Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 31-47

Predictive Modelling of Lapsation of Life Insurance Policies in India

Gurprit Grover[1], Vajala Ravi[2], Richa Saini[1] and Manoj Kumar Varshney[3]

[1]*Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, India*
[2]*Department of Statistics, Lady Shri Ram College for Women, University of Delhi, India.*
[3]*Department of Statistics, Hindu College, University of Delhi, India*

Received: 06 April 2021; Revised: 29 May 2021; Accepted: 03 June 2021

Abstract

Retaining customers is the biggest challenge in Indian insurance sector. Customer Relationship Management (CRM) department in every company plays a role of a platform from customer to company which informs insurance company about the needs of customers to be satisfied. Moreover, all the insurance companies are facing the problem of low persistency or high lapsation. Predictive modelling helps in classification of any policy as in force or lapsed. Predictive classification may be of help to insurer in identifying the groups of insured with various characteristic which may not frequently experience the lapsation. Such information can be found as useful for targeting the segment of society which is insurance minded and afford to keep policy in force. We have applied four different techniques for predictive modelling namely Cox PH model, Gompertz law of mortality, Bayesian networks using naïve Bayesian technique and random forests technique and the best fitted model is identified based on confusion matrix and error rate of misclassification. On applying the best fitted model to the data set of 3663 policies, it is found that nearly 54 per cent policies are classified as lapsed and 46 per cent policies are classified as in force. The major contributors in classifying the policies are age of the policyholders, sum assured (SA), policy term, occupation and income of the policyholders.

Key words: Predictive modelling; Naïve Bayesian model; Random forest; Gompertz; Cox PH model; Persistency.

1. Introduction

1.1. Outline

Loss of customers' confidence in company is distressing for any type of business. From a management perspective, quality of sales is to be ensured highly. For life insurance Company, quality of sales is ensured if its customers are persistent or retained with the same company till the end of the term of insurance contract. If policies are being terminated due to non-payment of premiums then it costs high to all who are involved in a contract including insured, insurer and the agent. Over the decade, the severe problem of low persistency has been experienced by every life insurance company in Indian sector. CRM understands the need of customers and provides the same information to the company to retain its customers. The services provided by CRM mainly involve consultation, execution, subcontracting and

[1]Disclaimer: The views expressed in the paper are those of the authors and not necessarily those of the institution to which they belong. Moreover, data have also been obtained solely by surveys of different types and do not pertain to any specific institution.

Corresponding Author: Richa Saini
Email Id: richarawat55@gmail.com

training. But still the insurance companies are facing the challenge of customers' churn. For handling the problem of low persistency, a large sample of policies that are procured in a particular financial year, is observed every year until the fifth year after inception. For a deep dive in to the problem, it is important to identify the contribution of each working factor and better techniques for predictive modelling. In such a case, data mining techniques have evidently established the reliable results with great accuracy. These techniques are valid in our case too. In the present study we have made a comparison of different techniques used for predictive modelling and have applied the famous Cox PH model, Gompertz Curve, Naïve Bayesian model and Random forest technique. These four models are chosen to compare the convention with the advancement. Former two models are the conventional survival and actuarial models which have been established as most popular for modelling of survival data and often they are used for predictions also. For classification purpose, these two models have utility but the algorithm is complex. Over the last two decades, the latter two models have gained much attention by researchers for predictive modelling and classification. These models are the results of advancement in machine learning algorithms. These models do not assume any form of underlying distribution or specific structure but learn the features of the best fit model from the data itself. The analysis is performed using the statistical software *R*.

1.2. Review of literature

Previously many pioneering investigations have been conducted for predictive modelling of customers' churn in various fields including insurance. (Tirenni *et al*., 2007) drew out the general methodology for categorization of customers according to their life time value and also forecasted their lifetime values based on demographic and behavioral characteristics. They addressed the problem of lifetime values of airline customers using decision trees and classified the customers as long, medium and short term customers. Their predictions are based on limited information. (Jasek *et al*., 2018) provided the life time value models in the field of E commerce and discussed their forecasting abilities. They applied extended Pareto, Markov chain and status quo models and compared their results. (Huigevoort, 2015) carried out the predictive modelling of customer churn for a health insurance company. He utilized the four data mining techniques namely logistic regression, decision trees, neural network and support vector machine (SVM) to identify important churning variables and characteristics. (Adebiyi *et al*., 2016) explored the blend of two models namely Analytic Hierarchy Process (AHP) and Markov chains (MCM) for tackling the problem of customers' churn. Their study recommended the organizational strategies that reverse the churn alternatives with high priority and improves service delivery. (Zhang *et al*., 2017) employed the Deep and Shallow model for predictive modelling of insurance churn. Their proposed model yields enhanced performance as compared to Deep models or Shallow models, if applied individually. (Lariviere and Van Den Poel, 2005) predicted customers' withholding and profitability using Random forest and Regression forest techniques. In their study they considered three important measures of customer outcome *viz.* subsequent buy, fractional defection and customers' prosperity by employing random forest and regression techniques. They concluded that the results will be improved if behavioral outcome variables are also taken into consideration along with demographic factors. (Bandyopadhyay *et al*., 2014) offered a machine learning approach based on Bayesian networks (BN) with an application on Electronic health data (EHD) to forecast the probability of encountering a cardiovascular incident within next five years. In EHD, censoring exists as an inevitable feature and they described how to alter both the modelling and estimation techniques to account for censoring. Their proposed model is an improvement over Cox PH model or Bayesian networks with informal approach to right censoring. (Onisko *et al*., 2001) proposed a method that utilizes the Noisy OR gates to minimize the data requirements in learning

