CUSTOMER ATTRITION MODELING AND FORECASTING

By

Zehan Xu

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Engineering Management

Charlotte

2019

Approved by:

____
Dr. Tao Hong

____
Dr. Shaoyu Li

____
Dr. Linquan Bai

ABSTRACT

ZEHAN XU. Customer Attrition Modeling and Forecasting.
(Under the direction of DR. TAO HONG)

Customer relationship management has shown its critical role in the success of a company in today's increasingly competitive service industry. While modern machine learning techniques are widely adopted due to their advantages in working with extensive databases, organizations with little customer information and low-quality data have limited choices when analyzing customer data to retain existing customers. This thesis proposes two approaches to model and forecast existing customer attrition, survival analysis, and regression analysis. The proposed methodologies are demonstrated through customer data from 10 retail energy companies at different data quality. Results from both proposed models show superior performance in terms of Mean Absolute Error, Mean Squared Error and Mean Absolute Percentage Error, compared to a commonly used non-parametric model that forecast attrition rate based on the average of known customers.

DEDICATION

To My Family.

ACKNOWLEDGEMENTS

Dr. Hong is my mentor to forecasting and my advisor during my master study in the Engineering Management program. I wasn’t the best behaving student at the time, but he accepted me into his group and guided me both academically and personally. He introduced me to a new area of study and provided the necessary tools and environment for me to succeed in it. I would like to express my sincere gratitude to Dr. Hong for his trust and guidance through my master program.

I also want to thank my committee members, Dr. Shaoyu Li and Dr. Linquan Bai for their generously offering of time and guidance in preparation to this thesis.

Finally, I want to express my deepest love and thank to my parents: Ming Xu and Qingrong Li, and my girlfriend: Yangyang Xi, they provided financial and emotional support as well as motivation for me to continue my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1: INTRODUCTION

1.1. Customer Relationship Management

The competition between companies becomes increasingly intense as the market matures, more companies have realized their most valuable asset is the existing customer base. This evolution caused customer retention becoming a critical factor in companies' strategy. [1] When the event of customer loss happens, known as customer attrition or customer churn, companies are not only losing potential future sales but also creates the need to acquire new customers, which is about six times more expensive than customer retention. [2] In this context, a small improvement in customer retention can result in a significant cost saving, which means increased profit. [3] The following three rates are often referred to when analyzing customer attrition. Customer attrition rate is a calculation of the percentage of customers moves out of a company or organization in a specified time frame. Customer retention rate, a complimentary figure to attrition rate, is a calculation of the percentage of customers retained during the specified time frame. Customer acquisition rate, on the other hand, is the percentage of customers acquired during the specified time frame. All three are calculated based on the number of customers at the beginning of the specified period. [4]

The need for customer retention causes companies to adopt customer relationship management (CRM) to build closer relationships with customers. [5] CRM can be viewed as the development of a customer-oriented culture and the creation of a strategy for acquiring, retaining and enhancing the profitability of customers with information technology. [6] Various industries depend heavily on customer relationship management, namely insurance, financial service, telecommunication, and retail, etc. For the organization, successful implementations of CRM can increase customer retention, loyalty, and profitability. For customers, successful application of CRM means customization of the services and products, as well as higher quality services and products. Thus, there are mutual benefits for both the organization and the customers to implement CRM in organizations. [7]

1.2. Survival Analysis

Customer attrition analysis and forecasting are crucial parts of effective CRM. There are two types of attrition; voluntary attrition refers to a customer's decision to terminate the connection with the organization or switch to another company or service provider. Voluntary attrition is usually the focus of attrition modeling since it happens due to changes in the company-customer relationship, which can be controlled by the company. On the other hand, involuntary attrition refers to a decision made due to customers' circumstances such as death and relocation, etc. It is usually beyond the control of the company and therefore excluded from the attrition model if possible. [3]

In recent decades, companies experienced an explosion of data collection and storage, resulting in an ever growing database containing customer-related data, "big data." As more information on individual customers being collected, customer-level attrition analysis (customer profiling) is made possible and can help create comprehensive customer retention strategy. [6] This change also requires modeling techniques to process a massive amount of data efficiently. However, this change in research focus requires bigger data sets and higher data accuracy, models used in recent studies are relying on extensive data set to provide a valid conclusion. Smaller data sets with few covariates still existing among organizations, such as startup companies, new programs with the short history, low resolution data gathering, customer privacy requirement, etc. In this paper we will be focusing on forecasting existing customer count under the assumption of no covariate is available and compare forecast accuracy to higher resolution data.

CHAPTER 2: LITERATURE REVIEW

2.1. Survival Analysis

Survival analysis is used initially to study the time until death in medical applications. Since then its usage has been extended to examine the length of time until an event occurs, some application includes time to medical events in clinical research, time to system failure in reliability engineering, and time to customer attrition in the service industry, etc. [8] Take service industry as an example, customer information, promotion, and time to the customer attrition, etc. are recorded. [5] Survival analysis is then used to identify significant characteristics causing the attrition event and forecast customers attrition by modeling the time to customer attrition with the data gathered.

Among previous research in literature, attrition models are tailored to the need of a specific company or industry. Because every company's situation is unique, a different model is selected to incorporate the task in need. The critical factors of model selection are the customer characteristics and operating environments. Operating environment includes a company's industry, data availability, and data quality, etc. The difference in the operating environment is reflected in the data preprocessing and overall forecasting precision. Customer characteristics include customer behavior, customer perception, and customer demographics, etc. [3] While not all customer characteristics are available and useful in every situation, therefore, feature selection and attrition model selection are widely used in the past paper. [9] [10] [11] There are two approaches to attrition, one being the statistical method, which is a process for estimating the relationship between dependent and independent variable(s). This approach is easy to implement and can produce good performance after proper data transformation. Another method is data mining, it is powerful and can extract meaningful information from large databases without intimate knowledge of the data.

Different modeling techniques are developed based on the assumption of causes behind the event occurrence; they can be catalogized under three approaches: parametric, non-parametric, and semi-parametric, the appropriate method is selected based on forecast precision goal, available data, and understanding of the data.

2.1.1. Parametric Approach

The parametric approach assumes that the underlying customer survival probability with respect to time follows a specific known probability distribution. This distribution can be the exponential, Weibull, and log-normal distributions, etc. [12] This approach requires intimate prior knowledge of the application since the results can be misleading when fitting in correct distribution to the data. Model parameters in these settings are usually estimated using an appropriate modification of maximum likelihood. While often used in medical researches and engineering disciplines with established knowledge of the causes behind the event, parametric models are not commonly used when analyzing customer attrition since human behavior is everchanging and the assumption about the underlying distribution of customer survival probability does not hold the in service industry.

One of the reason parametric approach is not often selected when analzing customer attrition is the difficulty to incoorprate covariates. The end products of customer attrition analysis in the service industry are usually in forms of customer specific promotion recommendation which means identify significant attrition-causing covariates.

2.1.2. Semi-Parametric Approach

Cox proportional hazards regression model (semi-parametric) is often used when the model needs to incorporate covariates. The Cox regression model provides useful and easy to interpret information regarding the relationship of the probability density function to customer survival function (the hazard function) to predictors (information on customers). By using the cox model, a linear relationship between the base hazard function and the predictors is assumed, the hazard ratio comparing any two observations is constant over time if the predictor variables do not vary over time. This assumption is referred to as the proportional hazards assumption and checking this assumption is the first step of a Cox regression analysis. [3]

Cox proportional hazards regression model is by far the most popular model in survival analysis of customer attrition and is often used as a benchmark model. [13] The basic cox model assumes constant hazard ratio effects by the covariates. However, certain factors may be critical to distinguish between customers during a period, but they may become irrelevant as a discriminating factor later on. This changing effect of covariates over time can be overcome by incorporating time-dependent parameters and potential time-dependent covariates into the Cox model. [5]

2.1.3. Non-Parametric Approach

For the non-parametric approach, the Kaplan-Meier estimator is often used when indexed total customer count is available with little information on individual customers. It estimates retention probabilities as a function of time and can obtain univariate descriptive statistics on time to attrition. [14] This approach is useful when the sample size is large, and the time interval is precise since Kaplan-Meier estimator considers censored data at the end of each time interval. However, if the time interval is crude, then life table method (nonparametric) is more appropriate since it assumes censored data is distributed uniformly across each time interval. [15]

Logistic regression analysis is a type of probabilistic statistical classification model. [6] [9] [16] [17] [18] It can produce a binary forecast of a categorical variable (e.g., customer attrition) which depends on the independent predictor variables (e.g., customer features). Logistic regression analysis can be applied to the data after proper data preprocessing, it usually cannot produce an as good performance as the techniques mentioned above due to its regression nature. A particular limitation of logistic regression is that it assumes that the functional form of the relationship between the (log-transformed) target variable and the input variables is known and. [19] Similar to Decision Trees, logistic regression is usually not used alone but instead combined with other models or used as benchmark model.

2.2. Data Mining Methods

Data mining is continuously evolving due to the increased computerization of business transactions, improvements in storage and processing capacities of computers, and advances in discovery algorithms. [20] As the amount of data available increases, data mining is preferred by many companies as its advantages on working with a extensive database. The following is four well established and accessible techniques for customer attrition prediction. [9] [16] [17] [18] [19] [21] [22]

Artificial Neural Networks is a common approach to address customer attrition modeling. [9] [18] It attempts to simulated biological neural systems which learn by changing the strength of the connection between two neurons when experience repeated stimulation by the same impulse. There two types of neural networks: single-layer perceptron and multilayer perceptron. While single-layer perceptron networks only contain input layer and out layer, multilayer perceptron networks also contain multiple hidden layers with embedded hidden nodes, each layer can interact with each other by using weighted connections. Customer attrition forecasting is reduced to classification problem (e.g., whether the customer will attrite or not) to apply Artificial Neural Networks. Neural networks can use a variety of learning algorithms, [18] used a supervised feed-forward multilayer perceptron neural networks model trained with back-propagation rule and a 10-fold cross validation procedure to check performance. Neural networks with appropriate algorithm selection can produce a better result than Logistic Regression and Decision Trees regarding forecast precision . [14]

Support Vector Machines (SVMs), also known as Support Vector Networks, was first introduced in B.E. Boser et al., (1992). It is a supervised learning model with associated learning algorithms that can recognize and extract data patterns, and can be used for both classification and regression analysis. [9] [17] [18] These techniques are advantageous when applied to customer attrition data because they are based on the Structural Risk Minimization principle that minimizes the upper bound on the actual risk, they are usually employed with Kernel functions to improve performance (He et al., 2014). In customer attrition forecast, SVMs construct a hyperplane or set of hyperplanes in a high-dimensional space (depending on the number of independent variables) to optimally discriminate between the churners and non-churners, by maximizing the margin between two hyperplanes separating both classes. [18] There are ongoing researches for selecting the optimal combination of kernels to obtain best forecast results. [9] used a hierarchical multiple kernel support vector machine to model both static and longitudinal behavioral data and was able to produce better results than Logistic Regression and Decision Trees regarding forecast precision.

Decision trees, also known as Classification Trees or Regression Trees, are tree-shaped structures constructed by many nodes and branches on different stages and can generate classification rules. [23] It is a process of dividing a large dataset into successively smaller sets of data. Due to the nature of this process, Decision Trees are unable to produce excellent results when handling data with a complex and non-linear relationship between attributes. For customer attrition forecast, Decision Trees have surprisingly good performance when used appropriately. [19] used an ensemble of Classification & Regression Trees algorithm (CART) and Generalized Additive Model (GAM) on actual gambling behavior data and was able to produce a more robust model with better forecast precision compare to CART model alone. It is common to either use Decision Trees with other model or as benchmark model in the reviewed papers. [9] [19] [22] [24]

CHAPTER 3: THEORETICAL BACKGROUND

3.1. Survival Analysis

Survival analysis is a class of statistical model describing time to the event occurrence, which in this thesis is customer attrition. All common approaches to survival analysis are probabilistic or stochastic, which assumes the time to event T is a continuous random process following certain probability distribution.

Let's denote probability density function (p.d.f.) as $f(t)$ and cumulative distribution function (c.d.f.) as

$$F_{(t)} = Pr\{T \leq t\}.$$

(1) - Cumulative Distribution Function of Time to Event

Survival function, which is the probability that customer has not left by duration t, is denoted as

$$S_{(t)} = Pr\{T > t\} = 1 - F_{(t)} = \int_{t}^{\infty} f_{(t)}\, dt.$$

(2) - Survival Function

Hazard function is the instantaneous rate of customer attrition at time t, defined as

$$h_{(t)} = \lim_{\Delta t \to 0} \frac{Pr\{t \leq T < t + \Delta t \mid t \leq T\}}{\Delta t} = \frac{f_{(t)}}{1 - F(t)},$$

(3) - Continuous Hazard Function

which represents the conditional probability of customer attrition time in the interval $[t, t + \Delta t)$ given it is in the interval $[t, \infty)$.

In a discrete setting with a fixed time interval, which is the case of this study, hazard function is defined as

$$h_{(t)} = Pr\{t \leq T < t+1 \mid t \leq T\},$$

(4) - Discrete Hazard Function

which is the conditional probability that the customer will leave in period t given he/she has not left before period t yet. The method of estimating hazard function is the key difference between different models in survival analysis.

