

# Mining Operational Databases To Predict Short-Term Defection Among Insured Households

Noe Tuason and Rajesh Parekh

Allstate Research and Planning Center  
Menlo Park, CA 94025  
ntuas@allstate.com, rpare@allstate.com

**Abstract.** Customer retention is a key problem in the insurance industry. As new customers are generally not profitable for the first few years, minimizing defection is critical. The objective of this study is to mine the company's operational databases to predict the insured households that will most likely defect within the next 12 months. The operational databases available for mining consisted of all active policies as of January 1994 and new policies written thereafter in a particular business region. Building the analysis dataset presented several challenges. Policy level data had to be aggregated into household level information and matched with demographics from other databases. Customers who moved to a different address had to be tracked. We constructed snapshot files of active customers for each of the years 1994-1998. Each snapshot file contained information on about 600,000 households and was used to build models using logistic regression (mainly) and decision trees. Each year's model was used to predict defection for the following year. Results showed that the yearly models had little variations in terms of model fit, gains, relative importance of predictors, and other measures. Further, they were uniformly accurate in identifying short-term defection. This means that, barring major changes in the marketplace, a model based on 1998 data should reliably predict the likely defectors in 1999.

## 1 Introduction

*Customer retention* is a key problem in the insurance industry. In the United States insurance rates from several hundred insurers are highly competitive. The wide spread use of direct mail, telemarketing, and the Internet has enabled even regional insurers to have a nationwide presence. The availability of a wide variety of choices and the convenience of switching prompt an increasingly larger number of customers to defect. *Defection* occurs when the customer switches to another company for insurance coverage. The cost of prospecting new customers is very high. Besides, new customers are typically not profitable for the first few years. Minimizing defection among existing customers is thus critical. Some of the key drivers of defection include increased premium, unsatisfactory claim handling, and significant life events like marriage, and a beginning teenage driver in the family. Presently, these drivers of defection are only *rule-of-thumb* information. No systematic models are in place for predicting short-term defection among insured households. The objective of our study is to mine operational databases and develop models to predict short-term defection (i.e., defection in the following year).

The modeling project presented the following challenges:

- *Householding policy level information.*

Our goal was to predict defection on a *household* level. However, the data available in the operational databases is on an individual policy level. Collecting information on all policies that belong to a particular household was a difficult task.

- *Appending suitable demographic information.*

Apart from price and dissatisfaction with customer service, customer defection is often triggered by important *life-stage* events such as marriage, birth of a child, and purchase of a new home or automobile. The demographic information for capturing these important life-stage events needs to satisfy two key criteria: *currency* (the information must be as current as possible) and *accuracy* (the information must accurately capture the customers' demographic profiles).

- *Constructing appropriate analysis files.*

Appropriate analysis file that would enable us to design models to predict short-term defection (i.e., which of the presently active households would defect by the year-end) had to be constructed from operational data. The householding of our customer information alone was not sufficient. Different households came to the company at different times and terminated their relationship with the company (if they defect) at different times. A uniform base file was needed to enable our models to predict yearly defection.

- *Selecting the right tools for analyzing the data and evaluating the trained models*

Several analysis techniques are available for predicting short-term defection. Our choice of the technique was driven by several considerations. The chosen technique should be able to handle missing information, accept both discrete and continuous valued attributes, and deal with noise in the dependent variable (i.e., two or more observations in the dataset have exactly the same attributes but belong to different output classes). Most importantly, the learned model must be interpretable by an insurance domain expert. The performance of the learned models can be measured using different yardsticks including overall classification accuracy, error in predicting the event of interest (in our case defection), and model gains and lift curves.

Although the problem we studied (*customer defection*) and the analysis techniques we used (*logistic regression* and *CHAID-style decision tree*) are not new, the unique challenges outlined above and the magnitude and scale of the problem on hand make our methodology interesting. The approach is fairly general and can be applied to other similar datamining and knowledge discovery projects.