conditional probabilities. Their proposed method is proved to be better than the simple multiple disorders model and a single disorder diagnosis model. (Rudolph, 2002) presented in first of his paper, an analysis of part of the long-term care insurance portfolio using Cox Proportional Hazard model to estimate transition intensities. Rudolph showed that the approach allowed the inclusion of censored as well as time dependent factors at risk. In his second part of the paper, it was shown how the evaluated intensities can be utilized in a multiple state model to calculate premiums. (Gustafsson, 2009) applied survival analysis to predict low retention in non-life insurance industry. He applied statistical models to estimate the survival probabilities on customer level in a competing risk framework, where low retention could be the consequences of different types of causes. (Richmond and Roehner, 2016) used the famous Gompertz law for prediction of human mortality beyond the age of 100 years and established that the law also works well for ages over 100 years. They carried out the transversal analysis for a sample of industrialized and developed countries. All these revolutionary researches inspired the authors to carry out the present study in Indian scenario.

The paper ahead is organized as follows: Section 2 deals with the brief discussion of the evaluated and applied techniques for predictive modelling. Section 3 provides the description of data and methodology. In section 4, the results are discussed followed by conclusions in section 5.

2. Survival, Actuarial and Machine Learning Techniques for Predictive Modelling

2.1. Cox proportional hazard (PH) model (Lee and Wang, 2003)

Cox PH model is a popular tool in survival analysis for modelling the relationship between survival time and covariates. This is a semi parametric approach and assumes that the form of hazard function for the random lifetime T be given by,

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \le T \le t + dt \mid T > t)}{dt} \tag{1}$$

which is the product of a baseline hazard function $\lambda_0(t)$ and a specific scaling factor depending upon covariates and it is of the form given below,

$$\lambda(t \mid Z = z) = \lambda_0(t) \exp(\beta^t Z) \tag{2}$$

where, $z \in R^p$ denotes the observed vector of covariates and $\beta \in R^p$ denotes the unknown regression coefficient.

Assuming no tied survival times, the estimation procedure is as follows:

Suppose that the k ($< n$) survival times of n individuals are distinct and uncensored, and the survival times of remaining $n - k$ individuals are right-censored. Let $t_{(1)} < t_{(2)} < t_{(3)} \ldots < t_{(k)}$ be the k distinct ordered survival times with corresponding predictors $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$. Let the risk set at time $t_{(i)}$ be denoted by $\boldsymbol{R(t_{(i)})}$. Then, $\boldsymbol{R(t_{(i)})}$ consists of all persons whose survival times are at least up to $t_{(i)}$. Then, the conditional probability of failure at time $t_{(i)}$ is given by

$$\frac{\exp\left(\underset{\sim}{b}' X_{(i)}\right)}{\sum_{l \in R(t_{(i)})} \exp\left(\underset{\sim}{b}' X_{(l)}\right)} \tag{3}$$

And hence the partial likelihood function is given by,

$$L(\underset{\sim}{b}) = \prod_{i=1}^{k} \frac{\exp\left(\underset{\sim}{b}' X_{(i)}\right)}{\sum_{l \in R(t_{(i)})} \exp\left(\underset{\sim}{b}' X_{(l)}\right)} \tag{4}$$

Then, (5) will give the maximum partial likelihood estimates (MPLEs) $\hat{\underset{\sim}{b}}$ by employing Newton Raphson iterated procedure.

$$\frac{\partial\left(log\, L\left(\underset{\sim}{b}\right)\right)}{\partial \underset{\sim}{b}} = 0 \tag{5}$$

2.2. Gompertz mortality law

The probability density function (pdf) of Gompertz distribution with location parameter *a* and shape parameter *b* is given by,

$$f(x) = ae^{bx - \frac{a}{b}(e^{bx}-1)}, \; x[0, \infty) \tag{6}$$

with the distribution function *F(x)*and the hazard function μ:

$$F(x) = 1 - e^{-\frac{a}{b}(e^{bx}-1)} \tag{7}$$

$$\mu(x) = ae^{bx}; a, b > 0 \tag{8}$$

Then, the revised maximum likelihood estimators (Lenart, 2012) of parameters *a* and *b* are given by (9) and (10).

$$\hat{a} = \frac{\sum_{x} D_x}{\sum_{x} E_x e^{bx}} \tag{9}$$

Estimator of *b* can be obtained solving numerically the following equality,

$$\frac{\sum_{x} D_x x}{\sum_{x} D_x} = \frac{\sum_{x} E_x e^{\hat{b}x} x}{\sum_{x} E_x e^{\hat{b}x}} \tag{10}$$

They provided the estimators for both the discrete and continuous ages. For discrete ages, if the number of deaths and the number of person years exposed to the risk of dying should be available.

2.3. Bayesian networks using Naïve Bayes classifier (Rao and Rao, 2014)

The Bayesian networks are probabilistic in nature and can model the joint probability density function (pdf) over a finite number of random variables. Bayesian network is represented as directed acyclic graph and it is often abbreviated as DAG. Its nodes are the random variables and the directed arcs express the dependencies among the random variables. A conditional probability table (CPT) of a variable *X* consists of probability distributions over the different states of random variable *X* for all possible combinations of

the states of *X's* parents. The joint probability distribution over all the random variables under the network can also be calculated by taking the product of all the priors and conditional pdfs. These networks deal with the complexity in the model and offer great flexibility. Therefore, these are more comprehensive than the other conventional survival models. Bayesian networks can be used in situations where little or no data are available. These models are based on the Bayes' theorem which is mathematically stated as,

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} \tag{11}$$

where A and B are events and $P(B) \neq 0$.