3.2. Life Tables

Life tables describe the customer attrition time in terms of the discrete fixed time interval and proportion surviving between each interval without any pre-assumption to the data. *FIGURE 1 is* an example of the customer survival data with time as the x-axis, each line indicates an individual customer, the start of a line indicates customer entering, the end of the line indicates customer attrition (event), and the arrow at the end of the line indicates censoring. Censoring refers to a technique used in survival analysis to represent unknown information, meaning we are uncertain when these customers are leaving. Since we often use up-to-date data in customer attrition analysis, censoring in this study are all happening at the current time point and indicates the customers remain in the organization.

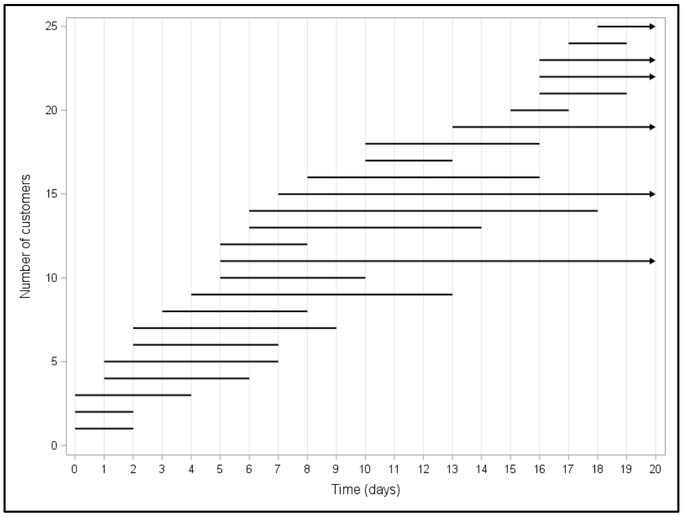


FIGURE 1 - Example of Survival Data

Since life tables consider survival data in a discrete fixed interval setting and only information about the number of periods that they stayed in the organization is used, it is more transparent to represent

FIGURE 1 by aligning all the customers' entry time to the left. As shown in *FIGURE 2*, the x-axis is the number of periods that customers stayed in the organization before the event. Note that we refer period t as time point $[t, t + 1)$.

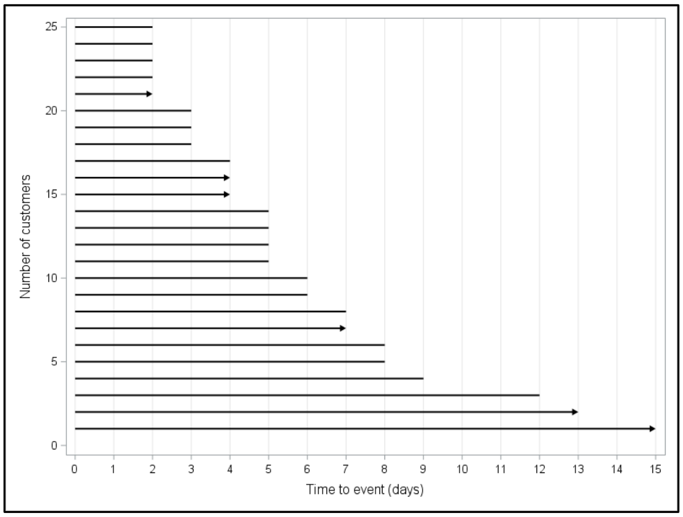


FIGURE 2 - Example of Survival Data (Left Aligned)

If we let N_t denotes the number of customers remaining in the organization for at least $t - 1$ periods, D_t denotes the number of customers left during period t, and as C_t denotes customers censored at period t (meaning whether they left the organization is unknown starting from period t). Life tables handle these customers by excluding them from the data during the next period with the assumption that the time of censoring is uniformly distributed in period t, and estimate the hazard rate at the middle point t.

The probability of a customer leaves in period t given he/she has not left before period t. Which is the hazard function, is estimated as

$$\hat{h}_{(t)} = \frac{D_t}{N_t - \frac{C_t}{2}}.$$

(5) - Life Tables Hazard Function Estimation 1

and the survival function is estimated as

$$\hat{S}_{(t)} = \prod_{t=0}^{T-1} (1 - \frac{D_t}{N_t - \frac{C_t}{2}}).$$

(6) - Life Tables Survival Function Estimation

3.3. Kaplan-Meier Estimator

Kaplan-Meier estimator is created to handle continuous survival time, its hazard rate changes only when customers leave. Let's denote "$t -$" as just before time t and "$t +$" as just after time t. For any time t with customer leaving the organization, let N_t denotes the number of customers remains at time $t -$ and D_t the number of customers left at time t, then the probability of leaving from $t -$ to $t +$ is estimated as

$$\hat{h}_{(t)} = \frac{D_t}{N_t}.$$

(7) - Kaplan-Meier Hazard Function Estimation 1

The survival function is estimated as

$$\hat{S}_{(t)} = \prod_{t \le T} (1 - \frac{D_t}{N_t}),$$

(8) - Kaplan-Meier Survival Function Estimation

which is the product of individual survival probabilities at time points with customer leaving, this is why Kaplan-Meier estimator is also known as product limit estimator. An example of the Kaplan-Meier survival curve is shown in *FIGURE 3*, which is a monotone decreasing function.

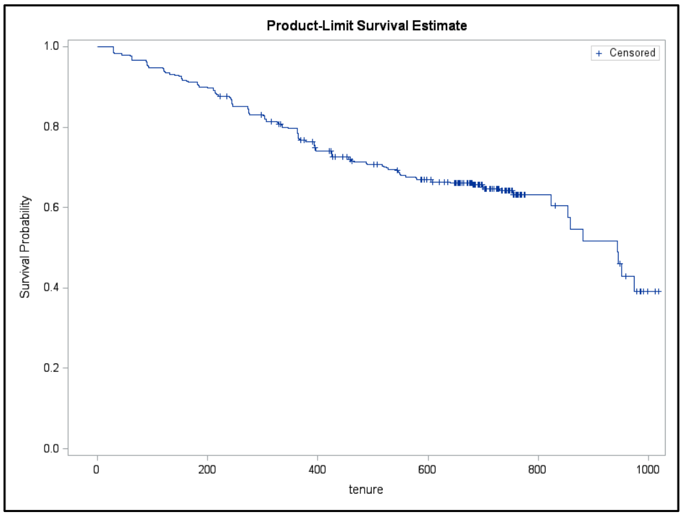


FIGURE 3 - Example of Kaplan-Meier Survival Curve

Continuous survival time means customers can be censored between customers leave, Kaplan-Meier estimator exclude censored customers from the study at the time of censoring, which is why the Kaplan-Meier hazard function does not include C_t. In discrete settings with a fixed time interval, the difference between Kaplan-Meier estimator and life tables is how they handle censoring.

3.4. Multiple Linear Regression

Regression models are extensively used in a variety of studies and industries; it is a model of relationships between the predictors and an outcome. Multiple linear regression refers to a class of regression model with more the one predictor and fulfills the assumptions of linear regression. Let's assume X_1, X_2 are the two predictors (independent variables) and Y_i is the outcome (dependent variable) with n pairs of observation (Y_i, X_{1i}, X_{2i}). Multiple linear regression model assumes that

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

(9) - Multiple Linear Regression Model

where ε_i is the stochastic error term defined as

$$\varepsilon_i \sim N(0, \sigma^2 \cdot I_{n\times n}).$$

(10) - Stochastic Error Term Assumption 1

Ordinary least square is a standard method of estimating $(\beta_0, \beta_1, \beta_2)$, it selects these parameters by minimizing the sum of the squares of the differences between the observed outcome and the prediction of the model as following

$$\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right) = \underset{(\beta_0, \beta_1, \beta_2)}{argmin} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2.$$

(11) - Principle of Ordinary Least Square

The prediction is the conditional mean of the dependent variable given specific values of the independent variables. There are many other methods of estimating parameters in a linear model, one of which is weight least square, the approach we will use in this paper.

Weighted least square is a particular case of the generalized least square model; it can be interpreted as each observation is given a weight, the multiplication of the weight and the sum of the squares of the differences between the observed outcome and the prediction of the model is minimized to select the parameters. Let's denote W_i as the weight for the i^{th} observation, the principle of weighted least square is defined as

$$\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right) = \underset{(\beta_0, \beta_1, \beta_2)}{argmin} \sum_{i=1}^{n} W_i (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2.$$

(12) - Principle of Weighted Least Square

3.5. Model Evaluation

We will use three different metrics to evaluate the model performance as there is no universal metrics used in the previous papers, mean absolute error (MAE) as defined in *(13*, mean squared error (MSE) as defined in *(14*, and mean absolute percentage error (MAPE) as defined in *(15*. *(13*, *(*14, *(*15 will use the notation mentioned in 3.4. Multiple Linear Regression.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|$$

(13) - Mean Absolute Error Definition

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

(14) - Mean Squared Error Definition

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100\%$$

(15) - Mean Absolute Percentage Error Definition

All evaluations are made on the out-of-sample basis, a sliding simulation with fixed validation and test set length is used in the case study.

CHAPTER 4: METHODOLOGY

4.1. Tenured Customer Forecasting

As we have discussed in *3.2.* Life Tables, customers don't all enter at the origin, customers remaining in the organization at the end of the analysis period is right-censored to indicate unknown survival time. Tenured customer refers to the customer already joined the organization and tenure refers to the number of periods stayed in the organization under discrete setting, we will be forecasting the tenure customer count in this paper and exclude customers joined the organization after the forecast origin.

The method of estimating hazard function will be discussed in 4.2. Modified Non-Parametric Models and 4.3. Regression Hazard Model, let's denote $\hat{h}_{(t)}$ as the estimated hazard function for different tenure t. Once we estimate the hazard function, the conditional probability of customers with tenure of T periods stay for additional K periods can be derived as

$$\hat{S}_{(T+K|T)} = \prod_{i=0}^{K-1} (1 - \hat{h}_{(T+i)}).$$

(16) - Survival Function Estimation

Let's denote N_T as the number of customers with tenure of p periods at the forecast origin. The tenure customer forecast, which is the expected number of existing event-free customers that will stay in the organization for additional K periods can be derived as

$$Tenured\ Customer\ Forecast = \sum_{T} [N_T \prod_{i=0}^{K-1} (1 - \hat{h}_{(T+i)})],$$

(17) - Tenured Customer Forecast Definition 1

note that new customers entered after the forecast origin is excluded.

4.2. Modified Non-Parametric Models

In order to explain the proposed model, we will use *TABLE 1* and *TABLE 2* as they provide a better view of data with entry time information. *TABLE 1* is an example of the survival data summary; each column refers to a period with the number indicates starting time point of the period, each row relates to how long a customer has stayed in the organization. The numbers are the starting customer count of the certain period with some tenure, so the starting total number of customers in the organization for each period is the sum of each column.

TABLE 1 - Example of Survival Data - Customer Count

		Period										
		0	1	2	3	4	5	6	7	8	9	10
Tenure	0	1555	1705	1276	599	522	524	914	300	599	153	26
	1		1555	1705	1276	599	522	524	914	300	599	153
	2			1541	1693	1271	594	518	519	904	295	591
	3				1470	1610	1197	556	489	494	848	275
	4					1377	1453	1095	500	459	452	773
	5						1087	1149	841	412	378	402
	6							798	922	717	322	308
	7								673	825	642	288
	8									611	745	582
	9										550	669
	10											506

Let's denote $N_{p,t}$ as the starting number of customers for period p with tenure t. We denote $H_{p,t}$ as the probability of attrition for each of $N_{p,t}$, which is defined as

$$H_{p,t} = 1 - \frac{N_{p+1,t+1}}{N_{p,t}}.$$

(18) - Hazard Rate Definition

TABLE 2 shows the resulting table for $H_{p,t}$, which is the dependent variable we are aiming to forecast when forecast tenured customers, tenured customer forecast for the next period is the multiplication of the stating customer count and the estimated hazard rate. Note that hazard rates for $Tenure = 0$ are all zero because the minimal time a customer can stay in the organization is one. *FIGURE 4* shows the variation of hazard rate from period to period at different tenure.