Section 2 presents the approaches and methods we used to prepare the data for analysis. Section 3 outlines the predictive modeling phase and summarizes the results. Section 4 concludes with a look at the limitations of our work and suggests future directions.

## **2 Data Preparation Issues and Methodologies**

### **2.1 Household Policy Level Information**

The first challenge for us involved *householding* of policy level information. Our goal was to predict defection on a *household* level. However, the data in the operational databases are on a policy level. The operational database available for mining consisted of policies that were active as of January 1, 1994 and all new policies issued thereafter. A household might own several different policies such as automobile, homeowner, and life. These policies might be purchased at different times. The household might have added new policies or terminated existing ones during its tenure. Further, customers who moved to a different address had to be tracked. Information about movers is not very precisely captured in our databases. Data from external data sources are costly and often incomplete. For the current project we restricted ourselves to information available in our company databases. Policies belonging to the same household were identified using a name and address match. Further, on most policies we have data on supporting (or concurrent) policies issued to the same household. These policy numbers were used to pull together additional policies belonging to the household. Tracking movers was by far the most difficult task. In certain cases, the new policies issued following a move contained the previous policy number (i.e., the identification number for the policy that was terminated as a result of the move). Movers for whom previous policy information was not available were not identified (and thus potentially treated as separate households).

After identifying the household, we assigned to it a unique identification number. All data on individual policies were assembled to construct household level information. This included the number of policies of each type (note that a household might hold more than one automobile policy or more than one home owner policy), the *earliest original date* for each type of policy (i.e., the date when this type of policy was first purchased), the *latest termination date* for each type of policy, and the length of the household's relationship with the company. A household was treated as active as long as it had at least one active insurance policy with the company. It was considered to have defected when it terminated all its policies.

## 2.2 Choice of Demographic Variables

Customer defection is often triggered by *important life-stage* events such as marriage, addition of a new teenage driver to the policy, and purchase of a new automobile or home. These life-stage events might result in a change in insurance requirements and/or an increase in the insurance premium. This prompts customers to shop for a better deal which, if available, could result in defection. The demographic information available to us was sketchy. Often the information was collected at the time the policy was first issued and seldom updated after that. Ideally, we would like to know the most current demographic status of the household while predicting its short-term defection. Further, the demographic data contained a lot of missing information. Purchasing demographic data from external data sources was one possible way to address both the issues of *data currency* and *data accuracy*. However, for the purpose of this project, we chose to use only internal demographic data taking precautions to pick up the latest demographic information whenever it was available.

## 2.3 Building the Analysis Files

The next challenge for us was to build an analysis file that would enable us to model short-term defection. Specifically, we were required to predict which among the current active households would defect within the next 12 months. The data available for analysis included customer information over a 5-year period (1994-1998). After householding we had information about households whose relationships with the company began and terminated at different points in time. For most applications this would be sufficient. A *defection flag* (1 or 0) could be assigned to the household based on whether or not it had defected. However, there were two problems with this approach. Firstly, the predictive models developed using this data would not be able to directly predict yearly defection. Secondly, the data constructed this way would be biased since we did not have any information about households that had defected prior to 1994. Our files did contain information on households that had established a relationship with the company prior to 1994 and had stayed active at least until January 1, 1994. Completely eliminating households prior to 1994 would not be feasible since these longer-term customers are usually more profitable and it is required to predict defections among these customers very accurately.

In light of the above, we decided to take a snapshot by including all active households as of January 1, 1998. Each household was assigned a *defection flag* (1 or 0) depending on whether or not it had defected by December 31, 1998. Predictive models were then built using a randomly drawn training sample from the snapshot data. One way to test these models was to score an independent test sample (that was not used during the model construction) and to measure the accuracy of the model in predicting defectors among the households in the test sample. This would, in a sense, validate the models. An important question still remained to be answered. How could we be confident that these models would perform well when used to predict 1999 defectors? We constructed 5 snapshot data files, one for each of the years 1994-1998. Each yearly file included all the active households as of January 1<sup>st</sup> of that year and determined whether or not they had defected by the end of the year. Note that

households that first established a relationship with our company in, say, June 1996 and stayed active until March 1998 would appear in both the 1997 and 1998 snapshot files. We built models for each yearly file and tested them by predicting defection in the snapshot file of the following year. We hypothesized that if a model for a particular year predicts the following year's defectors with reasonable accuracy and, further, if the yearly variations are minimal in terms of model fit, gains, and the independent variables appearing in the model, then barring any major upheaval in the market dynamics, the model constructed using the 1998 snapshot file could reliably predict the 1999 defectors.