For an instance, the simple Bayesian network in life insurance can be represented as follows:

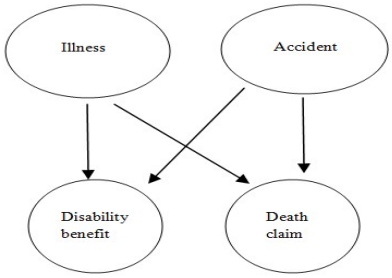


Figure 1: Simple Bayesian network

Figure 1 above shows the simple cause and effect form of Bayesian network which has two causes *viz.* illness or accident and two effects as claim for disability benefit or death. Suppose that the following probabilities are known:

P(accident), P(illness), P(disability benefit | accident), P(disability benefit | illness), P(death claim | accident) and P(death claim | illness), P(disability benefit | no accident), P(disability benefit | no illness), P(death claim | no accident) and P(death claim | no illness). Then, the conditional probability table (CPT) corresponding to any effect or node (disability benefit or death claim) given the cause or parent node information (illness or accident) can be constructed. One of such CPTs can be evaluated corresponding to disability benefit which is cited below in Table 1:

Table 1: Conditional probability table for disability benefit

Accident	Illness	Disability benefit
True	True	P_1
True	False	P_2
False	True	P_3
False	False	P_4

In Table 1 above the conditional probabilities for disability benefit have been denoted by P_1, P_2, P_3 and P_4. These probabilities are conditionally dependent upon their parent nodes, accident and illness. Other such CPTs are also evaluated for rest of the effect nodes and then any type of joint or conditional probabilities can be evaluated.

The best Bayesian network is generally fitted by learning from the data and maximizing the entropy scoring function (Cheng and Greiner, 2013). In the present study, we have

utilized the Bayesian networks (BN) Naïve Bayes classifier. Naïve Bayes classifiers have proved to be influential tools for solving classification problems in every field. Naïve Bayes classifier is the simplest probabilistic model based on the Bayes' theorem with strong independence assumptions between the characteristics (also known as predictors in case of regression).

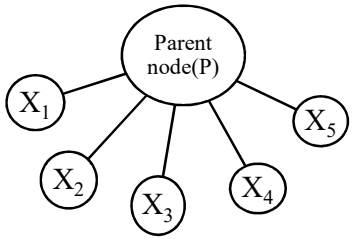


Figure 2: Structure of Naïve Bayes classifier

The Naïve Bayes classifier is the easy structure that has only one parent node of all the other next generation nodes. Graphically its structure is represented above in Figure 2.

In Figure 2 above, there is only one parent node denoted by *P* and all the other nodes X_1, X_2, X_3, X_4 and $X_5$are the next generation nodes (Cheng and Greiner, 2013).

For instance, consider the above Bayesian network with disability benefit and death claim as next generation nodes and accident as their only parent. Then, the structure of Naïve Bayes classifier is as follows (Figure 3):

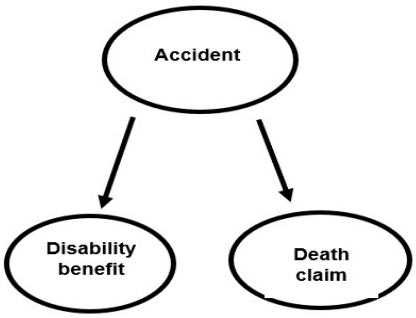
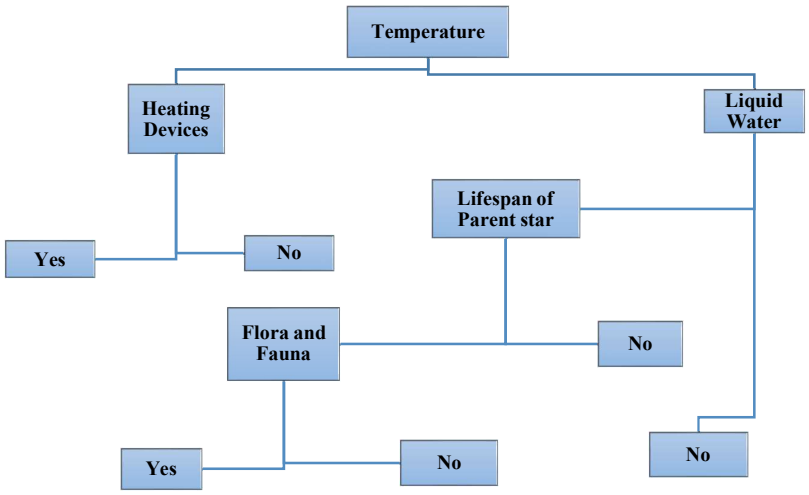


Figure 3: Structure of Naïve Bayes classifier

2.4. Random forest technique

The random forest technique is first introduced by Bell lab researcher (Ho, 1995). Random forest or random decision forest is a collaborative learning method for classification, regression and other jobs, that function by building a multitude of decision trees at training time and producing the class that is the type of the classes (classification) or mean forecasting (regression) of the individual tree. Random forest technique is more accurate for prediction, avoid over fitting and include bagging also. (Breiman, 2001) has compared empirically the results of the random forest technique with two different random features of selection from

the original inputs and the random linear combinations of inputs. When we make a prediction, the new observation gets strapped down each decision tree and allocated a predicted value. For example, suppose we want to know whether the planet is habitable or not. Then, the following type of decision tree helps in predicting. If the temperature on a planet is 150 Kelvin and there are no heating devices, then can it be predicted as habitable or not habitable? Figure 4 below shows the decision tree.



