Mining Life Insurance Data for Customer Attrition Analysis

T. L. Oshini Goonetilleke
Informatics Institute of Technology/Department of Computing, Colombo, Sri Lanka
Email: oshini.g@iit.ac.lk

H. A. Caldera
University of Colombo/School of Computing, Colombo, Sri Lanka
Email: hac@ucsc.cmb.ac.lk

***Abstract*— Customer attrition is an increasingly pressing issue faced by many insurance providers today. Retaining customers who purchase life insurance policies is an even bigger challenge since the policy duration spans for more than twenty years. Companies are eager to reduce these attrition rates in the customer-base by analyzing operational data. Data mining techniques play an important role in facilitating these retention efforts. The objective of this study is to analyse customer attrition by classifying all policy holders who are likely to terminate their policies. These customers who are at high risk of attrition can then be targeted for promotions to reduce the rate of attrition. Data mining techniques such as Decision trees and Neural Networks are employed in this study. Models generated are evaluated using ROC curves and AUC values. Our research also adopts cost sensitive learning strategies to address issues such as imbalanced class labels and unequal misclassification costs.**

***Index Terms*—classification, cost sensitive learning, customer attrition, data mining, life insurance.**

I. Introduction

The Insurance Industry in Sri Lanka is growing rapidly and it has become more and more important to keep pace with the growth of the industry through technological advancements and innovative ideas to market the organization to the masses. Portfolio of products offered by insurance providers has diversified, over the years, attracting more customers than ever. Accumulation of operational data inevitably follows from this growth in industry. There exists an increasing need to convert their data into a corporate asset in order to stay ahead and gain a competitive advantage.

Data mining technologies are emerging in the industry which attempts to extract knowledge from large collections of data. The exploration of data is greatly facilitated by these data mining efforts. Many commercial data mining tools like SAS Enterprise Miner, MineSet by Silicon Graphics and Intelligent Miner by IBM are available to facilitate the exploration process. However in the Sri Lankan context, insurance companies are reluctant to use off-the-shelf software. This is because they are very costly and companies are not convinced of what data mining can do for them.

This opens up many opportunities and challenges for the data mining researcher to convince these companies of the commercial viability of data mining efforts. This work is motivated by a real-world problem of customer attrition faced by life insurance providers.

This paper aims to demonstrate how data mining can be applied, highlighting the issues and suggested solutions at each phase of the methodology.

A. Customer Attrition in Life Insurances

The aim of life insurance is to pay an agreed sum, known as a *premium*, on a specific date to the insurer so as to provide financial benefit to the dependents after the death of the policy holder. Interim uncertainties like accident to a person, critical illnesses, surgeries and medical charges could also be included in this kind of contract.

Like in many other industries, customers have the option to choose from a wide range of life insurance providers giving attractive benefits. A main problem faced by the life insurance sector is customer attrition. Policy holders discontinue payments of the premium which results in a policy termination or in insurance terminology - a lapse. It is far more expensive to acquire a new customer than to retain an existing one. This is especially true in life insurances, since cost of losing a customer who has bought a policy that spans for over ten years, is very high.

After a period of time, customers discontinue payment of premiums and eventually terminate their policies. It is not possible to detect the lapses early unless customer explicitly contacts the company with some concerns in suitability of the policy, which is rare. It is important for any insurance provider to conduct analysis on customer attrition so they can take proactive measures to retain them.

B. Problem Definition

The goal of this analysis is to address the problem of customer attrition in the life insurance domain by attaining the following objectives:

Manuscript received January 23, 2013; revised March 30, 2013.

doi: 10.12720/jiii.1.1.52-58

- Identify factors that contribute most to the customer's decision of a policy termination.
- Using these factors, perform data mining to understand customer retention patterns by classifying policy holders who are likely to continue or terminate their policies.
- From the classification, identify a group of customers who have a high probability to attrite.

These customers who are at high risk of attrition can then be targeted for promotions to reduce the rate of attrition.

