonko, J., et al., Modeling of Average Survival Time for a Loss to be Handled in Insurance Company, *American Journal of Mathematical and Computer Modelling,* 2020.

Borucka, J., Extensions of Cox Model for Non-Proportional Hazards Purpose, PhUse *Annual Conference*, 2013.

Burez, J. and den Poel, D. V., Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department. Expert Systems Application, 2008.

Coussement, K., Benoit, D. F., and Poel, D. V., Preventing customers from running away! Exploring generalized additive models for customer churn prediction. In M.Dato-on (Ed.). The sustainable global marketplace. Developments in marketing science, *Proceedings of the academy of marketing science*, Cham, 2015.

Ekinci, Y., Uray, N., and Ülengin, F., A customer lifetime value model for the banking industry: A guide to marketing actions, *European Journal of Marketing*, vol. 48, pp. 761–784, 2014.

Evrensel, A., Banking Crisis and Financial Structure: A Survival Time Analysis, *International Review of Economics and Finance,* vol. 16, 2008.

Goyal, M., et al., A Review on K-Mode Clustering Algorithm, *International Journal of Advanced Research in Computer Science,* 2017.

Gunter, C., Tyete, I., Aas, K., Sandnes, G., and Borgan, O., Modelling and Predicting Customer Churn from An Insurance Company, *Scandinavian Actuarial Journal*, vol. 1, pp. 58-71, 2014.

Havrylovvych, M. and Kuznietsova, N., Survival Analysis Methods for Churn Prevention in Telecommunications Industry, vol. 2577, 2019.

Lariviere, B. and Poel, D. V., Customer Attrition Analysis for Financial Services Using Proportional Hazard Models, *European Journal of Operation Research*, vol. 157, pp. 196-217, 2004.

Lariviere, B. and Poel, D. V., Investigating the role of product features in preventing customer churn, by using survival analysis and choice modelling: The case of financial service. Experts Systems with Application 27, 2004.

Qadadeh, W., et al., Customer Segmentation in the Insurance Company (TIC) Dataset, *INNS Conference on Big Data Deep Learning*, 2018.

Saefuddin, et al., Survival Analysis of Customer in Postpaid Telecommunication Industry. *Indonesian Journal of Statistics*, vol. 18, no. 1, pp. 1-10, 2013.

Shirazi, F. and Mohammadi, M., A big data analytics model for customer churn prediction in the retiree, *International Journal of Information Management*., 2018.

Spiteri, M. and Azzopardi, G., Customer Churn Prediction for a Motor Insurance Company, *13th International Conference on Digital Information Management,* 2008.

Wadikar, D., Customer Churn Prediction, 2020.

Zaqueu, J., Customer Clustering in the health Insurance Industry by Means of Unsupervised Machine Learning, 2019.

Zaki, M. and Wagner, W. Data Mining and Analysis Fundamental Concepts and Algorithms, 2014.

Zhang, B., et al., Application of Survival Analysis for Predicting Customer Churn with Recency, Frequency, and Monetary, 2017.

## Biographies

**Joshua James A. Bravante** has been with the insurance industry for almost 6 years, both in life and non-life. Currently, he is the Technical Assistant and Claims Statistician of one of the leading non-life insurance companies in the Philippines, Malayan Insurance Co., Inc. He also worked as an actuarial assistant under Philippine Prudential Life Insurance, Co. Inc. Due to his passion and interest in research and analytics, he graduated from Southern Luzon State University and holds a degree in BS Mathematics, minor in Statistics and is a recent graduate of the Master in Business Analytics program of Mapua University, Manila, Philippines.

**Rex Aurelius C. Robielos** is the Department Manager of Operations Research Group at Analog Devices General Trias. Before joining Analog Devices, he was the Dean of the School of Industrial Engineering and Engineering Management at Mapua University. He has a BS in Applied Mathematics from the University of the Philippines Los Baños, and a Diploma and MS in Industrial Engineering from the University of the Philippines Diliman. He is pursuing Ph.D. in Industrial Management (candidate) at National Taiwan University of Science and Technology in Taiwan. He is the Director of Human Factors and Ergonomics Society of the Philippines and the Philippine Institute of Industrial Engineers.