TABLE 2 - Example of Survival Data - Hazard Rate

		Period										
		0	1	2	3	4	5	6	7	8	9	10
Tenure	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1		0.009	0.007	0.004	0.008	0.008	0.010	0.011	0.017	0.013	0.013
	2			0.046	0.049	0.058	0.064	0.056	0.048	0.062	0.068	0.102
	3				0.063	0.098	0.085	0.101	0.061	0.085	0.088	0.091
	4					0.211	0.209	0.232	0.176	0.176	0.111	0.173
	5						0.266	0.198	0.147	0.218	0.185	0.231
	6							0.157	0.105	0.105	0.106	0.146
	7								0.092	0.097	0.093	0.090
	8									0.100	0.102	0.065
	9										0.080	0.067
	10											0.069

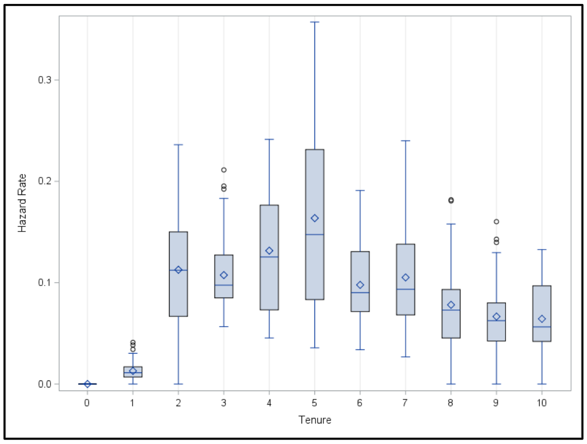


FIGURE 4 - Example of Hazard Rate Variation

Life tables and Kaplan-Meier estimator both use customers entered before the end of training data to estimate the hazard function for customers with the same tenure. We transform Life Tables hazard function estimation at the beginning of period p, which is *(5*, into

$$\widehat{hlt}_{p,t} = 1 - \frac{\sum_{i=t}^{p-1} N_{i,t}}{\frac{1}{2} N_{p,t} + \sum_{j=t+1}^{p} N_{j,t-1}}, t < p.$$

(19) - Life Tables Hazard Function Estimation 2

Furthermore, we can incorporate $H_{p,t}$ into the hazard function estimation with a percentage weight based on their starting number of customers $N_{p,t}$, which can be derived as

$$\widehat{hlt}_{p,t} = \sum_{i=t}^{p-1} \left(H_{i,t} \frac{N_{i,t}}{\frac{1}{2} N_{p,t} + \sum_{j=t}^{p-1} N_{j,t}} \right), t < p.$$

(20) - Life Tables Hazard Function Estimation 3

This provides us a link from hazard rate from all previous customers to hazard rate during each period, which is what we are aiming to forecast. Kaplan-Meier estimator estimates hazard function excluding the censored customers $N_{p,t}$, so we can transform *(7* into

$$\widehat{hkm}_{p,t} = 1 - \frac{\sum_{i=t}^{p-1} N_{i,t}}{\sum_{j=t+1}^{p} N_{j,t-1}}, t < p.$$

(21) - Kaplan-Meier Hazard Function Estimation 2

There is no term associated with $N_{p,t}$ used here, therefore, censored customer excluded from the estimation. Similar to life tables, Customer count based hazard estimation can be expressed as a weighted sum of individual hazard rate,

$$\widehat{hkm}_{p,t} = \sum_{i=t}^{p-1} \left(H_{i,t} \frac{N_{i,t}}{\sum_{j=t}^{p-1} N_{j,t}} \right), t < p.$$

(22) - Kaplan-Meier Hazard Function Estimation 3

Since by period p, a customer can only stay in the organization for a maximum of $p - 1$ periods, hence the restriction $t < p$. *(20, (*22 shows that the hazard function estimation of life tables and Kaplan-Meier can be interpreted as the weighted mean of hazard rates at different tenure with weights based on the percentage of customers at the start of the periods.

The above methods assume the probability of the customers leaving is associated with their tenure in the organization only, while additional covariates can estimate the hazard function more accurately, it is not likely that covariates are available in a low data resolution setting. There is additional information we can use, which is entering time, customers' entering period information is not used in life tables and Kaplan-Meier hazard estimation (see *FIGURE 2*). The intuition behind incorporating entering time into hazard function is promotion and special rates may be given to new customer as they enter, while the exact time range of these promotion is not available, we can assume that customers entered in the same period will behave similar to each other. However, this assumption is does not hold for all periods and tenure is still the dominating factor to estimate most of hazard rates, so we will need to evaluate each period separately.

To compare how valid are tenure associated assumption and entry period associated assumption, we will be assuming the hazard rate in different periods follows a normal distribution and comparing the sample variance of the hazard rate for all customers included by each assumption. *TABLE 3* gives an example of the corresponding customers included in each assumption to estimate a single hazard rate, with green cell as the target, yellow cells as tenure associated hazard rates and blue cells as entry period associated hazard rates. Note that hazard rate at tenure zero is exclude since it is always zero.

TABLE 3 - Example of Hazard Rate Sample Variance Comparison Range

		Period										
		0	1	2	3	4	5	6	7	8	9	10
Tenure	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1		0.009	0.007	0.004	0.008	0.008	0.010	0.011	0.017	0.013	0.013
	2			0.046	0.049	0.058	0.064	0.056	0.048	0.062	0.068	0.102
	3				0.063	0.098	0.085	0.101	0.061	0.085	0.088	0.091
	4					0.211	0.209	0.232	0.176	0.176	0.111	0.173
	5						0.266	0.198	0.147	0.218	0.185	0.231
	6							0.157	0.105	0.105	0.106	0.146
	7								0.092	0.097	0.093	0.090
	8									0.100	0.102	0.065
	9										0.080	0.067
	10											0.069

The actual hazard rate $H_{p,t}$ is applied all customers at the start of period p with tenure t, which is $N_{p,t}$. As the result, when computing the sample variance, each hazard rate $H_{p,t}$has a size of $N_{p,t}$. The total number of customers involved in tenure associated assumption is denoted as $NT_{p,t}$,

$$NT_{p,t} = \sum_{i=t}^{p-1} N_{p,t}.$$

(23) - Tenure Associated Total Customer Definition

Weighted mean hazard rate is estimated in *(20*, *(22* and can be simplified by replacing part of the equation with $\mathrm{NT}_{p,t}$. The mean hazard rate involved in entry period associated assumption is denoted as $NE_{p,t}$,

$$NE_{p,t} = \sum_{i=1}^{t-1} N_{p-t+i,i},$$

(24) - Entry Associated Total Customer Definition

Note that hazard rate at tenure for period 0 is excluded since it is always zero. The corresponding weighted mean hazard rate, which is the expected hazard rate for all customers entered in the same period, can be written as two equations adapting life tables or Kaplan-Meier approach to censored customers. The life tables corresponding entry period based hazard can be estimated as

$$\widehat{hlte}_{p,t} = \sum_{i=1}^{t-1} \left(H_{p-t+i,i} \frac{N_{p-t+i,i}}{\frac{1}{2} N_{t,p} + NE_{p,t}} \right), 3 \leq t < p,$$

(25) - Life Tables Entry Associated Hazard Function Estimation

and the Kaplan-Meier corresponding entry period based hazard can be estimated as

$$\widehat{hkme}_{p,t} = \sum_{i=1}^{t-1} \left(H_{p-t+i,i} \frac{N_{p-t+i,i}}{NE_{p,t}} \right), 3 \leq t < p.$$

(26) - Kaplan-Meier Entry Associated Hazard Function Estimation

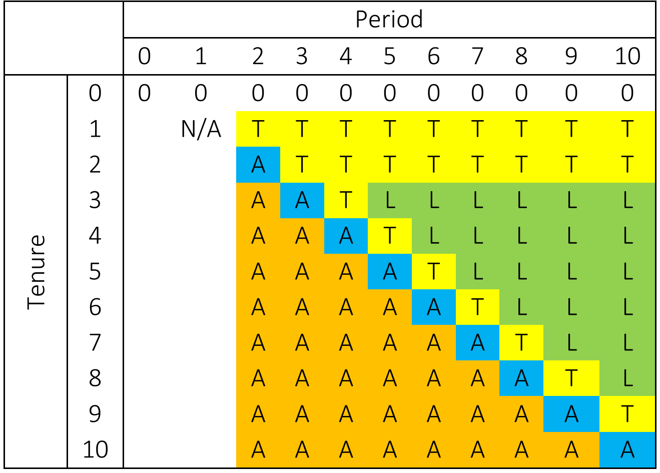
We now can calculate the sample variance using the mentioned customers in each assumption, let's denote $SVT_{p,t}$ as the sample variance of hazard rates in tenure associated assumption, which is used in life tables and Kaplan-Meier hazard estimation, and $SVE_{p,t}$ as the sample variance of hazard rates in entry period associated assumption. Since the entry period associated effect on the hazard rate is not always present in a data set, we will exclude the entry period estimated hazard rate when $SVT_{p,t} < SVE_{p,t}$ and combine hazard rate estimated by each assumption by assigning weight based on one over their corresponding sample variance when $SVT_{p,t} > SVE_{p,t}$. *(27* shows the above scheme with life tables approach, we can also use Kaplan-Meier Estimator by replace the corresponding hazard estimation and sample variance.

$$\hat{h}_{p,t} = f(x) = \begin{cases} \widehat{hlt}_{p,t}, & SVT_{p,t} < SVE_{p,t} \\ \dfrac{SVE_{p,t}\widehat{hlt}_{p,t} + SVT_{p,t}\widehat{hlte}_{p,t}}{SVT_{p,t} + SVE_{p,t}}, & SVT_{p,t} > SVE_{p,t} \end{cases}, t < p$$

(27) - Combined Hazard Function Estimation

An example of the hazard rate estimation at different tenure is shown in *TABLE 4*, with "T" indicates hazard rate estimated by tenure estimated hazard rate (life tables or Kaplan-Meier), which is *(20*, *(22*. "L" indicates hazard rate estimated based on the *(27*. "A" indicates hazard rate estimated based on average of all estimated hazard rate in the same period, which will be discussed below. Note that there are two color in the cells mark with "A", blue cells refer to the hazard rate estimation needed for one step ahead forecast and orange cells refer to multi-step ahead forecast.

TABLE 4 - Example of Hazard Function Estimation Source



		Period										
		0	1	2	3	4	5	6	7	8	9	10
Tenure	0	0	0	0	0	0	0	0	0	0	0	0
	1		N/A	T	T	T	T	T	T	T	T	T
	2			A	T	T	T	T	T	T	T	T
	3			A	A	T	L	L	L	L	L	L
	4			A	A	A	T	L	L	L	L	L
	5			A	A	A	A	T	L	L	L	L
	6			A	A	A	A	A	T	L	L	L
	7			A	A	A	A	A	A	T	L	L
	8			A	A	A	A	A	A	A	T	L
	9			A	A	A	A	A	A	A	A	T
	10			A	A	A	A	A	A	A	A	A

Finally, we will need hazard rate for larger tenure for longer forecasting horizon since the customer will have larger tenure as time moves forward, but there is no customer in the history with the same tenure. [15] used the average of all hazard rate with different tenure for all hazard rate not estimated, we will apply the same approach in this paper.

4.3. Regression Hazard Model

In 4.2. Modified Non-Parametric Models, we modified non-parametric models, namely life tables and Kaplan-Meier estimator, to include the assumption that customers entered in the same year will have similar behavior. The end product of the modified models is the tenure customer forecast as shown in *(17*, while the forecast horizon can be across several periods and the actual hazard rate varies from period to period, the hazard rate for the forecasting horizon are estimated only once at the forecast origin.

FIGURE 5 shows an example of how hazard rate at the same tenure varies from period to period. In the previous chapter, we assume that the hazard rates at a certain period consist of a constant expected hazard rate $\bar{h}_t$ and a stochastic error term following normal distribution.

$$h_{p,t} = \bar{h}_t + \varepsilon_{p,t}$$

(28) - Non-parametric Model Hazard Function Definition

Under this assumption, we compared sample variance of two different series of hazard rates to test their estimation accuracy.

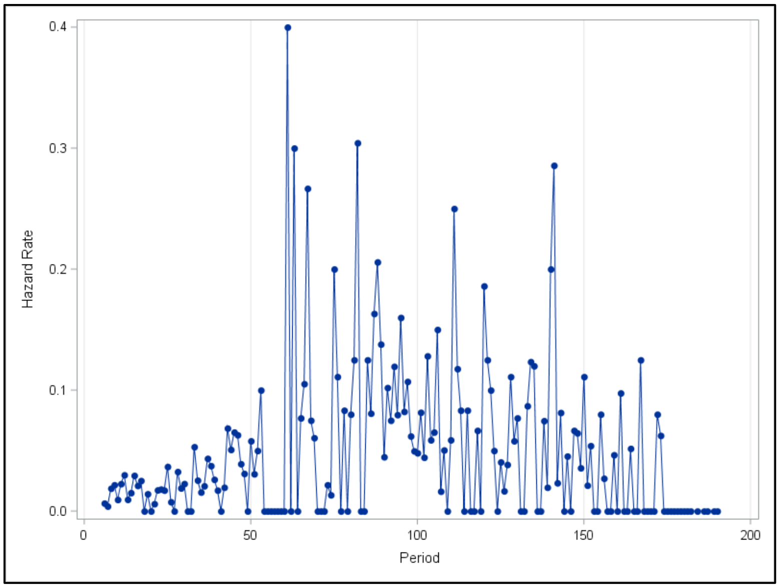


FIGURE 5 - Example of Hazard Rate Variation

Since the hazard rate is representing the decision made by a collection of customers, we can assume these decisions are affected by previous decisions, which is considered as auto correlation in time series analysis. *FIGURE 6* shows the autocorrelation function plot (ACF), inverse autocorrelation function plot (IACF), and partial autocorrelation function plot (PACF) of the hazard rates in *FIGURE 5*. These plots show the degree of correlation with past values of the series as a function of lag values at which the correlation is computed. The blue bar is a statistical test of the hypothesis that none of the correlations of the series up to a given lag are significantly different from 0. As the correlation of some lag value has a correlation higher or lower than the blue bar, there is a sign of autocorrelation in the hazard rate time series.