Each snapshot dataset contained information of about 500,000 to 600,000 households. Each observation was described in terms of 86 attributes (independent variables) capturing the policy level and demographic profile of the household<sup>1</sup>. The policy level information included household activities such as whether each particular type of policy (automobile, home, life, etc.) was currently active, previously owned but presently terminated, or not owned at all; a count of the total number of policies owned by the household; the household's relationship length until the beginning of the snapshot year; and so on. The household demographics were described by variables such as gender, age, marital status, estimated income level, and home ownership status of the head of the household and other information such as the presence of children and length of residence.

Considerable time and effort were dedicated to data cleaning. Several demographic variables were poorly populated. In some cases (e.g., estimated income level) we had data from two independent sources. In cases like this we used the attribute coming from the most reliable source (based on experience) for modeling. Data from the other source(s) were used to fill in missing values wherever feasible. Some demographic variables with too many missing values (in some cases as high as 80% missing data) were dropped from the analysis. We looked at the correlation matrix of the remaining independent variables and selected just one from among each group of highly correlated variables. The data cleaning effort left us with 28 independent variables that could be used as potential predictors.

The choice of the training and test sample sizes involved a trade-off between the size of the data files that were available and the ability of the analysis techniques and our computational infrastructure to handle the data. After some initial experimentation with samples of different sizes, we settled on randomly drawn training and test samples of 50,000 records each from each of the snapshot files. Thus, each training and test sample was about 10% of the size of the corresponding yearly snapshot file. The tools available to us for doing logistic regression and decision tree analysis were both able to handle datasets of this size comfortably. Each model was also validated on the entire data (nearly 500,000 odd households) of the following year. Care was taken to ensure that the proportion of defectors and survivors in the training and test samples was the same as the proportion in the entire snapshot file for the corresponding year. However, the distribution of the defectors and non-defectors among each snapshot file was skewed. In each case we had a very small proportion of defectors as compared to the survivors. When the event of interest (defection) is

---

<sup>1</sup> We are unable to provide more details about the independent variables as they constitute proprietary information.

significantly under-represented, predictive models tend to simply predict the *default class* (in our case, survival) for all observations. Such a model would be completely useless to us. These effects can be partially alleviated by incorporating differential misclassification costs or by training the models using a *balanced* training sample. We constructed *balanced* training samples containing an equal proportion of defectors and non-defectors. Separate models were constructed using these training samples. These models were then tested on a holdout sample containing the original proportion of defectors and non-defectors.

## 2.4 Overview of the Analysis Techniques

**Logistic Regression.** Each observation can be represented as a vector  $(x_1, x_2, x_3, \dots, x_n)$  in  $n$  dimensional Euclidean space where the  $x_i$ 's correspond to the independent variables. The dependent variable  $y$  is a binary response variable (with values 1 and 0 indicating defection and survival, respectively). The logistic regression model has the following form:

$$\text{logit}(p_i) = \log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

where,

$p_i = \text{Prob}(y_i = 1 \mid \mathbf{x}_i)$  is the probability of the response variable modeled

$\beta_0$  is the intercept parameter

$\beta_i$  ( $i = 1 \dots n$ ) are the slope parameters

$\mathbf{x}_i$  is the vector of independent variables

Thus, logistic regression models the logit transformation of the  $i^{\text{th}}$  observation's event probability,  $p_i$ , as a linear function of the independent variables  $x_i$ . The logit function is most commonly used as it can be more easily interpreted as opposed to some other functions such as the *complementary log-log function*.