II. Data Mining Methodology

A range of researches have been conducted which attempts to systematically analyze the reasons and predict the likelihood of customer attrition. The analysis poses different types of challengers in different industries during various stages. However the approach taken in essence is similar. Given below are standard high-level steps in any data mining effort:

1. Problem definition
2. Data Preparation
3. Attribute selection
4. Knowledge discovery using data models
5. Take action based on results

First and the last step are more of business issues. Step two, data preparation requires extensive subject knowledge of an expert in the domain. Tasks under step two include defining the class label for classification task, extraction of the data for a significant period, initial descriptive statistics to visually explore the data, understanding data dependencies and combining attributes if necessary. In third step, attribute selection identifies a good subset of attributes that can classify terminations well. This analysis will unveil much directly actionable information that might be useful for the insurance provider. Knowledge discovery using models such as neural networks and decision trees, attempts to identify a good set of customers who are of high probability to attrite. Finally the data miner can recommend the insurance providers to use the subset to conduct various targeted campaigns in the hope of reducing their attrition.

Researchers engage in various attempts to provide a successful data mining solution, concentrating more on some steps than others. For the experiments in stage 3 and 4, Weka package [1] by University of Waikato version 3.7.4 was used, mostly with their default configurations. Weka package is manipulated to a great extent facilitating the process of model building so the focus can be on the problem.

III. Data Preparation

Appropriate data need to be collected and pre-processed to meet the objectives of the problem at hand. Initial data is extracted from an operational database with the help of a domain expert who can provide valuable insight into the type of attributes required for analysis of customer attrition.

A life insurance policy can be in any one of the stages shown in Table I depicting the status of one's policy.

TABLE I. Stages of a Life policy

OPEN	Policy is open when the customer is successfully making payment of premium on an agreed mode (monthly, quarterly, half yearly or yearly)
LAPSE	If premium payments have discontinued before the policy has acquired a surrender value[1], the policy is said to be lapsed.
PDUP	If premiums are not paid after the policy has acquired a surrender value, it becomes a paid-up policy.
MAT	If all premiums are paid for the entire period as specified in contract, policy is said to be matured.

Lapsed and paid-up are the two main behaviors of customer attrition. Out of the two behaviors of customer attrition, insurance providers are mostly interested in customers who 'lapse' than 'paid-up' policies. The reason, according to experts in the domain, is that the cost an insurer has to incur for the first three years of a policy is significantly high. Due to this reason, primary focus of attrition analysis is to take measures to prevent policies becoming lapsed.

Class label is identified to be one of the statuses shown above; OPEN, LAPSE, PDUP and MAT.

Dataset need to be extracted from a significant period of time to build a meaningful model. Unlike in a motor policy, contract in a life insurance is for a period of 18-20 years, on average. This period is known as the policy *term*. A subset of records was selected where the policy start date is within the year 2002-2003 for initial analysis.

The class distribution of the extracted dataset from year 2002-2003 is 19% of OPEN, 68% of LAPSE, 12% PDUP and 1% of MAT policies as depicted in Fig. 1. The class distribution is similar in consequent years. It is assumed that the 2002-2003 dataset is representative of all policy holders.

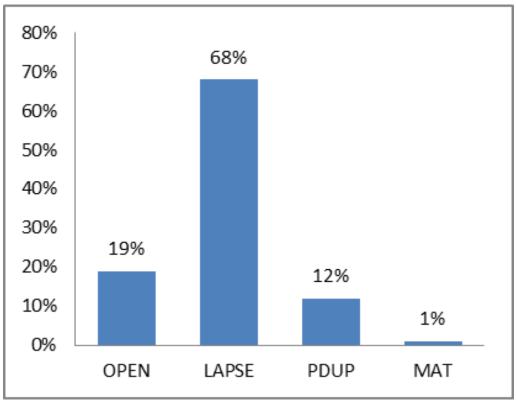

Figure 1. Distribution of classes

It has to be noted here that the class distribution is skewed. This is common for many industries dealing with attrition analysis. As a consequence, a learning model will be eager to classify every customer for the frequent class and still achieve high accuracy. Researchers have

[1] When a life insurance policy is maintained for a minimum period of 3 years continuously it acquires a cash value, known as a ‘surrender value’. This is the minimum cash entitlement available when the policy is terminated before the maturity date.

dealt with this in many ways. In a study conducted by Hu [2] in the banking industry, there were a high percentage of non-attriters (97.2%) vs. a very low attriter rate (2.2%). As an intuitive solution, data from the rare class (attriters) was included to make a more balanced representation of the dataset.