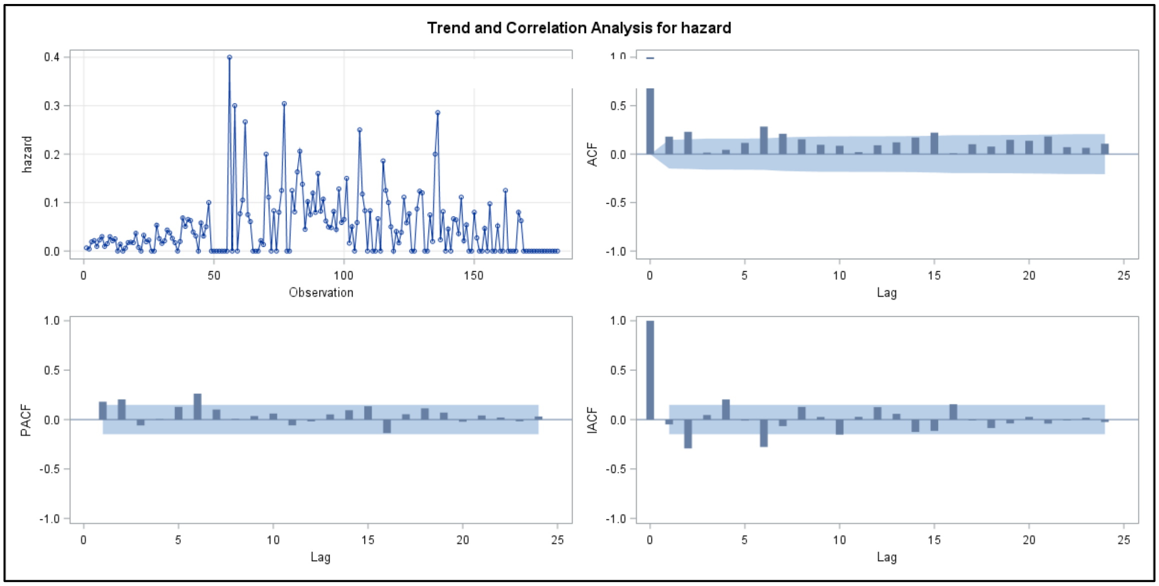


FIGURE 6 - Trend and Correlation Analysis for Hazard Rate

We will include a categorical variable $Tenure_i$ as one of the features to capture the average hazard rate for different tenure, where the numerical variable $Tenure$ is separated into i binary variables $Tenure_i$. Additionally, lagged hazard rates from previous periods as features in the regression model, which is similar to the auto regressive portion in ARIMA model, hazard rate is defined as a linear combination of the variables mention above,

$$H_{p,t} = \alpha_0 + \sum_i \alpha_i Tenure_i + \sum_j \beta_j LH_j + \varepsilon_{p,t},$$

(29) - Tenure Associated Auto Regressive Hazard Model

where LH_j is the hazard rate with the same tenure j periods ago. Note that certain lag range of hazard rate may not exist for hazard rate with higher tenure, in *TABLE 5*, we show an example of the range of hazard rate used in regression. Suppose hazard rates marked in green are used as the training set, to avoid using non-existing lagged hazard rates as features in out model, the maximum number of lagged hazard rates we can include is one. The hazard rate in yellow are the features for observation of hazard rate in period 6, we will discuss the technique to control this restriction in the case study.

TABLE 5 - Example of Features in Regression Hazard Model

		Period										
		0	1	2	3	4	5	6	7	8	9	10
Tenure	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1		0.009	0.007	0.004	0.008	0.008	0.010	0.011	0.017	0.013	0.013
	2			0.046	0.049	0.058	0.064	0.056	0.048	0.062	0.068	0.102
	3				0.063	0.098	0.085	0.101	0.061	0.085	0.088	0.091
	4					0.211	0.209	0.232	0.176	0.176	0.111	0.173
	5						0.266	0.198	0.147	0.218	0.185	0.231
	6							0.157	0.105	0.105	0.106	0.146
	7								0.092	0.097	0.093	0.090
	8									0.100	0.102	0.065
	9										0.080	0.067
	10											0.069

For hazard rate forecast in higher forecast horizon, we adopt the same methodology used in univariate time series analysis. Forecasts in the previous periods are filled as features, but model estimators α's and β's are only estimated at forecast origin, which means hazard rates are estimated for each period with different tenure. Therefore, expected number of existing event-free customers that will stay in the organization for additional K periods with forecast origin period P, which was *(17,* is modified into

$$Tenured\ Customer\ Forecast^{*} = \sum_{T} [N_{P+i,T} \prod_{i=0}^{K-1} (1 - \widehat{H}_{(P+i,T+i)})].$$

(30) - Tenured Customer Forecast Definition 2

Because we include forecasted value in independent variables, the model is subject to error in regressor, hence the forecasts accuracy decreases as forecast horizon increases. We will discuss problems associated with it specifically in the case study.

The regression model estimators will be solved with weight least square; weights are the corresponding starting customer count ($N_{p,t}$) associated with each hazard rate ($H_{p,t}$).

Each hazard rate $H_{p,t}$ is the mean probability of attrition for a collection of customers, each customer should be an observation and has equal weight instead of each hazard rate is an observation. Therefore, we change our assumption in the error term, which is *(10* to

$$\varepsilon_{p,t} \sim N\left(0, \sigma^2 \cdot I_{n\times n} \cdot N_{p,t}\right).$$

(31) - Stochastic Error Term Assumption 2

We refer to *(12* as the principle to estimate the model with weighted least square, weight W_i is replaced with $N_{p,t}$.

CHAPTER 5: CASE STUDY

5.1. Data Description

The case study of this paper is from a fast growing retail electric provider operating in deregulated electricity markets, customers in these markets are able to switch electric provider freely. Compare to other parts of the country where electricity is regulated, these electric providers' customers have a higher hazard rate and total number of customers can subject to high volatility over time. Its territory is divided into 14 different zones and we will be using data from 10 of them to illustrate the proposed method, data from the rest 4 zones are discarded. We selected 10 data sets with history longer than 6 months since we want to compare model performance in low resolution and high resolution setting.

High volatility in customer count propose challenges to forecast electricity load for the retail electric provider since load are affected by the number of customers, and the provider has adopted a conservative strategy to schedule generators based on existing customer load demand. [15] provided a solution to this challenge by forecasting tenured customer count and load per customer separately.

The original data set includes customer entering, leaving date, and additional information on individual customers. We will use customer entering and leaving date only in our analysis. *FIGURE 7* shows daily customer count from entering date of first customer to study end date in one of the 10 zones, A. Data set from each zone is aggregated into 5 lower resolution, 7, 14, 21, 28, 35 days to test the proposed methods' performance on data with different resolution, *FIGURE 8* shows the aggregated customer count of zone A under different resolution. Since the interval is higher for the aggregated data sets, we consider customers that disconnected before end of each period as staying in the organization until the end of the period. This approach simulates the real life scenario where customer leave the organization at the end of an interval, but the exact time of the decision making is uncertain.

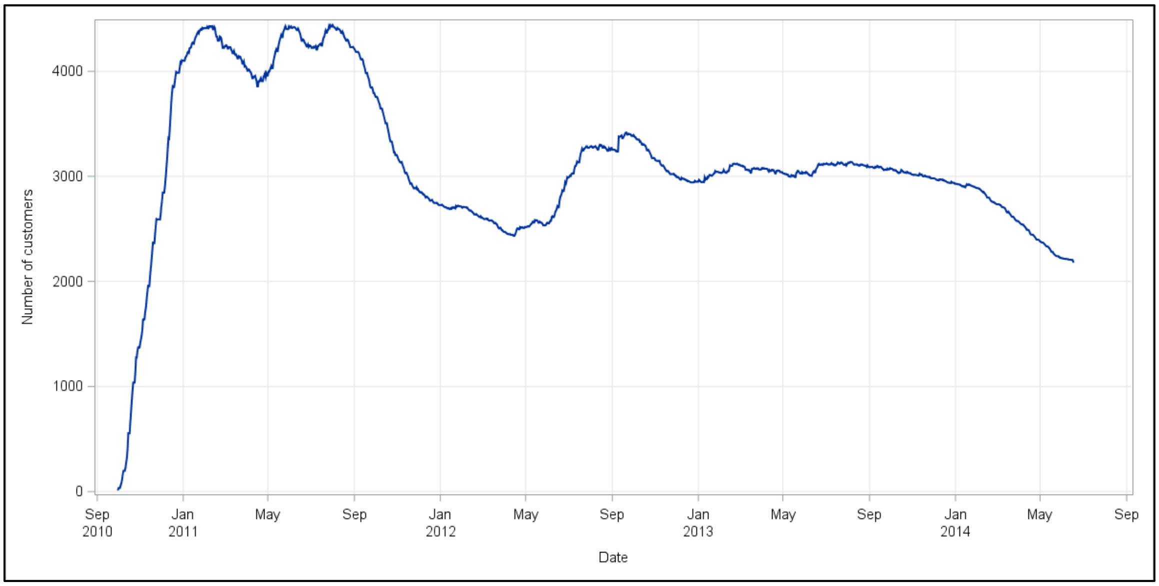


FIGURE 7 - Daily Customer Count of Zone ACE

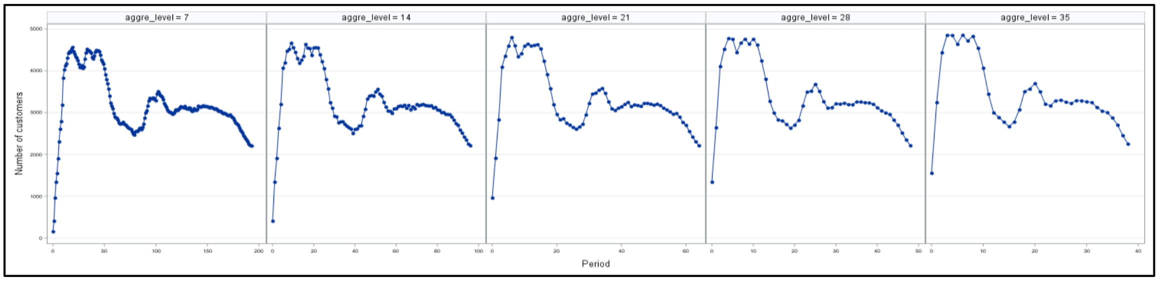


FIGURE 8 - Customer Count at Different Aggregation Level

We split the data set into three different sets based on the total length of the history and a sliding simulation as shown in *FIGURE 9* is used to verify the performance of the proposed method.

1. We start with using 60% of the data as the training set to estimate parameters in the model, certain hyperparameters will be discussed in individual models.

2. The estimated model then is fitted to the next 13% of data, the validation set, which we will compare the forecasts with the actual value to acquire out-of-sample result and its corresponding hyperparameters.
3. At last, the model will be estimated again using the training set and validation set, out-of-sample result in the test set (next 13% of the data), and its corresponding hyperparameters are recorded to compare with results in validation set.

The above process is repeated by increasing the training set period by one and ends when 87% of the data is used as training set, the process is aiming to simulate the real life forecasting task with forecasting origin sliding forward. The green bar is the training periods, orange bar is the validation periods, and yellow bar is the test periods, with each horizontal bar being one step in the simulation.

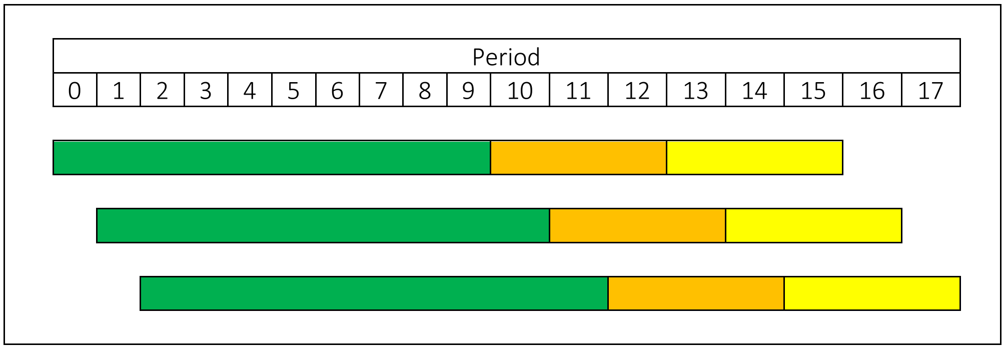


FIGURE 9 - Example of Sliding Simulation

It is important to understand that since we are forecasting tenured customer count, customers entered after the training set ends is not included in the validation set result. Similarly, customers entered after the validation set ends is not included in the testing set result. *FIGURE 10* provides an illustration for the forecast period customer count; Red line indicates validation set tenured customer count and green line indicates test set tenured customer count.

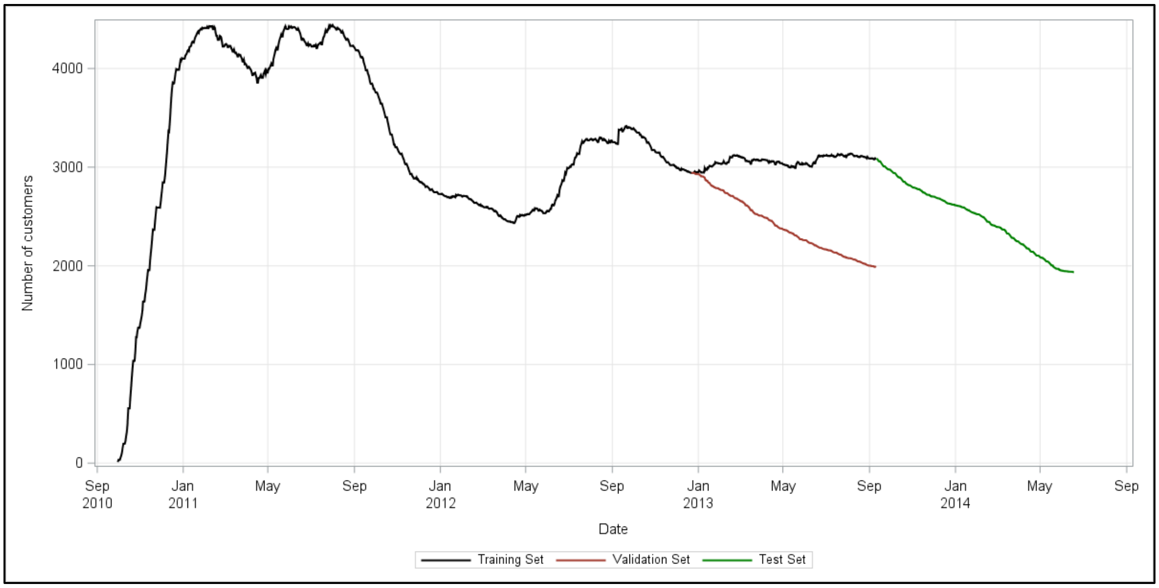


FIGURE 10 - Example of Customer Count Sets

The summary statistics of the 10 datasets included in this study is shown in *TABLE 6*, which include aggregation level, total number of periods, minimum training end period, forecasting horizon, sliding steps for each data sets.