Logit analysis estimates coefficients by maximizing the *likelihood* that an event will occur. Given this method of estimation, we assess the model fit by the likelihood value, i.e., -2 times the log of the likelihood (-2LL). The significance of the coefficients (that they are non-zero) uses the Wald statistic [3,4].

**Decision Trees** A decision tree partitions the  $n$  dimensional input space (defined by the  $n$  independent attributes) into mutually exclusive regions, each of which is assigned a particular class label. The decision tree mechanism is transparent and a tree can be easily interpreted to determine how a decision was made. The decision tree structure contains *internal* and *external* nodes connected together by *branches*. Each internal node implements a decision function that determines which of the child nodes should be visited next. An external node (also called *leaf* or *terminal* node) has no children and is associated with a class label that characterizes all the data that has reached that node. Consider that we have a decision tree and are required to classify an observation (described as a vector of  $n$  attributes). Starting at the root node, apply the decision function and, based on the result of the decision, select one of the child nodes to visit next. This process of applying the decision function at each child node

is continued until the node reached is a terminal node. At this point the class label associated with the terminal node is returned as the class of the given observation.

Decision tree induction is widely studied in both the statistics (CART [2] and CHAID [1,5]) and the machine learning communities (ID3 [6] and C4.5 [7]). A typical decision tree learning algorithm starts with all the training examples and determines which among the different independent variables (attributes) provides the *best* split among the different classes denoted by the dependent variable. The decision tree learning algorithms differ in terms of the criterion used to evaluate the best split. For instance, the CHAID algorithm ranks the different attributes using the chi-squared test of independence between each independent variable and the dependent variable. On the other hand, ID3 uses the entropy measure to determine the amount by which the split on each independent variable can reduce the uncertainty associated with the dependent variable. The attribute that results in maximum reduction in entropy is selected for splitting at the node.

Issues such as preventing over-fitting of training data, handling missing values, splitting on a continuous valued attribute, and dealing with noise in the dependent variable are standard. The interested reader is referred to [6,7] for a more complete treatment of these issues.

## 2.5 Evaluating the Performance of the Learned Model

Several metrics can be used to evaluate the performance of a learned model. A *confusion matrix* depicts the overall classification accuracy and the accuracy of predicting each individual class. In certain applications such as predicting defection it is acceptable to trade-off overall classification accuracy for a higher accuracy in predicting the class of defectors. Further, in order to design a suitable marketing campaign, it is not enough to simply predict whether a customer will defect or not. Instead, it is required to assign a metric that would allow the marketing department to rank order the entire list of active households in descending order of their propensity to defect. Both logistic regression and decision trees are capable of assigning a probability that an observation belongs to the class of interest (defectors in our case). This was another reason why we adopted these two techniques for our analysis. Equipped with a rank-ordered list, the marketing campaign can be designed around, say, the top three or five (or some such suitable number) deciles among the active households. This would allow management to focus its efforts on the customers that are most likely to defect in the short-term. *Cumulative lift* and *cumulative gains* curves are two other metrics used to measure the performance of a model. In our application, the cumulative lift chart plots the cumulative fraction of the total defectors that have been identified by the observations belonging to the top decile, the top two deciles, the top three deciles, and so on. The cumulative gains chart, on the other hand, plots the cumulative defection rate for the deciles. Examples of the cumulative lift and gains charts are shown in section 3.

### 3 Results

As mentioned earlier, we used the 1998 snapshot file to build the initial models. The 1998 file was selected due to its proximity to our target year (1999). We used two modeling approaches for this effort: *decision tree* and *logistic regression* analysis. The results of the two techniques were compared based on the number and type of independent attributes used, percentage of defectors and survivors correctly classified (confusion matrix), gains and lift curves, the simplicity and parsimony of the models, and the interpretability of the results.

**Description of Predictor Variables:** 28 potential independent variables were available for modeling after data cleaning. The CHAID style decision tree algorithm constructed a tree using 16 variables (9 policy and 7 demographic variables). Logistic regression constructed a model with only 7 variables (5 policy and 2 demographic variables). All the 7 variables used in logistic variables were also in the decision tree.