Another approach to solve this problem was to change the evaluation criteria. When the dataset is unbalanced, predictive accuracy alone will not be a good measure to evaluate the models. Alternative evaluation measures were taken as mentioned in Smith *et al* [3], where there was only 7.1% of the rare class. They have explored the varying effect of the decision threshold to find a model that suits best for the purpose of the study. This unbalanced dataset need to be taken into consideration when applying models and evaluation.

Next, a set of attributes were selected that would represent a life insurance policy. They include demographic information such as age, occupation, gender etc.; policy details such as agent, term of policy, sum assured, premium etc. another attribute was constructed to represent the number of other policies a policy holder may have with the company.

Preliminary cleansing was done removing static attributes (values that are mostly unique), null values (having more than 95% of null) and constants(values that are mostly the same within a column).

At the next step, initial descriptive statistics attempts to understand the data. The purpose is to ascertain if any significant patterns can be identified visually. Different types of visual techniques had to be utilized for numeric and nominal type of attributes.

For each nominal attribute, 100% stacked percentage column chart is drawn (based on a pivot table) as shown in Fig. 2 which shows distribution of class label, status, for each value of payment mode (annual (A), half-yearly (H), monthly(M) and quarterly(Q)).

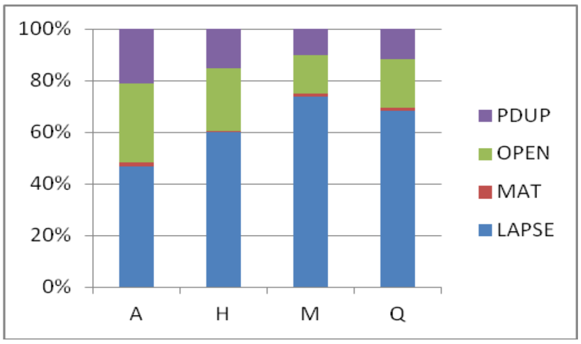

Figure 2. Stacked column chart for mode and status

Domain experts had intuitively thought when in monthly mode, since the premium is smaller, there would be less tendency for customers to lapse. However, observing the data in Fig. 2, lapsed policies are highest in the monthly (M) payment mode. This was not in total agreement with the domain experts. We have to be cautious when analyzing and interpreting relationships, as we cannot quantify them.

Similar to this, other nominal attributes (product type, gender etc.) were plotted against the status and reasoning was given based on the visual behavior with subject expertise. This exploration process did expose some valuable insight into the dataset in the initial study which was later used to explain patterns found. An association with the class label could be visualized for some nominal attributes. Not having visual patterns for other nominal attributes does not give enough reason to remove them from the analysis.

Nominal attributes such as agent, branch and occupation need to be combined due to high number of unique values in each of them. A comparison of averages in the class labels was conducted for the numeric attributes.

In the next step, a variety of mechanisms are used to select a good subset of attributes for model building. It is likely that all these attributes may have more impact when combined rather than individually having an effect on status. This analysis is further improved by running a methodical approach to attribute selection.

IV. Attribute Selection

After initial field reduction, we ended up with 21 attributes. It is a crucial task to pick the best set of attributes that has most predictive ability to classify a policy as lapsed. It is almost impossible and often not useful to build models using all of these attributes. Most researchers undertake various attribute selection techniques to find a suitable subset. With further analysis, insurance provider can act upon these selected attributes alone to take measures to reduce customer attrition.