TABLE 6 - Summary Statistics of REP Customer Data

Zone	A					B					C					D					E				
Aggregation Level	7	14	21	28	35	7	14	21	28	35	7	14	21	28	35	7	14	21	28	35	7	14	21	28	35
Total Periods	190	93	61	45	35	118	57	37	27	21	78	37	24	17	13	111	54	35	25	19	115	56	36	26	20
Minimum Training End	113	55	36	26	20	70	34	22	16	12	46	22	14	10	7	66	32	20	14	11	68	33	21	15	11
Forecasting Horizon	20	10	7	5	4	12	6	4	3	3	8	4	3	2	2	12	6	4	3	2	12	6	4	3	3
Sliding Steps	20	10	7	6	5	13	7	5	4	3	9	5	3	3	2	12	6	5	4	3	13	7	5	4	3

	F					G					H					I					J				
Aggregation Level	7	14	21	28	35	7	14	21	28	35	7	14	21	28	35	7	14	21	28	35	7	14	21	28	35
Total Periods	179	88	57	42	33	66	31	20	14	10	176	86	56	41	32	183	90	59	43	34	64	30	19	13	10
Minimum Training End	107	52	34	25	19	39	18	11	8	5	105	51	33	24	19	109	53	35	25	20	38	17	11	7	5
Forecasting Horizon	18	9	6	5	4	7	4	3	2	2	18	9	6	5	4	19	10	6	5	4	7	4	2	2	2
Sliding Steps	19	10	7	5	4	8	4	3	2	2	19	10	7	5	4	19	10	7	5	4	7	4	3	2	2

5.2. Hazard Function Estimation

Hazard function is the key function to estimate in survival analysis and essential difference between methods. We use life tables and Kaplan-Meier estimator as the benchmark models to compare result with the proposed methods.

5.2.1. Benchmark Models

Refer to *(5*, *(7* for life tables' and Kaplan-Meier's respective hazard function estimation model. Hazard function are estimated at the end of training set for validation result and at the end of validation set for test set result. In 4.2. Modified Non-Parametric Models we mentioned the hazard function for customers stayed in the organization since the beginning of the study period is not estimated by non-parametric methods. For the benchmark models, we will be using a naïve approach by applying the average of estimated hazard rates. As the result, for $t \geq p$, hazard rate estimation will take on the value of the average of estimated hazard rate, which can be defined as

$$\hat{h}_{p,t} = \sum_{i=1}^{p-1} \frac{\hat{h}_{p,i}}{p-1}, t \geq p.$$

(32) - Benchmark Hazard Rate Estimation for Large Tenure

We include both model in this study to test their assumption on censored customer on data set with difference resolution, since the sore difference between the two approach is censored customer handling in discrete fixed interval setting. Kaplan-Meier estimator excludes all censored customers while life tables include half of the censored customers for one more period with zero hazard rate.

5.2.2. Modified Non-Parametric Models

In 4.2. Modified Non-Parametric Models we define the modified non-parametric models as life tables or Kaplan-Meier models with portion of hazard rate estimation associated with weighted average hazard rate of its hazard rates in previous periods. The hazard function estimation scheme for life tables is defined in *(27*, which we will use to estimate the hazard function.

Similar to non-parametric models, we are still unable to estimate future hazard rates for customers entered at the beginning of the study. We will use the same naïve approach for benchmark models, which is defined in *(32*.

. Let's denote L_1 as the maximum number of recent periods included in the assumption with $L_1 \leq t$ and life tables entry associated hazard estimation, *(25* is transformed into

$$\widehat{hlte}^*{}_{p,t} = \sum_{i=1}^{L_1} \left(H_{p-i,i} \frac{N_{p-i,i}}{\frac{1}{2} N_{t,p} + NE^*{}_{p,t}} \right), 3 \leq t \leq p,$$

(33) - Modified Life Tables Entry Associated Hazard Function Estimation

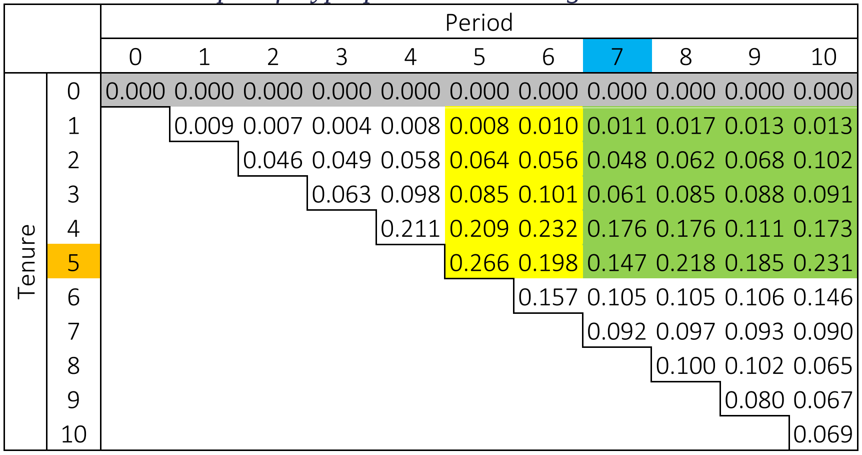
where $NE^*{}_{p,t}$ is the new total number of customers included in the assumption. Kaplan-Meier entry associated hazard estimation is also changed in the similar fashion.

5.2.3. Regression Hazard Model

Regression hazard model is defined in 4.3. Regression Hazard Model, lagged hazard rates of customers at the same tenure and lagged hazard rates of customers entered in the same period are included in the multiple linear regression model, this is the same as auto regressive model in time series analysis. However, while regression model requires each observation in the data set has the same number of regressors, there are limited number of lagged hazard rate available for higher tenure.

We use two hyperparameters to restrict the maximum number lagged hazard rates included in the model. Let's denote L_2 as the minimum period included the training set and L_3 as the maximum tenure included. *TABLE 7* shows an example of the hazard rate estimation training set, the red cells are hazard rate for tenure zero, which are excluded from the training set. Refer to *TABLE 2* for green and yellow cells description, blue and orange cells are the hyperparameter L_2 and L_3 respectively, where number of columns in the yellow cells, which we will denote as Lag_{max}, is the maximum number of lagged hazard rates in the model.

TABLE 7 - Example of Hyperparameters in Regression Hazard Model



		Period										
		0	1	2	3	4	5	6	7	8	9	10
Tenure	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1		0.009	0.007	0.004	0.008	0.008	0.010	0.011	0.017	0.013	0.013
	2			0.046	0.049	0.058	0.064	0.056	0.048	0.062	0.068	0.102
	3				0.063	0.098	0.085	0.101	0.061	0.085	0.088	0.091
	4					0.211	0.209	0.232	0.176	0.176	0.111	0.173
	5						0.266	0.198	0.147	0.218	0.185	0.231
	6							0.157	0.105	0.105	0.106	0.146
	7								0.092	0.097	0.093	0.090
	8									0.100	0.102	0.065
	9										0.080	0.067
	10											0.069

Let's assume that we are estimating the hazard rate at period P, then we can define L_1, L_2, Lag_{max} as the following,

$$L_2 \in [1, P-1], \qquad L_3 \in [1, L_2], \qquad Lag_{max} = L_2 - L_3.$$

(34) - Restriction on Hyperparameters in Regression Hazard Model

Compare to the modified non-parametric models, we made two changes in the validation process.

1. Rather than finding the pair of L_2, L_3 with the least validation period forecast accuracy for each forecast period, we will give hazard rate estimation at different tenure different pair of hyperparameters. Meaning hazard rate at each tenure is estimated separately.
2. We will use the MAE of hazard rate estimation and actual hazard rate in the validation set to select hyperparameters instead of the tenure customer forecast.

Due to the fact that max tenure increases as the period moves forward, there are tenure in the test set larger than the tenure in the validation set. Therefore, no validation is done on these tenures. We will use the life tables estimation of the hazard rate for the hazard estimation at these tenure and hazard rates with tenure larger than maximum tenure at the test set forecast origin will be estimated using the average of all previous hazard rate estimation in the same period. *TABLE 8* shows an example of the hazard estimation for one step in the sliding simulation, with period 7-8 as the validation set and period 9-10 as the test set. Light blue cells indicate hazard rates are estimated using regression hazard model, with the dark blue cells being the validation set to select hyperparameters. Light green cells indicate hazard rates estimated using life tables, and dark green cells indicate hazard rates estimated using average of estimated hazard rates in the same period.

TABLE 8 - Example of Regression Hazard Model Estimation

		Period										
		0	1	2	3	4	5	6	7	8	9	10
Tenure	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1		0.009	0.007	0.004	0.008	0.008	0.010	0.011	0.017	0.013	0.013
	2			0.046	0.049	0.058	0.064	0.056	0.048	0.062	0.068	0.102
	3				0.063	0.098	0.085	0.101	0.061	0.085	0.088	0.091
	4					0.211	0.209	0.232	0.176	0.176	0.111	0.173
	5						0.266	0.198	0.147	0.218	0.185	0.231
	6							0.157	0.105	0.105	0.106	0.146
	7								0.092	0.097	0.093	0.090
	8									0.100	0.102	0.065
	9										0.080	0.067
	10											0.069

Combine *(29* with the hyperparameter L_2 and L_3 assuming validation set starts at period P_v and test set starts at P_t, results in the following equation for regression hazard model estimation.

$$\widehat{H}_{p,t} = \begin{cases} \alpha_0 + \sum_{i=1}^{L_3} \alpha_i Tenure_i + \sum_{j=1}^{L_2-L_3} \beta_j LH_j^*, & 1 \le t < P_v - 1 \\ \widehat{hlt}_{p,t}, & P_v - 1 \le t < P_t \\ \sum_{i=1}^{P_t-1} \frac{\hat{h}_{p,i}}{P_t - 1}, & P_t \le t \end{cases}$$

(35) - Regression Hazard Model Hazard Function Definition

Note that LH_j^* indicates the lagged hazard rates are forecasted value from previous period as forecast horizon increases.

5.3. Tenured Customer Forecasting

The results are record in terms of number of tenured customer count for periods in the validation and test set, which is an aggregated value of tenure customers count in the period with different tenure. Metrics used in this paper, such as MAE, MSE, MAPE, are all based on the difference between forecasts and actual values in period of the forecast horizon under certain hyperparameter. Note that MAPE can be misleading when there are zeros in the tenure rates, therefore we will mainly use MAE and MSE as the primary metrics to compare forecast accuracy.

For benchmark and modified non-parametric models, one hazard function is estimated for all forecasting horizon, tenured customer forecast is generated using *(17*. The regression hazard model estimates hazard function for each period separately by using previously estimated hazard rate in the training set as forecast horizon increases, tenured customer forecast is generated using *(35*.

CHAPTER 6: FORECASTING RESULTS

6.1. Benchmark Results

FIGURE 11 show an example of the weekly tenured customer benchmark forecasts for validation and test set respectively. Life tables and Kaplan-Meier estimator produces similar tenured customer forecasts due to their small hazard estimation difference regarding censored customer. Since life tables assign half of the censored customer a hazard rate of zero, its final tenure customer forecast is always higher than Kaplan-Meier estimation.

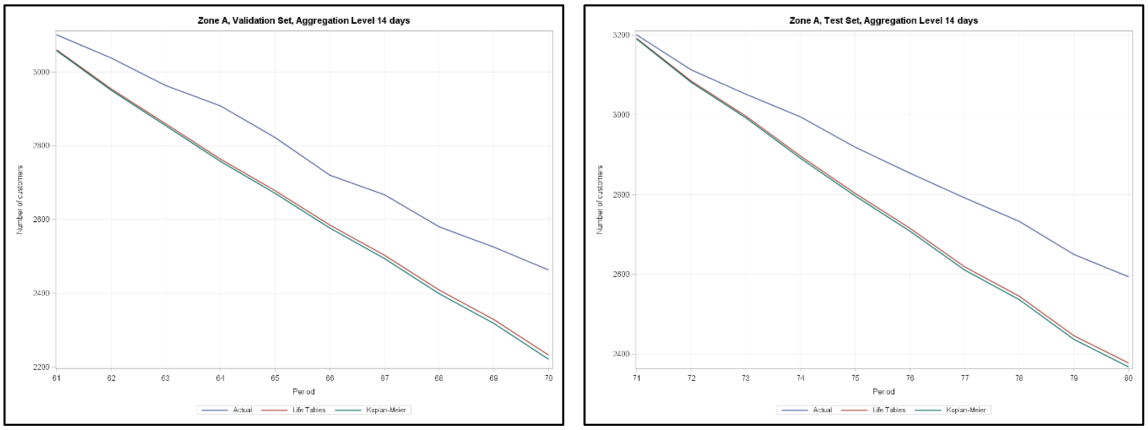


FIGURE 11 - Example of Benchmark Forecasts

TABLE 9 shows the validation and test summary performance of benchmark models with different metrics, detailed table can be found in *APPENDIX A* - BENCHMARK MODEL PERFORMANCE. Life tables performs better in 218 out of the 300 combinations of dataset, aggregation level, metrics, and periods. To simplify the result comparison, we will use life table as the benchmark to compare with the proposed models from now on.