(Source: http://www.machinelearningtutorial.net/2017/01/23/randomforest-basics/)

Figure 4: Decision tree

Once each of the possible trees as shown above (in Figure 4) in the forest has reported its forecasted value, the predictions are matched up for the final prediction. In the case of **decision trees**, a simple majority vote (mode) controls the output, whereas in the case of a **regression trees**, the mean predictions of all individual trees form the final prediction.

3. Data and Employed Methodology

Data has been obtained by conducting a survey around Delhi NCR and from commercial sources of business. Data remains confidential and unpublished as it has been obtained from different surveys and is not available in the public domain. Data is spatial too, as it has been collected over the northern region of India. Data consists of 3363 policies which were procured in 2014 - 2015 with a balanced mix of areas of habitat; various products offered by insurance companies like Term, Savings, ULIP, Health and Pension; various ages of policy holders, income levels, sum assured (SA) *etc*. The categorical variables are assigned appropriate codes according to each level of such variables.

As there is a facility of payment of premium during grace period of one month thus, policies with the date of first unpaid premium (FUP) in the months of May and June of any year (time of assessment of policies) are assumed to be in force policies.

The methodology for implementation of the four models discussed above is given below:

3.1. Cox PH model

Let the predictor age be denoted by Z_{1t}, area by Z_2, income band by Z_{3t}, occupation by Z_4, plan type by Z_5, mode by Z_6, channel by Z_7, sum assured (SA) by Z_{SA}, gender by Z_{Gen} and policy term by Z_{8t}.

We first obtained the pair wise correlations among the predictors and then fitted the following Cox PH model:

$$\lambda(t \mid Z = z) = \lambda_0(t)\exp(\beta^t Z) \tag{12}$$

where, $\beta^t = (\beta_1, \beta_2, \beta_3, ..., \beta_n)$ and $Z = (Z_{1t}, Z_2, Z_{3t}, Z_4, Z_5, Z_6, Z_7, Z_{8t}, Z_{1t}Z_4Z_{8t}, Z_{1t}Z_4Z_7Z_{8t}, Z_{1t}Z_2Z_{3t}Z_4Z_5, Z_{1t}Z_2Z_{3t}Z_5Z_6, Z_{1t}Z_2Z_4Z_5Z_6, Z_{1t}Z_2Z_{3t}Z_5Z_{8t}, Z_{1t}Z_4Z_5Z_7Z_{8t})$.

Equation (12) is the form of hazard function under the Cox PH model which is expressed as the product of baseline hazard $\lambda_0(t)$ and the exponential function of the prognostic factors Z_i's which is denoted by $exp(\beta^t Z)$. Elements of vector Z include main predictors and various interactions of among these predictors like age:occupation:Policyterm is the third order interaction among age, occupation of policyholder and policy term and it is denoted by $Z_{1t}Z_4Z_{8t}$. Similarly, other higher order interaction terms have been used. We have also obtained the residuals using Deviance approach and predicted values of the type "expected" for the fitted model. Deviance residuals are symmetric unlike Martingale residuals. Using deviance residuals and predicted values, we have obtained the plot of Residuals *v/s* Predicted values. Such a plot is used diagnostic checking of the Cox PH model. That means we can assess whether the assumption of linearity is fulfilled or not.

3.2. Gompertz mortality law

Gompertz law of mortality for various predictors namely age, SA, area, gender and certain combinations of these predictors to estimate the probability of death which is lapsation in our case is defined below in four different forms.

Form 1: When mortality depends upon age only –

$$\mu(x) = ae^{bx}; a, b > 0 \text{ where, } x \text{ denotes age only.} \tag{13}$$

Form 2: When mortality depends upon SA only –

$$\mu(x) = ae^{bx}; a, b > 0 \text{ where, } x \text{ denotes SA only.} \tag{14}$$

Form 3: When mortality depends upon Area, Age and Gender –

$$\mu(x) = ae^{bx}; a, b > 0 \text{ where } x \text{ denotes interactive effect area*age*Gender} \tag{15}$$

Form 4: When mortality depends upon Area and SA:

$$\mu(x) = ae^{bx}; a, b > 0 \text{ where } x \text{ denotes the interactive effect area*SA} \tag{16}$$

The validity of these forms is checked by standardized residuals plot.

3.3. Bayesian networks using Naïve Bayes classifier

We first obtained the pairs and panel graphs before applying Naïve Bayes model which helps in ascertaining whether the predictors are linearly independent among themselves. This assumption has been earlier verified using the pair wise correlations among the predictors before fitting the Cox PH model. We then split the data in to two datasets namely training and testing in the ratio70:30. After splitting, we fitted the Naïve Bayes to the training data set with all the predictors. We further evaluated confusion matrix and error rate of misclassification for the model and compared them with the predicted results for testing data set.

3.4. Random forest technique

The entire dataset was divided into two parts in the ratio 70:30. Here, the first 70 *per cent* of the data are treated as training set and remaining 30 *per cent* treated as testing data set. The predicted variable is lapsation status of a policy which is coded as a factor variable with two levels *viz.,* 0, if policy is in force and 1, if it is lapsed. We then fitted the random forests model with various numbers of decision trees but with 20 trees, the model is found to be fitted well. We also obtained the plots of error rate for the model, proportion of predicted values for testing data and training data both. We also evaluated confusion matrix for the model to assess the goodness of fit of the model for predictive modelling.