One of the methods used is Correlation-based Feature Selection (CFS) subset evaluation technique[4] which picks the best attribute(s), but this method does not show the ranking of remaining attributes. CFS technique selected ‘product type’ and ‘basic sum assured’ as the most important. Other measures such as Gain Ratio and Information gain (with respect to the class label) was used to rank the attributes in the order of importance. Underlying idea of these techniques is that attributes selected in the standard decision tree that gives high accuracy are considered as relevant and meaningful indicators.Out of the 8 most important attributes that were selected for each method, 5 of them were common in the two approaches. Various combinations of these selected attributes were used for the experiments in the next step.

V. Knowledge Discovery

Goal of our study is to classify policy holders who are likely to continue or terminate their policies. Common techniques to predict customer attrition using classifications are Decision Tree, Neural Network and Logistic Regression. It is important to select the model(s) that best suits the problem.

We found three main criteria that are in common practice when deciding on a model for a given problem. First criterion for selection is whether the rules generated from the model can be well understood by the end users. Considering this, decision trees and regression are clear winners compared to the neural network which is a black box in its nature. Second criterion for selection is better performance of a model across many domains. According to an extensive study conducted by Cerny[5] the predictive ability of a neural network is promising. In most cases they seem to outperform techniques to which it is compared. Final and the most appropriate seem to be building models using multiple classifiers, either combining them or selecting a single one which performs best on given data. But this can only be done with availability of tools and depending on the extent of the analysis. The models must be compared using suitable evaluation criteria.

In the ensemble classifier, instead of selecting one model, several data mining algorithms such as Na ïve Bayesian, decision tree, and neural network are performed and combined[2]. In this method, output of each classifier is combined together hoping that the combinations of answers of multiple classifiers result in better accuracy. As mentioned in another research by Hu[6], the combination can be done using various methods. The study claims that the use of an ensemble of classifiers have yielded better results.

Researchers that perform classification [3],[2],[7], mostly try to “rank” the customers instead of trying to classify a new customer as either an attrite or non-attrite. In order to perform ranking, it is required to build learning algorithms that predict with a confidence measurement. This confidence factor is also known as a probability estimation factor or certainty factor. Ranking is useful so that the potential attriters can be ordered and promotions can be varied depending on the certainty of termination.

A. *Evaluation of Classifiers*

Literature has proposed many different ways to evaluate a classification model. Prediction accuracy is the simplest measure that can be calculated to evaluate models.

$$\text{prediction accuracy} = \frac{\text{correctly classified instances}}{\text{total number of instances}} * 100\% \quad (1)$$

As pointed out in a research done by Ling and Sheng[8], prediction accuracy alone is not a good measure to evaluate the performance of classifiers. One reason is that the classification errors are treated equally, but in attrition analysis, false negatives (recognising non-terminators as terminators) and false positives (recognising terminators as non-terminators) have different impact and should be dealt with differently. Prediction accuracy will also give an erroneous view for skewed data sets as well. Alternative measures are discussed next.

a) ROC and AUC for two-class problems

A receiver operating characteristics (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance. A ROC Curve shows the trade-off between the true positive rate[2] (TP rate) and the false positive rate[3] (FP rate) for a given model of a two class problem. A simple ROC graph is shown in Fig. 3.

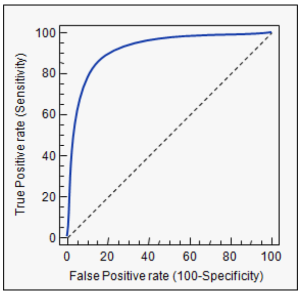

Figure 3. The ROC graph

Output of a discrete classifier (classifier that outputs only the class label) will yield only a single confusion matrix with values for TP rate and FP rate, which in turn corresponds to a single point in the ROC space. Fawcett[9] details a more practical method that would generate the ROC curve. The curve is created by thresholding a test set. Each point in the ROC space corresponds to a particular decision threshold.

In addition to being a generally useful performance graphing method, ROC analysis is used in this study due to its extended properties. ROC curves have an attractive property[9]: they are insensitive to change in class distribution. If the proportion of positive to negative instances changes in a test set, the ROC curves will not change. They work well for domains with skewed class distributions and unequal misclassification costs[9]. Our study attempts to address both issues. Therefore ROC analysis becomes the ideal candidate. These characteristics have become increasingly important as research continues into the areas of cost-sensitive learning and learning in the presence of unbalanced classes.