TABLE 9 - Benchmark Validation and Test Performance Summary

	MAE				MSE				MAPE			
	Validation		Test		Validation		Test		Validation		Test	
Aggregation Level	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier
7	8	2	7	3	8	2	6	4	7	3	8	2
14	8	2	7	3	8	2	6	4	7	3	8	2
21	8	2	6	4	8	2	6	4	6	4	8	2
28	8	2	8	2	8	2	8	2	8	2	8	2
35	5	5	8	2	4	6	9	1	9	1	5	5
Total	37	13	36	14	36	14	35	15	37	13	37	13

6.2. Compare Modified Models with Benchmark

Similar to the benchmark models, modified life tables and modified Kaplan-Meier produces forecasts with close performance. Since we use life tables as the benchmark model, we will compare it with the result of modified life tables and Kaplan-Meier result will be included in the appendix.

FIGURE 12 is the comparison example of the weekly tenured customer forecast between benchmark and modified models ($L_1 = 2$) for validation and test set respectively.

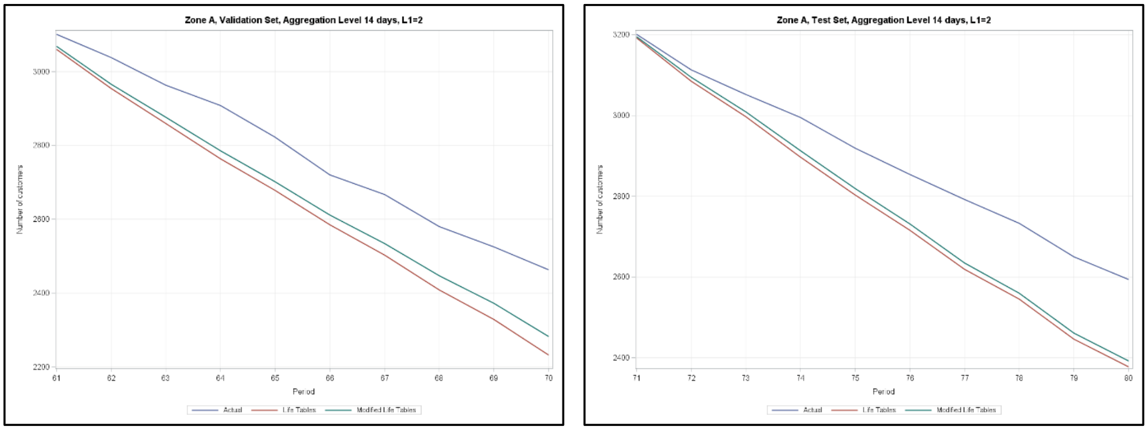


FIGURE 12 - Example of Modified Life Table Forecasts

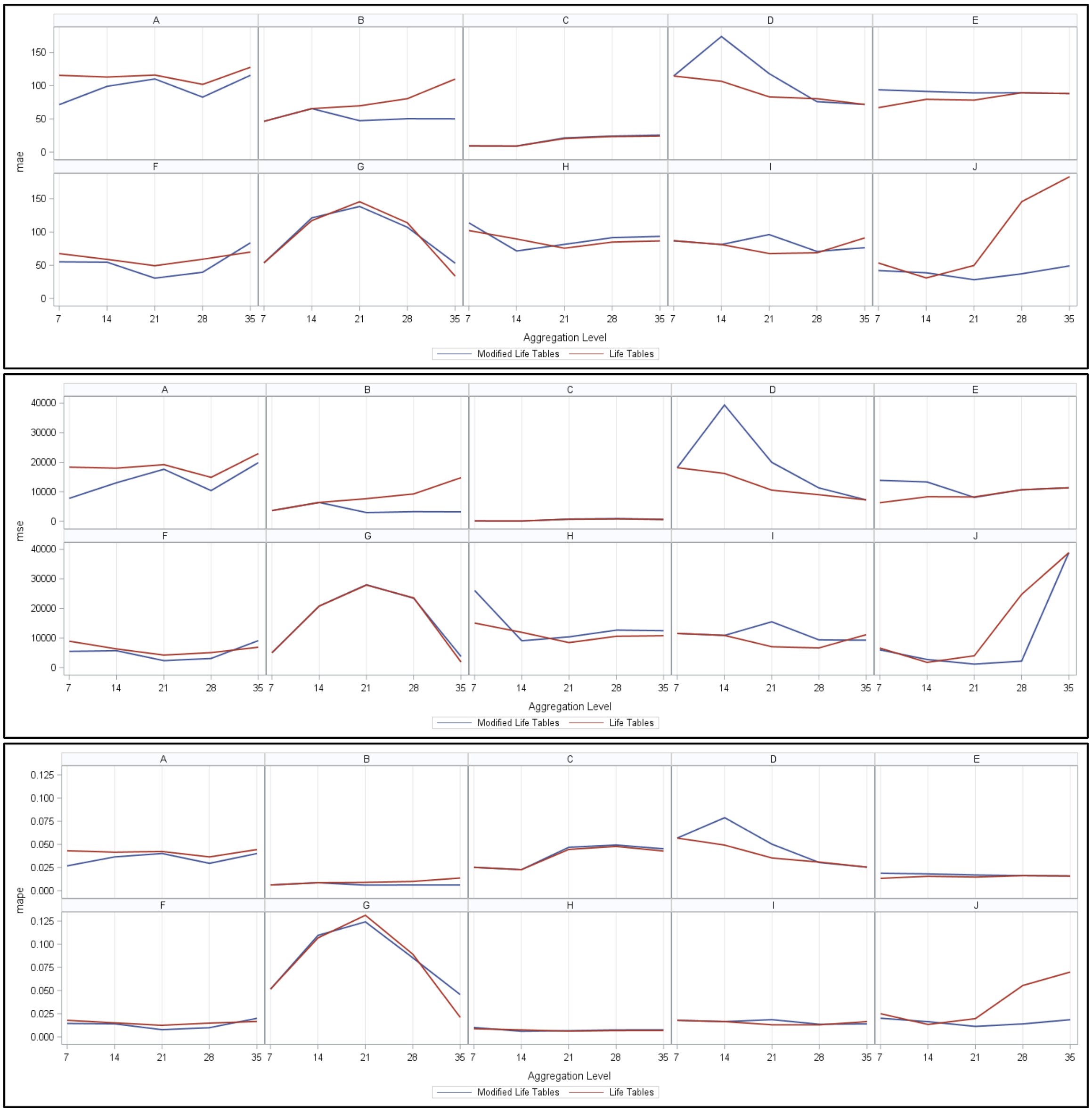


FIGURE 13 - Modified Life Tables Test Set Performance Comparison

As mentioned in 5.2.2. Modified Non-Parametric Models, hyperparameter L_1 is used as maximum number of recent periods included in the entry associated assumption. *FIGURE 13* shows the comparison of test period performance between the modified life tables model and life tables benchmark model under different metrics, the final forecast for the test set is based on the performance of the model in validation set. The best performing hyperparameter L_1 in the validation set is selected for the test set.

The relative performance of the proposed model compared to benchmark under different metrics align with each other. Summary of the performance is shown in *TABLE 10*, with number of the datasets proposed model performed better or equal compared to the benchmark model. A more detailed table is included in *APPENDIX B* - MODIFIED LIFE TABLES TEST PERFORMANCE COMPARISON with the highlighted cells being the lower error metric value.

TABLE 10 - Modified Life Table Test Set Performance Summary

	MAE		MSE		MAPE	
Aggregation Level	Better/Same	Worse	Better/Same	Worse	Better/Same	Worse
7	8	2	8	2	8	2
14	6	4	7	3	6	4
21	5	5	5	5	5	5
28	7	3	6	4	7	3
35	6	4	6	4	6	4
Total	32	18	32	18	32	18

It is interesting that proposed model performs better in high resolution data such weekly aggregation level. This contradicts our previous assumption that as resolution of the data increases, the effect of entry period associated hazard rate estimation perishes. One of the reasons of this unexpected result can come from the validation process, as the data resolution increases, error metrics on validation and test period have a higher correlation, meaning that we are more likely to find the appropriate hyperparameter L_1 for test period based on the model performance with the same L_1 at validation period.

6.3. Compare Regression Hazard Model with Benchmark

FIGURE 14 is the comparison example of the weekly tenured customer forecast between life tables benchmark and the regression model for test set.

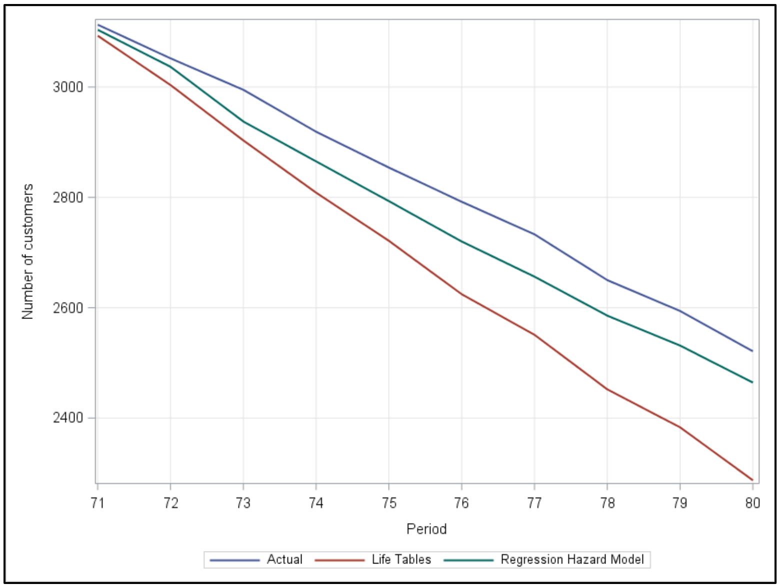


FIGURE 14 - Example of Regression Hazard Model Forecasts

As mentioned in 5.2.3. Regression Hazard Model, hyperparameters L_2, L_3 are used to control the training set length and lagged feature to maximize data usage and avoid using non-existing hazard rate as features in the model. *FIGURE 15* shows the comparison of test period performance between the regression hazard model and life tables benchmark model under different metrics, the final forecast for the test set is based on the performance of the model in validation set. The best performing hyperparameters L_2 , L_3 in the validation set are selected for the test set.

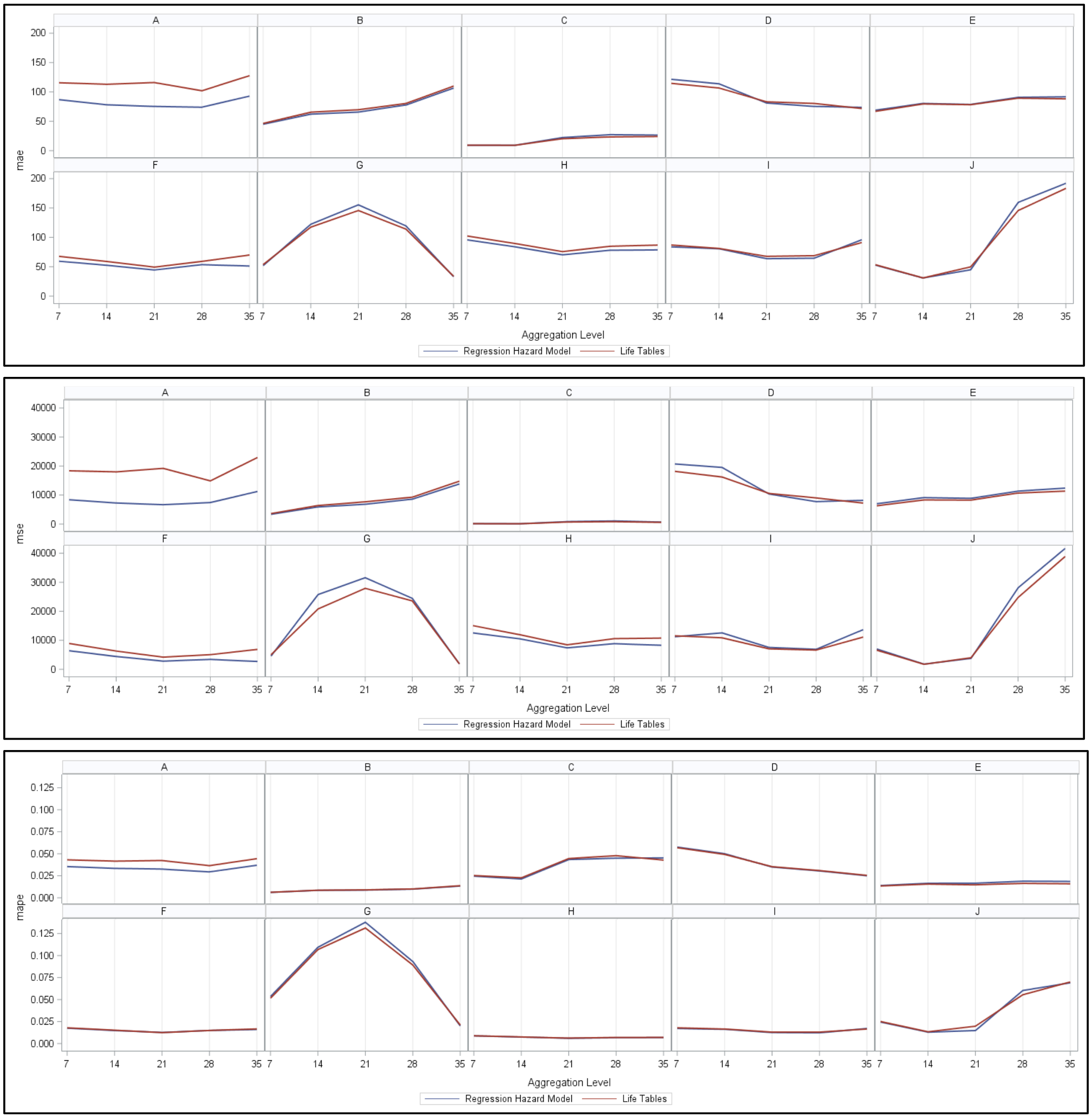


FIGURE 15 - *Regression Hazard Model Test Set Performance Comparison*

Summary of the performance is shown in *TABLE 11*, with the number of the datasets proposed model performed better compared to the benchmark model. A more detailed table is included in *APPENDIX C* - REGRESSION HAZARD TEST PERFORMANCE COMPARISON with the highlighted cells being the lower error metric value. In MAE results, proposed model performs better as data resolution increases, but performs similarly with different data resolution in MSE results. This is mainly because proposed model results are overall more accurate compared benchmark but at the same time produce more forecasts with higher absolute error. Proposed model results in MAPE performs similarly with different data resolution but is overall better than results in MAE. The higher performance with lower resolution in MAPE results shows that the proposed model error is better distributed based on the actual customer count.