4. Results and Discussions

4.1. Cox PH model

Pair wise correlations have been first calculated and Table 2 below shows the results:

Table 2: Pair wise Correlations

	Area	Plan type	Age	Mode	Channel	SA	Policy term	Gender	Occu.	Income Band
Area	1									
Plan type	0	1								
Age	–0.01	–0.02	1							
Mode	–0.02	–0.09	–0.01	1						
Channel	0	–0.02	0	0	1					
SA	0.07	–0.02	0.02	–0.08	–0.03	1				
Policy Term	–0.01	0.13	0.03	–0.01	–0.02	-0.01	1			
Gender	0.04	–0.03	0.03	–0.02	0.08	0.04	0	1		
Occupation (Occu.)	0.01	0.01	0.01	0	–0.01	0	0.02	0.02	1	
Income Band	0	–0.01	0	0.01	0.02	0.01	0.00	0	–0.01	1

The Table 2 above shows that all the pair wise correlations are very low, ensuring the linear independence among all the predictors and now we can proceed to fit the Cox PH model. Table 3 below presents the coefficients, exponential and *SE* of coefficients and *P* values:

Table 3:Results of Cox PH model

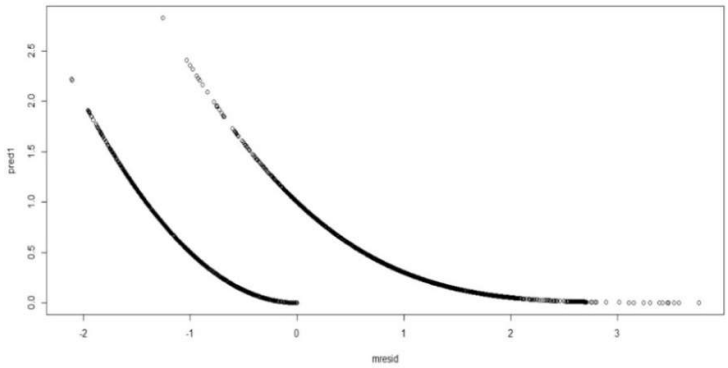
Effect	Notations	Coefficient	*SE* (Coeff)	*p*-value
Age	Z_{1t}	-2.828	64.58	0.9651
Area	Z_2	-10.860	3443.00	0.9975
Income_band	Z_{3t}	93.150	923.30	0.9196
Occupation	Z_4	-202.900	1034.00	0.8445
Plan_type	Z_5	56.790	3427.00	0.9868
Mode	Z_6	-56.170	1523.00	0.9706
Channel	Z_7	-320.800	3440.00	0.9257
Policy_term	Z_{8t}	-19.700	136.60	0.8853
age:occupation:Policy_term	$Z_{1t}Z_4Z_{8t}$	-0.305	0.168	*0.0702
age:occupation:channel:policy_term	$Z_{1t}Z_4Z_7Z_{8t}$	0.362	0.158	**0.0218
age:area:income_band:occupation:plan_type	$Z_{1t}Z_2Z_{3t}Z_4Z_5$	0.864	0.353	**0.0143
age:area:income_band:plan_type:mode	$Z_{1t}Z_2Z_{3t}Z_5Z_6$	1.326	0.472	***0.0050
age:area:occupation:plan_type:mode	$Z_{1t}Z_2Z_4Z_5Z_6$	1.866	0.676	***0.0058
age:area:income_band:plan_type:policy_term	$Z_{1t}Z_2Z_{3t}Z_5Z_{8t}$	0.113	0.036	***0.0017
age:occupation:plan_type:channel:policy_term	$Z_{1t}Z_4Z_5Z_7Z_{8t}$	-0.201	0.065	***0.0019

*denotes significance at 10*per cent* level of significance (los), **denotes significance at 5*per cent* level of significance, ***denotes significance at 1*per cent* level of significance (los).

The Table 3 above evidently exhibits that the main factors like age, area, income band *etc.* are all insignificant with probability values (*p*-values) more than 0.8 and all the interaction effects are significant at 10, 5 and 1 *per cent* level of significance. We observe that age has the coefficient – 2.828 and *p*-value is 0.9651 (> 0.05) which means that age is insignificant at 5 *per cent* level of significance for classifying the policies. Similarly, we can observe that the other main factors are also insignificant. On the other hand, the interaction effect of, say, age; occupation; and policy term is significant at 10 *per cent* level of significance with *p*-value 0.0702. The other interaction effects are also significant at different level of significance, like the interaction between age; occupation; channel; and policy term is significant at 5 *per cent* level of significance with *p*-value 0.0218 *etc*. This implies that the factors individually have no influence on the status of policy but taking together three or more factors is effective and helps in classification. In fact, as the order of interaction increases we observe that *p*-value is kept on decreasing and the effect is found to be significant.

The concordance in survival analysis is defined as measure of extent of agreement between risk score and time until failure. It was first popularized by (Harrell, Lee and Mark, 1996) in survival analysis and it is a statistic for Cox PH model. Now days, it is most widely used as a measure of goodness of fit in survival models. Not only in survival models but it has utility in logistic and ordinary linear regressions as well, (Therneau and Atkinson, 2020). The model with value of concordance more than 80 *per cent* is considered as a good fit but in our case the concordance value is evaluated as 62.8 *per cent* implying that the model is not a good fit. Other tests like Wald test, Likelihood ratio test and Score rank test with *p*-values 0.0001 also yield the same result. R^2 is obtained as 0.175 which seems to be inadequate. Figure 5 shows that the plot of deviance residuals also supports the same argument against the Cox PH model.