To compare the performance of different classifiers, visually, we may have to plot the ROC curves for each of them on the same graph. It would be better to reduce the ROC performance to a single scalar value representing its performance. The solution is to calculate the area under the ROC curve, abbreviated AUC. It is a value between 0 and 1. A random classifier has AUC value of 0.5, while a perfect classifier will have 1.0. No realistic classier should have an AUC less than 0.5. We refer the interested reader for a detailed discussion by Fawcett[9] and Vuk and Curk[10].

b) ROC and AUC for Multi-Class problems

All of the above are discussed for a problem with two class labels. Our study attempts to perform classification with 4 classes (Open, Lapse, Pdup and Mat). Lane has proposed [11] an intuitive method for handling n number of classes; to produce n different ROC graphs, one for each class. As mentioned in Fawcett[12], there are two main approaches in literature to handle the issue of generating AUC values for a multi-class problem. For simplicity of calculation, we go with Provost and

[2] TP rate is the proportion of positive cases that were correctly identified by the classifier

[3] FP rate is the proportion of negatives cases, incorrectly classified as positive

Domingos' method cited in [9]. They calculated AUCs for multi-class problems by generating each class reference ROC curve in turn, measuring the area under the curve, and then summing the AUCs weighted by the reference class's prevalence in the data.

$$AUC_{total} = \sum_{ci \in C} AUC(C_i).p(C_i) \qquad (2)$$

where $AUC\ (C_i)$ is the area under the class reference ROC curve for class (C_i), as in the equation given above. These measures are used to evaluate the performance of our classifiers.

B. Experiments

A series of experiments were conducted to find out the most suitable classifier. For purposes of evaluation, evaluation metrics were calculated for each experiment. Classifiers are compared using values in a confusion matrix, accuracy measures based on the confusion matrix, and AUC. Prediction accuracy is also listed in some cases, to have a simple idea about the percentage of correctly classified instances. Five datasets were created with various combinations of attributes. Attributes selected from;

Set 1. CFS subset evaluation,
Set 2. Gain Ratio measure,
Set 3. Information Gain measure,
Set 4. common attributes of set 2 and 3 and
Set 5. both Gain Ratio and Information Gain attributes.

A stratified 10-fold cross validation technique was used to select the training and testing samples due to its relatively low bias and variance in selection of data.

Primary classifiers we have used are Decision Trees and Neural Networks. Tests were conducted using a decision tree algorithm J48: an implementation of Weka. It generates a pruned or un-pruned C4.5 decision tree[13] implemented by Quinlan. Decision Trees are selected since it requires less training time and it is easier to interpret the rules generated.

Another experiment was conducted by discretizing all the numeric attributes, to nominal by equi-percentile binning method. Finally, a Neural Network was selected with a standard Multilayer Perceptron network structure using Backpropergation. Comparison of weighted AUC values generated from each classification is depicted in Fig. 4 below.

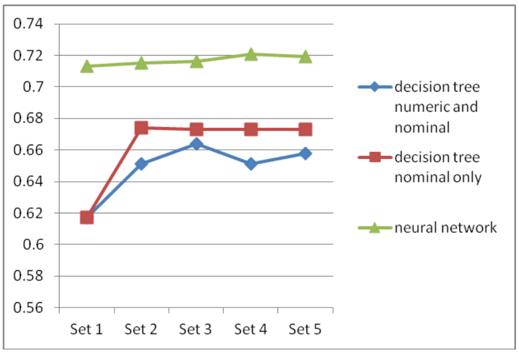

Figure 4. Comparison of AUC values for classifiers

The x-axis shows the attribute sets and the y-axis shows the weighted AUC values. It can be seen that, for decision trees, AUC values have been improved for all sets when all attributes were converted to nominal. However, there's a significant performance improvement when a neural network is used to classify the policies, confirming the superiority of the neural network classifiers.

