

# **Highlight Case Study 1:**

## ***High value Customer identification model development***

### **Executive Summary**

We helped our customer to build models for their insurance client who offers a wide range of property and casualty products through an extensive network of independent agents. Personal lines are distributed through 2,400 independent agents in 19 states. The Commercial Lines segment targets the small to mid-size business market including commercial multiple peril, commercial automobile, workers' compensation and other commercial coverage. The commercial lines are distributed through a network of over 2,000 agents. The client also independently deals with and manages hundreds of thousands of customers.

### **Business objective:**

Identify high value customers defined as those customers who might buy multiple products and stay with the company for a long time.

### **Methodology:**

Using the client's customer data along with demographic data, we developed a High Value Customer predictive model. Below are the steps that we followed

### **Data receiving mechanism**

Client sent the flat file with layout. This file consists of the client's customer's data. Client also has data on US demographics. We first converted these datasets into SAS format using the layouts provided by the customer. After receiving the above two files, we converted the flat files to SAS datasets using the given layouts.

### **Data Audit and cleansing**

We then conducted a data hygiene audit using the following checks:

- Cross checking the no. of records in the flat file
- Accuracy and completeness of variable information and values
- Number of duplicate records
- Number of records of deceased individuals
- Percentage of missing values and typo errors for each variable in the data.
- Check with the specifications list for whether all the variables mentioned in the layout files are available in the data.

**Data and the Statistics:**

1,951,508 records were received from Client

1,276,425 were used for the core analysis file after hygiene and deduplication processing.

1,051,922 records remained after enhancement matches.

1,047,761 were the sample size for this model building.

**Model building process:**

1. We first combined the data with a common variable from two data files and used a proprietary method to identify the most important variables that should be included in the model (which was a Logistic Regression)
2. We then applied transformations to some variables to improve their predictive ability.
3. After reviewing the outputs, we selected the most important variables
4. We then applied additional transformations for some selected variables.
5. We then tested and corrected for multicollinearity.
6. We then split the data file into a development sample and a validation sample
7. We then ran the model on the development sample
8. Using the estimated parameters, we developed a lift chart for the development sample.

**Validation of the model:**

We used the calibrated model to score the records in the validation sample and developed a lift chart for this sample. We, typically, use several methods to assess model accuracy:

- Concordance
- Cumulative lift chart
- Area under ROC curve
- Confusion matrix
- Divergence
- KS distance

In this project, we used Cumulative lift chart and area under ROC curve to test the accuracy of the model.

**Testing and Scoring of the model:**

Using a sample of 30% of the data, we tested the stability of the model and scored each record using the model.

We then compared both the lift charts and estimated the KS value.

**Software used:**

SAS 9.1, SAS/Stat, SAS Analytics PRO on Windows-XP server