It can be apparently observed in Figure 5that residuals are following pattern which is sloping downwards from left to right. This curvilinear pattern in residuals leads us to conclude that Cox PH model is not a good fit in insurance phenomenon.



Please note that both the curves show the plot of deviance residuals against predicted values of response variable Y. Some coordinates lie on the first curve and others lie on the second curve.

Figure 5: Deviance residuals under Cox PH model

4.2. Gompertz Curve

We employed four different forms of Gompertz law of mortality for which results are tabulated below in Table 4. The *AIC* values for form 1 (with age only) and form 3 (with SA only) are shown as they are generated by the system. But the *AIC* values for form 2 (with age, area and gender) and form 4 (with area and SA) are obtained using the weighted average of the *AIC* values under the different combinations of factors that are affecting the status of policies. These various possible combinations under each form of Gompertz law are also displayed in column 2 of Table 4. For instance, to compute the *AIC* value for form 2, we calculated the weighted average four *AIC* values obtained by fitting the Gompertz law to the four combinations given below:

Male lives in rural area with all ages; male lives in urban area with all ages; female lives in rural area with all ages; and female lives in urban area with all ages.

We can also calculate the probability of lapsation for these models as weighted average of the probabilities obtained under each combination using respective estimates of *a* and *b*.

From Table 4, we observe that all the *AIC* values are very low. Form 1 and form 3 have the lowest *AIC* values –12.1 and –12.17. These forms do not involve any combination like forms 2 and 4. It is also observed that the estimate of parameter *b* is too small for all the forms.

Figure 6 shows the standardized residuals (*y*-axis) plotted against the predicted values of hazard (*x*-axis).The plot is observed to be random without any fixed pattern or cycle. This is the plot obtained for the applied form 1 of Gompertz law of mortality. The other three plots also have the similar random plots. These aforesaid random plots of standardized residuals validate the Gompertz law of mortality as the good fit. We may also conclude that each form of Gompertz curve fits well to the data as compared to Cox PH model for which the similar plot is non-random supporting the fact. The same conclusion has been drawn by (Ravi *et al*., 2020). They have also recognized the robustness of the actuarial laws for modelling the survival time in insurance phenomenon.

Table 4: Results under Gompertz curve

Model Form	Combinations	Value of *a*	Value of *b*	Weights	*AIC*
Form 1 - With age Only	None	0.5594	0.0003	1	–12.1
Form 2 - With age, Area and Gender	Male lives in rural area with all ages	0.547	0.0002	0.2037	–10.5724
	Male lives in urban area with all ages	0.5673	0.0003	0.7158	
	Female lives in rural area with all ages	0.3737	0.028	0.0131	
	Female lives in urban area with all ages	0.4695	0.0087	0.0674	
Form 3 - With SA only	None	0.5744	0.0001	1	–12.17
Form 4 - With area and SA	Rural area with all SA	0.5478	0.0005	0.2168	–11.127
	Urban area with all SA	0.5747	0.0008	0.7832	

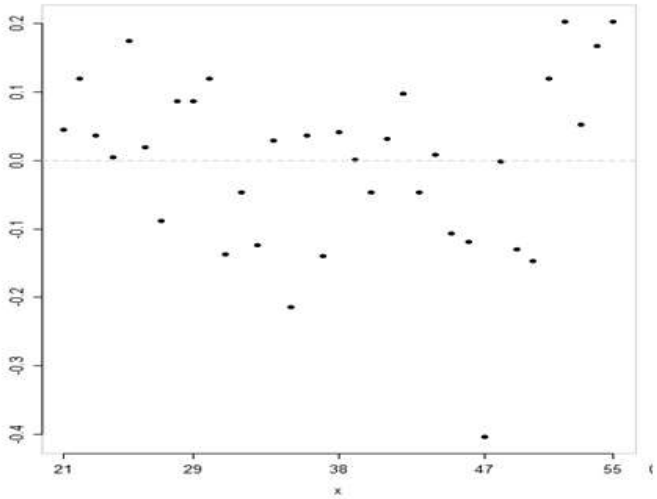


Figure 6: Residual plot under Gompertz law

Although the Gompertz law of mortality can be efficiently used for predicting lapsation probability but it is not suitable for classification purpose as it may require threshold value which in turn requires the complex algorithm to be employed. It predicts the probabilities of lapsation which may be categorized from low to high. To use this law for classification purpose, we are further required to find out a threshold value, such that if the predicted probabilities fall below it, then, the policies may be classified as lapsed and if the probabilities fall above it, then, the policies may be classified as in force. This makes the classification job more tedious. One more demerit of this law is that all the factors at role cannot be keyed in together to assess their effect on mortality (lapsation in our case). We must go on using various forms of the model one by one to assess their impact on the event of interest. So, we move ahead to apply other machine learning techniques of classification for predictive modelling.

4.3. Naïve Bayes classifier

We use the Kernel approach for fitting the Naïve Bayes classifier which is considered as more appropriate.

The a priori probabilities are evaluated as: 0.438 for in force policies and 0.562 for lapsed policies. We predicted few cases where certain combinations of values are given as input for factors at work like gender, SA, age, plan type *etc*. and model classified the policies as lapsed or in force. Then on comparing the predicted and observed values, the results are found to be consistent.

Table 5: Confusion matrix for training data

Status of policy (Lapsed: 1 and Inforce: 0) ↓ →	0	1
0	1130	2
1	0	1447

Table 5 above shows that out of 2579 policies in training data set, 1130 policies are correctly classified as in force and 1447 policies are correctly classified as lapsed whereas 2 policies are classified as lapsed while they were actually in force. The error rate for the model is calculated as 0.000775.