However, when observing the confusion matrix for the experiment, it was not deemed successful. Confusion matrix for Neural Network is shown in Fig. 5 for discussion. R. A and C. A referes to Row and Column Accuracies respectively. Prediction accuracy (from eq(1)) is 73.61%.

	Classified as				
	OPEN	LAPSE	PDUP	MAT	**R.A**
OPEN	1852	3473	104	14	34.0
LAPSE	521	19197	107	0	96.8
PDUP	377	3061	80	0	2.3
MAT	11	16	1	311	91.7
C. A	67.1	74.6	27.4	95.7	

Figure 5. Confusion matrix for neural network classification

It can be observed that MAT policies are classified well. The problem is that many instances of OPEN and PDUP policies are also classified into the LAPSE class. PDUP is another type of attrition. Therefore the cost of classifying a PDUP as a lapse is less than the cost of classifying an OPEN policy as a lapse. Cost sensitive learning attempts to introduce a cost for the misclassifications to improve this further by reducing the effect of class imbalance.

VI. Cost Sensitive Learning

It was noted when selecting the dataset that the class distribution is highly imbalanced. A skewed dataset is a common problem in attrition analysis. The data distribution is skewed towards one class label, in our case, towards a lapse.

One intuitive method to overcome the problem is to add more records to balance the data set. It is also possible to utilize different re-sampling techniques as summarized in the works of Chen[14]. However, it might not always be possible to do this. If the data set is not balanced, classifier tends to always predict a record as the class having a higher frequency.

As discussed in Elkan [15] and simplified by Ling[8], another reason for employing cost sensitive learning strategies is due to unequal costs assigned to each misclassification. Consider the two class problem, where “having a cancer” is positive. The cost of a person classified as positive when he’s actually negative is potentially less, than a person classified as negative when he’s actually positive (of having cancer). This phenomenon is also true for attrition analysis. By default, classifiers treat all misclassifications as having equal cost. In a realistic environment, this might not be the case.

Reliability of the classifier predictions can be improved by introducing a concept of ‘cost’ to the misclassifications. The goal is to adjust the tradeoff between such misclassifications.

A. Cost Matrix

In the problem of attrition prediction, there can be higher cost (or benefit) assigned to an actual attriter being classified as a non-attriter than the reverse.

Consider the cost matrix in Fig. 6 where costs are assigned to each classification. There is no cost assigned to the diagonal values, meaning there's no cost for correctly classified instances. Consider the value 10: cost of classifying an OPEN as a LAPSE is 10. This is twice as costly as classifying a LAPSE as OPEN. We need to give a higher value for the cost of misclassifying a rare class (OPEN) than misclassifying the majority class (LAPSE).

OPEN	LAPSE	PDUP	MAT	
0	**10**	2	2	OPEN
5	0	1	2	LAPSE
3	6	0	1	PDUP
2	3	0	0	MAT

Figure 6. Sample cost matrix

There's no easy mechanism to specify the optimal cost values. It must be noted that best values for a cost matrix must be determined empirically. Experiments were conducted in an effort to decide on these values. We try to achieve two goals when defining values for the matrix. Increase the row accuracy for OPEN and PDUP policies. If we achieve this, using this model, we can find a good subset of policy holders with high probability to LAPSE.

B. Experiment Results

We conducted a series of experiments, hoping to reduce the misclassifications for OPEN and PDUP policies. An interim result is shown in Fig. 7 for the cost matrix used in Fig. 6. By assigning a higher cost(10) for misclassifying OPEN as a LAPSE, we managed to improve the row accuracy of OPEN policies from 34% to 43%. There was not much of a change for PDUP policies.

	OPEN	LAPSE	PDUP	MAT	**R.A**
OPEN	2365	3068	3	7	43.5
LAPSE	1562	18255	8	0	92.1
PDUP	689	2820	9	0	0.3
MAT	43	25	0	271	79.9
C.A	50.8	75.5	45.0	97.5	

Figure 7. Improved confusion matrix

With further experiments, we managed to increase the row accuracy for OPEN policies to 46%. The further we increase the cost values; the increase of the row accuracy for OPEN policies was achieved at the expense of overall accuracy. So it was required to strike a balance between the two.