**Percentage of Cases Correctly Classified:** The percentage of defectors and survivors correctly classified by each of the two techniques on the training and test sets is depicted in Table 1. Note that the results for the two techniques differ by less than 3 percentage points.

**Table 1.** Percentage of cases correctly classified by each modeling approach.

	Decision Tree		Logistic Regression	
	Training (%)	Test (%)	Training (%)	Test (%)
Defectors	65.5	66.4	66.3	67.3
Survivors	72.9	73.2	70.9	70.9
Overall	69.4	72.1	70.2	70.3

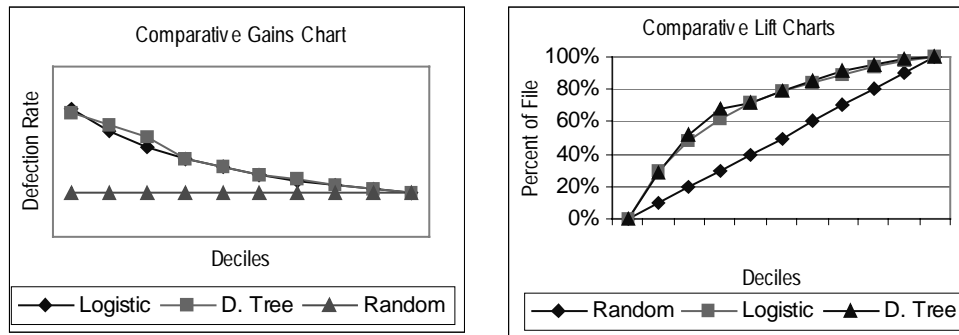
**Gains and Lift Curves:** We used gains and lift curves to further compare the results. The defection probabilities reported by the models were sorted in a descending order and arranged into deciles. The top decile contained customers with the highest probability of defection and the bottom decile the least.

Fig. 1 shows the cumulative gains and cumulative lift charts on the training set from the decision tree and logistic regression modeling. The cumulative gains chart compares the defection rate<sup>2</sup> at 10%, 20%, 30% up to 100% of the file. The gains represented by the chart show the relative predictive power of a model. Both techniques had a gain of more than 3 times in the top decile and almost 2.5 times in the top two deciles over the total defection rate. The cumulative lift curve shows the proportion of defectors correctly identified by the model at the top 10%, 20%, 30%, up to 100% of the file. Both techniques identified almost four-fifths (79%) of the defectors at the top 50% of the training set. The lift due to the models is compared to the random lift (the 45° line).

---

<sup>2</sup> Please note that the actual cumulative defection rates (along the Y-axis of the gains chart) constitute confidential information and hence have been omitted.

**Selecting Logistic Regression Over Decision Tree:** With very similar results, we opted for logistic regression to model defection in the other four snapshot years. The logistic regression model used only 7 independent variables as opposed to the 12 used by the latter. Besides, the former was also simpler to implement since it was described in terms of a single equation. The decision tree produced 128 terminal nodes. Deploying the tree for targeting purposes would be tedious. The coefficients in the equation produced by logistic regression showed the direction of defection and the partial correlation showed the relative importance of the independent variables.