Further the confusion matrix for testing data set which contains 1084 policiesis exhibited below in Table 6.

Table 6: Confusion matrix for testing data

Status of policy (Lapsed: 1 and Inforce:0) ↓ →	0	1
0	462	3
1	0	619

Table 6 above shows out of 1084 policies; 462 policies are correctly classified as in force and 619 policies are correctly classified as lapsed which are consistent with their observed values and the remaining 3 policies are misclassified. Error rate of the model for testing data for the model is 0.002767.

In both the cases it is observed that misclassification probability is negligibly small. Therefore, the Naïve Bayesian model is a better fit as compared to the other two models discussed above.

4.4. Random Forest technique

Figure 7 below shows the error rates of misclassification for the model (y - axis) plotted against the chosen number of decision trees (x - axis). We observe that the decline in error rates become stagnant between 15 and 20 numbers of decision trees. If we choose decision tree beyond 20 then there will be no significant reduction in error rate but the number of parameters to be estimated will redundantly increase.

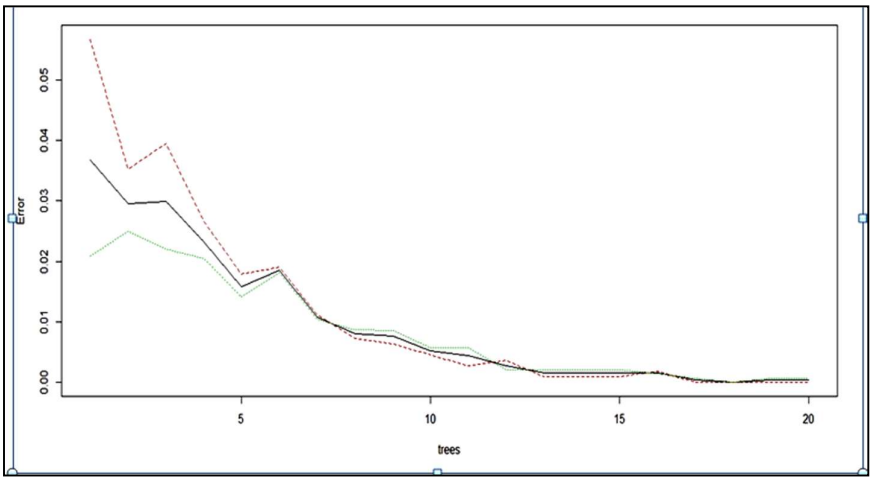


Figure 7: Error rate for fitted random forest

Following results are obtained by fitting the random forest with 20 decision trees:

(i) The Out of Bag (*OOB*) estimate of error rate is 0.04 *per cent.*

(ii) The confusion matrix (Table 7 below) for the fitted random forest on training data set.

Table 7: Confusion matrix on training data

Status of policy (Lapsed: 1 and In force: 0) → ↓	0	1
0	1111	1
1	0	1431

It is apparent that out of total 2543 policies in training data set, random forest correctly classified the 1111 policies as in force when they are actually in force and 1431 policies as lapsed when they are actually lapsed. The model misclassified only in one case where the policy is predicted as lapsed but it is actually an in-force policy.

(iii) The confusion matrix (Table 8 below) for the fitted random forest on testing data set is given below and the *OOB* error rate estimated as 0.067 *per cent*.

Table 8: Confusion matrix on testing data

Status of policy (Lapsed: 1 and In force: 0) ↓ →	0	1
0	480	2
1	0	638

The above confusion matrix in Table 8 gives correct classification for total 1118 policies out of 1120 policies in testing data set and misclassified only 2 policies. All these above results show that the model is a best fit to the data.

(iv) The proportions of the policies with the predicted status as lapsed (coded as 1) or in force (coded as 0), for both the training and testing data sets are shown below in Figure 8.

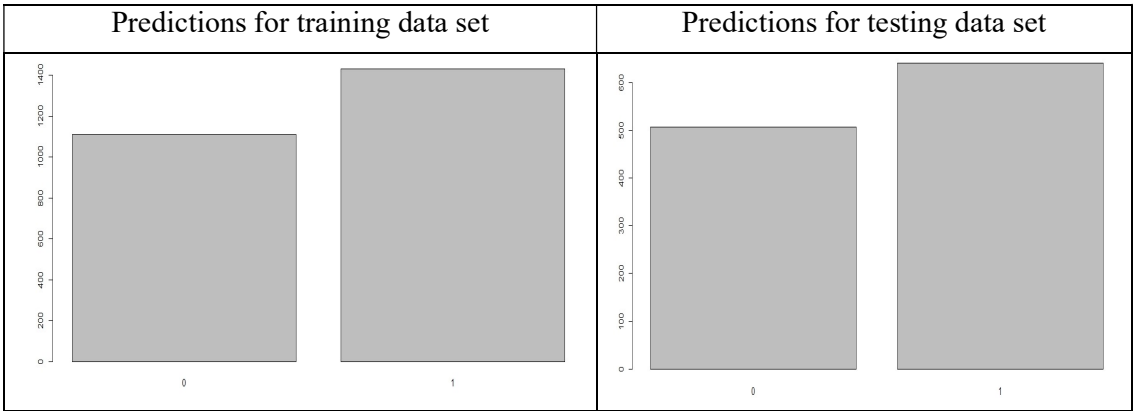


Figure 8: Prediction proportions for both training and testing data

The proportions for both the data sets are almost equal for lapsation as well as in force policies. For both training and testing data set the model predicted same proportion around 54 *per cent* for lapsed policies. The proportion of in force policies in both the data sets is approximately obtained as 46 *per cent*. Both data sets having almost same proportions of predictions for in force and lapsed policies specify that the model works well.