If it is required to conduct some direct marketing campaigns, it is required to have a model with higher row accuracy making sure to capture as many actual terminations as possible. So we finally arrived at the model where it captures 90% of the lapses correctly as shown in Fig. 8.

	OPEN	LAPSE	PDUP	MAT	**R.A**
OPEN	2478	2959	1	5	**45.5**
LAPSE	1924	17901	0	0	**90.3**
PDUP	793	2724	1	0	0.0
MAT	68	32	0	239	70.5
C. A	47.1	75.8	50.0	98.0	

Figure 8. Final confusion matrix

It must be noted that 45.5% is the best we could achieve for row accuracy for Open policies. Using this model we can conduct targeting campaigns aiming to reduce the rate of customer attrition. Row accuracy is high, making sure to capture as many actual terminations as possible. Using this subset of customers with a high attrition rate insurance provider can conduct direct marketing campaigns contacting the most likely attriters, where it is expensive to contact an entire list of customers.

VII. EVALUATION

This study sets the scope by trying to achieve three objectives. First is to find factors that contribute most to the customer's decision of a policy termination. Second is to classify policy holders as likely to terminate or not. Third and final is to identify a subset of policy holders who have high probability to attrite. Using this subset the goal is to take measures to facilitate customer retention. Such measures can be promotions at various target groups and adjustments of premiums.

To achieve the objectives mentioned above, researcher was faced with practical challenges at different stages of the project. There are many documented ways of addressing the problem of attrition analysis faced by many business domains. The focus of the project was to find a solution that can be implemented by the insurance provider eventually. Input from domain experts was required for almost all the stages of the project.

Data set considerations-Data was extracted having policies that started in year 2002-2003. Determining this period is not straightforward as the predicted label of status of the policy is dependent on the period. For example, policies that started in 2009 do not have room for LAPSES. It can be noted as an enhancement, data can be extracted as policies that are due for maturity instead of the policy start date. Whichever way, we are bound to end up with a data set where the class distribution is imbalanced. So this problem needed to be addressed in this paper.

Attribute selection-At the end of initial and visual analysis of data, we still end up with a large attribute set that cannot be directly used for data mining. There are many statistical techniques that can verify the association of each attribute with the class label. Univariate tests such as chi-squared tests can be done to verify certain hypothesis developed from the data. This effort did not yield satisfactory results in our study, as the chi-squared value for each attribute was almost always significant. This observation was in agreement with some literature that stated the difficulty in applying univariate tests for very large populations of data.

Evaluation of classifiers-In all the models built it was noted that the classifier is not capable of classifying PDUP policies well. They are mostly misclassified to be LAPSE policies. Classifier fails to distinguish the two classes properly. However, PDUP is also a type of method that customers terminate policies. Due to this reason, the misclassification can be somewhat justified. Similar problem was encountered with OPEN policies being classified as a LAPSE. This is not good as they differ in meaning. Due to this reason, better row accuracy in OPEN is also a factor to consider when comparing classifiers. It was observed that the MAT policies are classified with reasonable column and row accuracy.

It must be emphasized that the goal of this classification is certainly not to predict which class a new customer might belong to with some certainly measure. The goal is to find out a subset of customers where the percentage of attrition is high. As discussed in the study by Hu, if the model is good enough [2], it should find subset that has a high concentration of attriters or lapsed policies.

VIII. Recommendations & Future Work

The recommendations are given for various aspects of the problem of attrition analysis in the life insurances domain. There are various methods that can be suggested as classifier improvement strategies. Further investigation can be carried out to determine a set of optimal values for the cost matrix. We have approached the problem as a classification task. It is also possible to explore the dataset to find any strong associations among the data. An association analysis that satisfies a minimum support and a threshold will be useful (with the class fixed) for the insurance provider to act on these rules to promote to customer. This in fact, could've produced some meaningful results, considering the fact that exploratory analysis uncovered some patterns within the data.