**Fig. 1.** Comparative Gains and Lift Charts for Decision Tree and Logistic Regression.

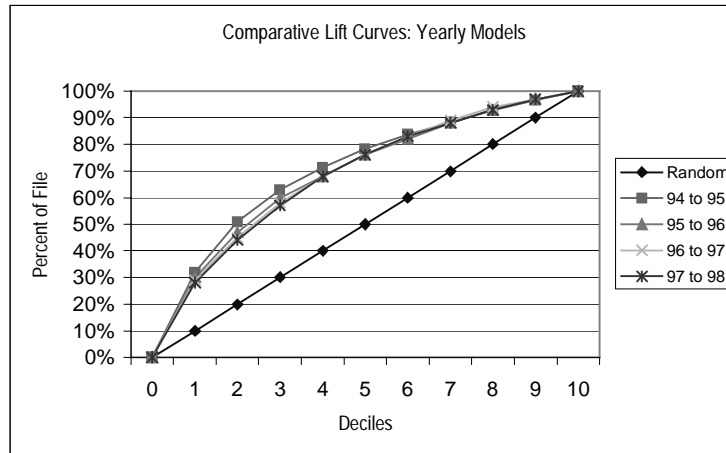
**Yearly Models 1994-1997:** These models were designed to test if each year’s model would uniformly predict the immediately following year’s known defectors and survivors. If that were the case, then the 1998 model would reliably predict the 1999 defectors assuming that there are no drastic changes in the insurance marketplace within that period.

Table 2 shows the percentage of defectors and survivors correctly classified when each year’s model was applied to the following year’s total file. The type and number of predictors in the yearly models varied somewhat but the percentage correct classification showed little variations. We also compared the cumulative lift curves of each year’s model applied to the following year’s total file to see if there is any variation.

**Table 2.** Percentage Correctly Classified Yearly Model Applied to the Following Year’s Total File.

Model Year	Defectors (%)	Survivors (%)	Overall (%)
1994	61.9	74.4	73.2
1995	58.9	74.2	72.4
1996	57.0	75.3	72.7
1997	60.9	71.9	70.2

Figure 2 shows the comparative lift curves. We see from these results that the yearly models are comparable in terms of their classification accuracy and the model lifts. Thus, our 1998 model should be able to reliably predict defectors in 1999.



**Fig. 2.** Comparing the cumulative lift curves obtained by applying each yearly model to the following year's data.

#### 4 Summary, Limitations, and Future Work

In this paper, we have presented the results of our study on mining operational databases to predict short-term defection among insured households. We have discussed issues that arose during our mining efforts such as householding policy level information, appending demographics, selecting the right analysis tools, and analyzing snapshot files and building yearly models. The results of our predictive modeling are encouraging. The models have very little variation from year to year and predict the following year's defectors with reasonable accuracy. Of course, the ultimate value of these models can be determined when the results are used in the field to identify likely defectors who may be targeted with retention campaigns.

We did not have access to data on defectors prior to 1994. Without these data we could not combine all the survivors and defectors into a single analysis file. This led us to the approach of building models on the individual snapshot files. We were able to track movers only within the same business region. Expanding our householding effort to capture movers all across the country is a computationally intensive and time-consuming process.

We will continue to improve our models in the near future. We plan to augment our demographic data (possibly external) that would help alleviate problems such as missing values and outdated information. Moreover, we will incorporate data from important trigger events such as bill payment and claims settlement and add them to

our list of possible predictors. Presently, we have decided to incorporate results only from the 1998 model. We are investigating techniques for combining multiple models (say, from the different years) for improving classification accuracy and constructing a more robust model. Another avenue that merits investigation is the use of survival analysis techniques to predict year-to-year defection (not just defection for the following year). It is also of interest to model the defection characteristics of households with particular types of insurance coverage.

## References

1. Biggs, D., de Ville, B. and Suen, E. A Method of Choosing Multiway Partitions for Classification and Decision Trees, *Journal of Applied Statistics* **18**: 49-62. 1991.
2. Brieman, L., Friedman, J., Olshen, R., Stone, C. *Classification and Regression Trees*. Wadsworth Inc., Belmont, CA. 1984.
3. Goodman, L. The Logit Model, *Analyzing Quantitative/Categorical Data*. Magidson, J. (editor), University Press of America, Lanham, MD. pp. 5-54, 1978.
4. Hair, J., Anderson, R., Tatham, R. and Black, W. *Multivariate Data Analysis* (Fourth edition). Prentice Hall, Englewood Cliffs, NJ. 1995.
5. Kass, G. V. An Exploratory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics* **29**: 119-127. 1980
6. Quinlan, J. R. Induction of Decision Trees. *Machine Learning*, 1:81-106. 1986.
7. Quinlan, J. R. *C4:5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA. 1993.