5. Conclusions

A comparison of all the four models for predictive modelling and classifying the policies as lapsed or in force is important before concluding and therefore, it is shown below in Table 9.

Table 9: Comparison of employed four models

Points for comparison	Cox PH model	Gompertz law	Naïve bayes classifier	Random forests
Does model fit well to the survival data of life insurance policies	No	Yes	Yes	Yes
Basis of assessment of a model as a good fit	Standardized Residuals plot and Concordance	Standardized residuals plot and AIC values	Confusion matrix and error rate of misclassification	Confusion matrix and error rate of misclassification
Is the model appropriate to classify the life insurance policies	No Not appropriate as it does not fit the data well	No It is tedious to use the law for classification which requires complex algorithm	Yes It is easy to use and interpret	Yes It is easy to use and interpret
Does model involve complex algorithm to classify	Yes	Yes	No	No

The conventional survival model (Cox PH model) and the actuarial law (Gompertz law) are not found to be suitable for the classification of policies as lapsed or in force and the reasons have been tabulated above in Table 9. Thus, we require some better predictive modelling techniques which can classify the policies with certain set of values of predictors/factors at work, as lapsed or in force. These techniques are also capable of taking into consideration all the factors affecting lapsation together. We see that Naïve Bayesian technique and Random forest technique both performed very well. Both the techniques have the almost same error rate of prediction. Under both the models, confusion matrix is depicting the same proportions of correctly classified policies. We conclude that both the techniques performed equivalently well on insurance data. Classification is important in predictive modelling to assess whether the policy holder with certain features like income band, SA opted, policy term opted, age etc. will be lapsing the policy or will continue the policy with certain probabilities attached to the outcome. It is found that nearly 54 per cent policies are classified as lapsed in the class of policyholders who are businessmen up to the age 45 and with income up to 5 lakhs, insured with low SA up to 5 lakhs and policy term between 25 to 30 years and they purchased savings/health/ULIP insurance products. This classification helps insurer to identify the segments in the society with the different combinations of various features where more prospects can be canvassed which will be retained for long time or up to the maturity. Insurers may also identify the segments where the prospects or policyholders are not being canvassed or retained for long time to analyze for probable reasons responsible for problem of low retention. The issues may be related to after sales service, low awareness of insurance in the segment or products which do not appropriately serve the needs in the segment. Some issues are also beyond the control of an insurer. But some of the issues like mentioned above may be improved by working upon them.

References

Adebiyi, S. O., Oyatoye, E.O. and Amole B. B. (2016). Improved customer churn and retention decision management using operations research approach. *Emerging Markets Journal*, **6**(**2**). DOI 10.5195/emaj.2016.101, http://emaj.pitt.edu

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P. E. and O'Connor, P. J. (2014). Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data.http://arxiv.org/abs/1404.2189v1.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5–32. Kluwer Academic Publishers. Manufactured in The Netherlands.

Cheng, J. and Greiner, R. (2013). *Comparing Bayesian Network Classifier*. Department of Computing Science University of Alberta Edmonton, Email: {jcheng, greiner} @cs.ualberta.ca.

Gustafsson, E. (2009). Customer duration in non-life insurance industry. *Mathematical Statistics Stockholm University*, Examensarbete, **3**.http://www.math.su.se/matstat.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

Huigevoort, C. (2015). *Customer Churn Prediction for an Insurance Company*. TUE School of Industrial Engineering. Series Master Theses, Operations Management and Logistics.

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z. and Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail. *Informatics*, **5**, 2. doi:10.3390/informatics5010002.

Lariviere, B. and Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, **29**, 472-484. 10.1016/j.eswa.2005.04.043.

Lee, E. T. and Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*. Third Edition. Wiley.

Lenart, A. (2012). *The Gompertz Distribution and Maximum Likelihood Estimation of its Parameters - a Revision*". Max Planck Institute for Demographic Research, Working Paper 2012-008.

Onisko, A., Druzdzel, M. J. and Wasyluk, H. (2001). Learning Bayesian network parameters from small data sets: application of noisy OR gates. *International Journal of Approximate* Reasoning, **27**,165-182.

Rao, M. B. and Rao, C. R. (2014). *Handbook of Statistics*. Elsevier (ISBN: 9780444634313).

Ravi, V., Saini, R., Varshney, M. K. and Grover, G. (2020). Modelling of survival time of life insurance policies in India: A Comparative study. *International Journal of System Assurance Engineering and Management.* DOI: 10.1007/s13198-020-01026-2.

Richmond, P. and Roehner, B. M. (2016). Predictive implications of Gompertz law. *Physics A*: *Statistical Mechanics and its Applications*. http://arxiv.org/abs/1509.07271v1.

Rudolph, C. (2002). *Application of Survival Analysis Methods to Long Term Care Insurance.* Sonderforschungsbereich386, Paper 268. http://epub.ub.uni-muenchen.de/.

Therneau, T. and Atkinson, E. (2020). *Concordance*. www.CRAN.r-project.org.

Tirenni, G., Kaiserand C. and Herrmann, A. (2007). Applying decision trees for value-based customer relations management: Predicting airline customers "future values". *Database Marketing and Customer Strategy* Management, **14,** 130–142.

Zhang, R., Li, W., Mo and, T. and Tan, W. (2017). Deep and shallow model for insurance churn prediction service. *IEEE Computer Society*, 2474-2473/17.

DOI 10.1109/SCC.2017.51