Conducting a field test with the insurance provider will evaluate the validity of the final model being developed. To conduct a field test first we need to select the policies that are likely to lapse along with its confidence facto and they should be ranked according to decreasing confidence factor. Then customers in the top 10 % (or 20%) in the list can be selected, let's say 2000 customers. Split the 2000 customers randomly into 2 groups: group 1 to have no promotion effort and group 2 to make promotions. On the due date, retention rates should be observed in both groups. If the percentage of customers who actually lapsed in group 2 is less than group 1, then we have built a model that indeed facilitated the retention effort of the company. We can also extend this, effort to target different types of promotions. For example, we can promote to the top 10 to 20 (in a list having a reduced confidence factor of a lapse) percentile of customers for marketing campaigns that are less costly.

References

[1] M. Hall, E. Frank, G. Holmes, B. Pfahringer, *et al.*, *WEKA Data Mining Software: An Update; SIGKDD Explorations*, vol. 11, no. 1, 2009

[2] X. Hu, “A data mining approach for retailing bank customer attrition analysis,” *Applied Intelligence*, vol. 22, no. 1, pp. 47-60, 2005.

[3] K. A. Smith, 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*, vol. 51, pp. 532-541, 2000.

[4] M. Hall, “Correlation-based Feature Selection for Machine Learning,” Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.

[5] P. A. Cerny, “Data mining and neural networks from a commercial perspective,” *Department of Mathematical Sciences*, University of Technology Sydney, Australia, 2001.

[6] X. Hu, “Using rough sets theory and database operations to construct a good ensemble of classifiers for data mining application,” in *Proc. IEEE International Conf. on Data Mining*, 2001, pp. 233–240.

[7] L. Yang and C. Chiu, “Knowledge Discovery on Customer Churn Prediction,” in *Proc. 10th SWEAS International Conference on Applied Mathematics*, Texas, USA, 2006, pp. 523-528.

[8] C. X. Ling and V. S. Sheng, “Cost-sensitive learning and the class imbalance problem,” in *Appear in Encyclopedia of Machine Learning*, C. Sammut Ed, Springer, 2008.

[9] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letter*, Elselvier, vol. 27, pp. 861–74, 2006.

[10] M. Vuk and T. Curk, “ROC curve, lift chart and calibration plot,” *Metodolo˘ski zvezki*, vol. 3, no. 1, pp. 89-108, 2006.

[11] T. Lane, “Extensions of ROC analysis to multi-class domains,” in *Workshop on Cost-Sensitive Learning*, T. Dietterich, D. Margineantu, F. Provost, and P. Turney, Eds, 2000.

[12] T. Fawcett, “ROC graphs: Notes and practical considerations for data mining researchers,” *HP Laboratories*, Kluwer Academic Publishers, 2004, vol. 12, no. 56.

[13] R. Quinlan, “C4.5: Programs for machine learning,” Morgan Kaufmann Publishers: San Mateo, CA, 1993

[14] Y. Chen, *Learning Classifiers from Imbalanced Only Positive and Unlabeled Data Sets,* Project Report for UC San Diego Data Mining Contest, Department of Computer Science, Iowa State University, 2008.

[15] C. Elkan, “The foundations of cost-sensitive learning,” in *Proc. Seventeenth International Joint Conference of Artificial Intelligence*, Seattle, Washington. 2001, pp. 973-978.

Oshini Goonetilleke received her Masters degree in computer science from University of Colombo, Sri Lanka in 2011. She is currently working as a Lecturer at Informatics Institute of Technology, Colombo. She is interested in pursuing research on the dynamic arena of Data Mining and its applications to Customer Relationship Management in various industries.

Oshini Goonetilleke received her Masters degree in computer science from University of Colombo, Sri Lanka in 2011. She is currently working as a Lecturer at Informatics Institute of Technology, Colombo. She is interested in pursuing research on the dynamic arena of Data Mining and its applications to Customer Relationship Management in various industries.