TABLE 11 - Regression Hazard Model Table Test Set Performance Summary

	MAE		MSE		MAPE	
Aggregation Level	Better	Worse	Better	Worse	Better	Worse
7	8	2	6	4	7	3
14	7	3	4	6	7	3
21	7	3	6	4	7	3
28	6	4	5	5	6	4
35	5	5	5	5	7	3
Total	33	17	26	24	34	16

Since we used past forecasted value to forecast periods with higher forecast horizon, we can expect accuracy decreases as the forecast horizon increases. *FIGURE 16* shows an example of the absolute error variation comparison between proposed model and benchmark as forecast horizon increases in one dataset. Each box plot consists of the forecast error of the corresponding forecast horizon and model. While the example shows a dominating performance from the proposed model for all forecast horizon, we want to find the performance comparison without the effect of past forecasted value, which in case is the one-step ahead forecast result.

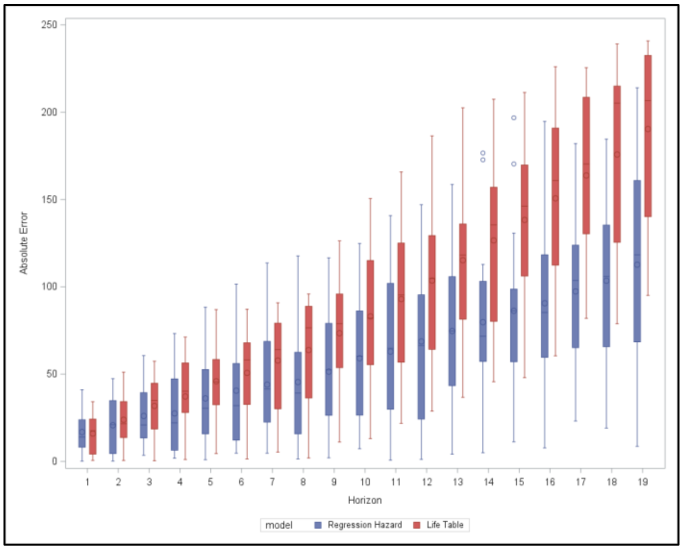


FIGURE 16 - Example of Forecast Error with Higher Forecast Horizon

TABLE 12 shows the summary of the proposed model performance in one-step ahead forecast. A more detailed table is included in *APPENDIX D* - REGRESSION HAZARD TEST ONE-STEP AHEAD FORECAST PERFORMANCE COMPARISON with the highlighted cells being the lower error metric value. The performance comparison is similar to the multiple-step ahead forecast with slightly higher overall performance of proposed model. This shows that the as the forecast horizon decreases, regression hazard model performance increases.

TABLE 12 - Regression Hazard Model Test Set One-step Ahead Forecast Performance Summary

	MAE		MSE		MAPE	
Aggregation Level	Better	Worse	Better	Worse	Better	Worse
7	9	1	5	5	9	1
14	8	2	6	4	9	1
21	6	4	5	5	6	4
28	7	3	7	3	7	3
35	7	3	7	3	6	4
Total	37	13	30	20	37	13

CHAPTER 7: CONCLUSION

As companies collects more information on individual customers to aid customer relationship management, recent customer attrition models are developing with the assumption of large customer base, high data resolution and extensive history. For start-up companies and programs, it is difficult to forecast customer attrition due to poor data quality. This thesis investigates customer attrition modeling without individual customer information on different data resolution. Two proposed models are tested on retail electric customer data. Both models out-perform benchmark model at all data resolution. Modified non-parametric model provides large improvement but with the higher forecast accuracy variance. Regression hazard model has a smaller forecast accuracy variance but at the cost of a smaller improvement compared to the modified non-parametric model. The performance of the both models increases as data resolution increases while overall still slightly out-perform the benchmark model at the lowest data resolution. Proves that as data resolution decreases, the effect of additional modeling dissipates.

Although information at lower resolution can be extracted when forecasting with higher resolution data, it is not utilized in either proposed models or benchmark model, resulting in a worse performance as data resolution increases in some of the data sets. It is possible that considering lower resolution forecast results in high resolution data can improve the forecast accuracy. Additional research could be conducted to analyze the effect of a hierarchal forecasting model to fully utilize the information contained in raw customer data.

Aside from the above future research direction, multi-step forecast performed in this thesis is cumulative, the forecast accuracy of periods at high forecast horizon is affected by the forecast accuracy of previous periods. Direct forecast of the periods with higher forecast horizon maybe able to improve accuracy on datasets with higher resolution.

REFERENCES

1. Athanassopoulos, A.D., *Customer Satisfaction Cues To Support Market Segmentation and Explain Switching Behavior.* Journal of Business Research, 2000. **47**(3): p. 191-207.

2. Colgate, M.R. and P.J. Danaher, *Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution.* Journal of the Academy of Marketing Science, 2000. **28**(3): p. 375.

3. Van den Poel, D. and B. Larivière, *Customer attrition analysis for financial services using proportional hazard models.* European Journal of Operational Research, 2004. **157**(1): p. 196-217.

4. Kumar, V., A. Leszkiewicz, and A. Herbst, *Are You Back for Good or Still Shopping Around? Investigating Customers' Repeat Churn Behavior.* Journal of Marketing Research, 2018. **55**(2): p. 208-225.

5. Guillén, M., et al., *Time-varying effects in the analysis of customer loyalty: A case study in insurance.* Expert Systems with Applications, 2012. **39**(3): p. 3551-3558.

6. Miguéis, V.L., A. Camanho, and J. Falcão e Cunha, *Customer attrition in retailing: An application of Multivariate Adaptive Regression Splines.* Expert Systems with Applications, 2013. **40**(16): p. 6225-6232.

7. Coussement, K., S. Lessmann, and G. Verstraeten, *A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry.* Decision Support Systems, 2017. **95**: p. 27-36.

8. Ahmed, M., et al., *A Survey of Evolution in Predictive Models and Impacting Factors in Customer Churn.* Advances in Data Science and Adaptive Analysis, 2017. **9**(3).

9. Chen, Z.-Y., Z.-P. Fan, and M. Sun, *A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data.* European Journal of Operational Research, 2012. **223**(2): p. 461-472.

10. Jahromi, A.T., S. Stakhovych, and M. Ewing, *Customer Churn Models: A Comparison of Probability and Data Mining Approaches.* Looking Forward, Looking Back: Drawing on the Past to Shape the Future of Marketing, 2016: p. 144-148.

11. Schweidel, D.A., Y.H. Park, and Z. Jamal, *A Multiactivity Latent Attrition Model for Customer Base Analysis.* Marketing Science, 2014. **33**(2): p. 273-286.

12. Ellsworth, R.K., *Attrition Analysis and Customer-Relationship Life Expectancy.* Business Valuation Review, 2005. **24**(4): p. 173-176.

13. Matsuno, S., et al., *A survival analysis of the Japanese information service industry.* Procedia Computer Science, 2017. **121**: p. 291-296.

14. Lariviere, B. and D. Van den Poel, *Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services.* Expert Systems with Applications, 2004. **27**(2): p. 277-285.

15. Xie, J.R., T. Hong, and J. Stroud, *Long-Term Retail Energy Forecasting With Consideration of Residential Customer Attrition.* Ieee Transactions on Smart Grid, 2015. **6**(5): p. 2245-2252.

16. Ballings, M. and D. Van den Poel, *Customer event history for churn prediction: How long is long enough?* Expert Systems with Applications, 2012. **39**(18): p. 13517-13522.

17. He, B., et al., *Prediction of Customer Attrition of Commercial Banks based on SVM Model.* Procedia Computer Science, 2014. **31**: p. 423-430.

18. Gordini, N. and V. Veglio, *Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry.* Industrial Marketing Management, 2017. **62**: p. 100-107.

19. Coussement, K. and K.W. De Bock, *Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning.* Journal of Business Research, 2013. **66**(9): p. 1629-1636.

20. Smith, K.A., R.J. Willis, and M. Brooks, *An analysis of customer retention and insurance claim patterns using data mining: a case study.* Journal of the Operational Research Society, 2000. **51**(5): p. 532-541.

21. Chen, S.-H., *The gamma CUSUM chart method for online customer churn prediction.* Electronic Commerce Research and Applications, 2016. **17**: p. 99-111.

22. Amin, A., et al., *Customer churn prediction in the telecommunication sector using a rough set approach.* Neurocomputing, 2017. **237**: p. 242-254.

23. Han, J.J., et al., *The Analysis of Logistic Regression in Customers' Churn of VIP Electronic Mailbox.* Icoscm 2007 - International Conference on Operations and Supply Chain Management in China, 2007. **1**.

24. Ballings, M., D. Van den Poel, and E. Verhagen, *Improving Customer Churn Prediction by Data Augmentation Using Pictorial Stimulus-Choice Data.* Management Intelligent Systems, 2012. **171**: p. 217-+.

APPENDIX A - BENCHMARK MODEL PERFORMANCE

		MAE				MSE				MAPE			
		Validation		Test		Validation		Test		Validation		Test	
Dataset	Aggregation Level	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier	Life Tables	Kaplan-Meier
A	7	125.9	130.6	115.6	120.4	24237	25928	18358	19884	0.043	0.045	0.048	0.050
A	14	157.2	167.5	113.0	122.6	35951	40515	17998	20971	0.042	0.045	0.059	0.063
A	21	178.6	196.1	116.0	128.0	43483	52035	19198	22965	0.042	0.047	0.066	0.072
A	28	171.6	190.5	101.9	116.5	40970	50469	14885	18785	0.036	0.042	0.061	0.068
A	35	145.3	161.9	127.7	147.3	32206	42017	22937	29353	0.044	0.051	0.051	0.057
B	7	30.9	46.3	46.4	39.1	1806	3219	3623	2439	0.006	0.005	0.004	0.006
B	14	47.0	54.8	65.5	43.5	4551	4286	6392	2945	0.009	0.006	0.006	0.007
B	21	83.3	62.4	69.7	51.9	11845	5286	7670	3735	0.009	0.007	0.010	0.007
B	28	171.4	54.1	80.4	59.6	39022	4536	9276	4868	0.010	0.007	0.019	0.006
B	35	254.2	93.3	110.0	65.4	79317	17149	14785	7402	0.014	0.008	0.028	0.011
C	7	16.5	18.8	9.3	12.5	516	729	128	264	0.025	0.034	0.033	0.037
C	14	19.6	18.3	9.1	9.6	528	556	102	176	0.023	0.025	0.037	0.033
C	21	25.3	36.5	20.4	16.4	952	2142	697	455	0.045	0.035	0.041	0.058
C	28	29.7	30.1	23.4	18.3	1043	1476	832	530	0.048	0.037	0.045	0.046
C	35	92.4	58.4	24.2	31.6	10536	5262	591	1559	0.043	0.053	0.115	0.071
D	7	61.2	64.8	114.6	103.0	5489	6115	18161	15137	0.057	0.051	0.025	0.026
D	14	60.9	74.2	106.7	91.1	5432	7542	16225	12395	0.049	0.042	0.024	0.029
D	21	60.7	65.7	83.2	81.4	5202	7169	10554	9736	0.035	0.034	0.023	0.025
D	28	50.0	75.2	80.4	91.9	4218	8311	9013	11267	0.031	0.035	0.019	0.028
D	35	86.4	82.2	71.8	80.9	14133	11469	7242	10330	0.025	0.029	0.030	0.029
E	7	111.3	126.1	66.9	63.4	17756	22510	6303	5618	0.013	0.013	0.020	0.023
E	14	103.3	136.2	79.5	75.6	15233	25666	8316	7589	0.016	0.015	0.019	0.025
E	21	86.7	143.8	78.3	93.6	11034	27464	8265	12912	0.015	0.017	0.016	0.026
E	28	61.5	144.8	89.3	137.7	6229	27158	10688	25750	0.016	0.025	0.011	0.025
E	35	43.0	142.0	88.2	183.4	2798	27080	11351	44156	0.016	0.033	0.007	0.025
F	7	65.9	76.7	67.6	77.4	6046	8055	8908	11145	0.018	0.020	0.016	0.018
F	14	69.7	90.4	58.9	79.8	6929	11282	6289	11357	0.015	0.021	0.016	0.021
F	21	78.4	108.2	49.3	68.4	8845	15788	4200	7355	0.012	0.017	0.017	0.024
F	28	87.7	127.4	59.1	81.6	11046	21388	5050	9625	0.015	0.021	0.019	0.028
F	35	75.4	112.9	69.9	109.7	8096	18141	6873	15152	0.017	0.026	0.016	0.024
G	7	108.0	98.4	53.7	55.8	18081	14855	5000	4490	0.052	0.054	0.091	0.083
G	14	90.0	100.6	117.4	120.6	12019	15504	20822	18971	0.107	0.111	0.065	0.069
G	21	56.0	108.0	145.7	107.5	4473	15637	27916	16663	0.131	0.098	0.035	0.069
G	28	47.2	55.8	114.0	147.2	3278	6110	23591	26658	0.089	0.107	0.026	0.033
G	35	120.7	47.8	33.7	135.9	17885	3103	1888	22537	0.021	0.085	0.063	0.025
H	7	68.7	84.6	102.3	106.4	7677	11226	15056	16299	0.009	0.009	0.007	0.008
H	14	65.4	96.6	89.6	107.1	7794	15765	11906	16692	0.008	0.009	0.006	0.009
H	21	63.0	98.4	75.7	112.4	6164	16101	8434	17739	0.006	0.009	0.006	0.009
H	28	75.2	104.4	84.9	132.3	8163	18348	10576	25022	0.007	0.011	0.007	0.010
H	35	56.6	100.7	86.7	153.1	4800	16440	10760	35062	0.007	0.013	0.006	0.010
I	7	75.5	65.6	86.9	97.6	8346	6662	11543	13942	0.018	0.020	0.015	0.013
I	14	118.0	87.3	81.2	107.6	18005	10940	10866	16917	0.017	0.022	0.023	0.017
I	21	139.3	95.4	67.7	84.1	24016	12352	7040	11166	0.013	0.016	0.026	0.018
I	28	146.7	81.2	68.8	78.3	27088	10000	6627	11425	0.013	0.015	0.027	0.015
I	35	172.7	118.8	91.2	90.9	42538	18886	11116	14278	0.017	0.017	0.031	0.021
J	7	50.7	60.2	53.4	56.8	4009	5444	6587	6841	0.025	0.026	0.021	0.025
J	14	132.9	162.6	30.9	45.5	23258	34171	1735	3202	0.013	0.020	0.051	0.063
J	21	123.0	172.0	49.7	79.9	19756	36178	3983	8753	0.020	0.032	0.046	0.064
J	28	55.7	143.6	145.8	218.7	4939	25167	24755	54002	0.056	0.083	0.019	0.049
J	35	306.6	312.9	183.5	286.4	156580	139722	38904	92631	0.070	0.109	0.097	0.101

APPENDIX B - MODIFIED LIFE TABLES TEST PERFORMANCE COMPARISON

			MAE		MSE		MAPE	
Dataset	Aggregation Level	Selected Entry Range (L_1)	Benchmark	Modified	Benchmark	Modified	Benchmark	Modified
A	7	50	115.6	71.5	18358	7789	0.043	0.027
A	14	2	113.0	99.1	17998	13071	0.042	0.037
A	21	2	116.0	110.2	19198	17637	0.042	0.040
A	28	2	101.9	82.8	14885	10405	0.036	0.030
A	35	2	127.7	115.6	22937	19867	0.044	0.040
B	7	66	46.4	46.4	3623	3623	0.006	0.006
B	14	26	65.5	65.5	6392	6392	0.009	0.009
B	21	14	69.7	47.3	7670	2954	0.009	0.006
B	28	12	80.4	50.2	9276	3251	0.010	0.006
B	35	8	110.0	50.0	14785	3200	0.014	0.006
C	7	17	9.3	9.3	128	128	0.025	0.025
C	14	8	9.1	9.1	102	102	0.023	0.023
C	21	10	20.4	21.3	697	744	0.045	0.047
C	28	4	23.4	24.1	832	899	0.048	0.049
C	35	2	24.2	25.6	591	666	0.043	0.045
D	7	9	114.6	114.6	18161	18161	0.057	0.057
D	14	14	106.7	174.1	16225	39397	0.049	0.079
D	21	5	83.2	118.1	10554	19986	0.035	0.050
D	28	4	80.4	75.9	9013	11306	0.031	0.030
D	35	6	71.8	71.8	7242	7242	0.025	0.025
E	7	37	66.9	93.8	6303	13877	0.013	0.019
E	14	6	79.5	91.4	8316	13317	0.016	0.018
E	21	2	78.3	89.1	8265	8094	0.015	0.017
E	28	2	89.3	89.3	10688	10688	0.016	0.016
E	35	2	88.2	88.2	11351	11351	0.016	0.016
F	7	24	67.6	55.1	8908	5472	0.018	0.014
F	14	2	58.9	54.7	6289	5678	0.015	0.014
F	21	2	49.3	30.6	4200	2327	0.012	0.008
F	28	2	59.1	39.5	5050	3061	0.015	0.010
F	35	6	69.9	83.8	6873	9108	0.017	0.020
G	7	14	53.7	53.7	5000	5000	0.052	0.052
G	14	6	117.4	121.4	20822	20822	0.107	0.110
G	21	4	145.7	138.6	27916	28036	0.131	0.124
G	28	2	114.0	107.2	23591	23468	0.089	0.085
G	35	2	33.7	53.1	1888	3765	0.021	0.046
H	7	101	102.3	113.9	15056	26120	0.009	0.010
H	14	38	89.6	71.6	11906	9049	0.008	0.006
H	21	2	75.7	81.5	8434	10389	0.006	0.007
H	28	2	84.9	91.6	10576	12661	0.007	0.008
H	35	2	86.7	93.6	10760	12461	0.007	0.008
I	7	101	86.9	86.9	11543	11543	0.018	0.018
I	14	50	81.2	81.2	10866	10866	0.017	0.017
I	21	30	67.7	96.2	7040	15485	0.013	0.019
I	28	20	68.8	70.8	6627	9352	0.013	0.014
I	35	17	91.2	76.5	11116	9260	0.017	0.014
J	7	14	53.4	42.0	6587	5989	0.025	0.020
J	14	8	30.9	38.7	1735	2704	0.013	0.016
J	21	6	49.7	28.2	3983	1154	0.020	0.011
J	28	2	145.8	37.2	24755	2162	0.056	0.014
J	35	2	183.5	49.1	38904	38904	0.070	0.019

APPENDIX C - REGRESSION HAZARD TEST PERFORMANCE COMPARISON

		MAE		MSE		MAPE	
Dataset	Aggregation Level	Benchmark	Regression	Benchmark	Regression	Benchmark	Regression
A	7	115.6	86.8	18358	8389	0.043	0.035
A	14	113.0	78.1	17998	7263	0.042	0.033
A	21	116.0	75.3	19198	6661	0.042	0.033
A	28	101.9	74.0	14885	7415	0.036	0.029
A	35	127.7	92.9	22937	11235	0.044	0.037
B	7	46.4	44.9	3623	3350	0.006	0.006
B	14	65.5	62.1	6392	5904	0.009	0.008
B	21	69.7	65.7	7670	6820	0.009	0.009
B	28	80.4	77.7	9276	8581	0.010	0.010
B	35	110.0	106.5	14785	13826	0.014	0.013
C	7	9.3	9.1	128	136	0.025	0.025
C	14	9.1	9.1	102	112	0.023	0.021
C	21	20.4	22.3	697	836	0.045	0.043
C	28	23.4	27.2	832	1070	0.048	0.045
C	35	24.2	26.5	591	683	0.043	0.045
D	7	114.6	121.4	18161	20704	0.057	0.058
D	14	106.7	113.8	16225	19524	0.049	0.050
D	21	83.2	81.1	10554	10353	0.035	0.035
D	28	80.4	75.3	9013	7733	0.031	0.031
D	35	71.8	73.8	7242	8178	0.025	0.025
E	7	66.9	68.9	6303	7011	0.013	0.014
E	14	79.5	80.4	8316	9126	0.016	0.016
E	21	78.3	78.7	8265	8872	0.015	0.017
E	28	89.3	90.9	10688	11353	0.016	0.019
E	35	88.2	91.7	11351	12396	0.016	0.019
F	7	67.6	59.3	8908	6403	0.018	0.017
F	14	58.9	52.5	6289	4417	0.015	0.015
F	21	49.3	44.4	4200	2814	0.012	0.013
F	28	59.1	53.6	5050	3421	0.015	0.015
F	35	69.9	51.1	6873	2696	0.017	0.016
G	7	53.7	51.8	5000	4571	0.052	0.053
G	14	117.4	122.3	20822	25777	0.107	0.109
G	21	145.7	155.4	27916	31575	0.131	0.138
G	28	114.0	119.4	23591	24460	0.089	0.093
G	35	33.7	33.0	1888	1861	0.021	0.020
H	7	102.3	95.8	15056	12532	0.009	0.009
H	14	89.6	83.9	11906	10497	0.008	0.007
H	21	75.7	70.3	8434	7395	0.006	0.006
H	28	84.9	78.1	10576	8835	0.007	0.007
H	35	86.7	78.6	10760	8275	0.007	0.007
I	7	86.9	83.8	11543	11241	0.018	0.017
I	14	81.2	80.6	10866	12561	0.017	0.016
I	21	67.7	63.7	7040	7532	0.013	0.013
I	28	68.8	64.5	6627	6889	0.013	0.012
I	35	91.2	95.9	11116	13675	0.017	0.017
J	7	53.4	52.7	6587	7010	0.025	0.024
J	14	30.9	30.8	1735	1791	0.013	0.013
J	21	49.7	44.8	3983	3763	0.020	0.015
J	28	145.8	159.6	24755	28080	0.056	0.060
J	35	183.5	192.1	38904	41665	0.070	0.069

APPENDIX D - REGRESSION HAZARD TEST ONE-STEP AHEAD FORECAST PERFORMANCE COMPARISON

		MAE		MSE		MAPE	
Dataset	Aggregation Level	Benchmark	Regression	Benchmark	Regression	Benchmark	Regression
A	7	115.6	100.6	18358	12959	0.043	0.038
A	14	113.0	95.7	17998	10974	0.042	0.036
A	21	116.0	96.3	19198	10321	0.042	0.037
A	28	101.9	83.8	14885	8084	0.036	0.032
A	35	127.7	91.7	22937	9019	0.044	0.037
B	7	46.4	40.1	3623	2585	0.006	0.006
B	14	65.5	61.2	6392	5316	0.009	0.008
B	21	69.7	71.1	7670	7757	0.009	0.009
B	28	80.4	78.1	9276	8986	0.010	0.010
B	35	110.0	103.5	14785	14042	0.014	0.013
C	7	9.3	9.0	128	131	0.025	0.024
C	14	9.1	8.9	102	95	0.023	0.021
C	21	20.4	21.5	697	724	0.045	0.041
C	28	23.4	26.8	832	1032	0.048	0.044
C	35	24.2	25.4	591	647	0.043	0.048
D	7	114.6	104.3	18161	15808	0.057	0.051
D	14	106.7	109.7	16225	18843	0.049	0.048
D	21	83.2	71.7	10554	6689	0.035	0.031
D	28	80.4	71.7	9013	7779	0.031	0.028
D	35	71.8	76.0	7242	7012	0.025	0.026
E	7	66.9	62.6	6303	5937	0.013	0.012
E	14	79.5	75.5	8316	7961	0.016	0.015
E	21	78.3	73.0	8265	8194	0.015	0.015
E	28	89.3	85.0	10688	9354	0.016	0.018
E	35	88.2	88.7	11351	12790	0.016	0.018
F	7	67.6	65.0	8908	8969	0.018	0.017
F	14	58.9	57.9	6289	6203	0.015	0.015
F	21	49.3	49.9	4200	4329	0.012	0.013
F	28	59.1	58.2	5050	4543	0.015	0.015
F	35	69.9	57.5	6873	3427	0.017	0.017
G	7	53.7	53.3	5000	5358	0.052	0.052
G	14	117.4	123.8	20822	25982	0.107	0.106
G	21	145.7	155.7	27916	32127	0.131	0.140
G	28	114.0	121.0	23591	24528	0.089	0.095
G	35	33.7	32.9	1888	1809	0.021	0.020
H	7	102.3	92.2	15056	12936	0.009	0.008
H	14	89.6	81.4	11906	9942	0.008	0.007
H	21	75.7	68.6	8434	7414	0.006	0.006
H	28	84.9	79.8	10576	10186	0.007	0.007
H	35	86.7	79.4	10760	9420	0.007	0.007
I	7	86.9	87.1	11543	12270	0.018	0.017
I	14	81.2	80.6	10866	11034	0.017	0.016
I	21	67.7	63.5	7040	6498	0.013	0.013
I	28	68.8	60.4	6627	6074	0.013	0.012
I	35	91.2	85.9	11116	11947	0.017	0.016
J	7	53.4	52.0	6587	9360	0.025	0.020
J	14	30.9	30.1	1735	2023	0.013	0.011
J	21	49.7	48.8	3983	4598	0.020	0.015
J	28	145.8	161.4	24755	28245	0.056	0.063
J	35	183.5	178.9	38904	38717	0.070	0.061