Acquiring Insurance Customer: The CHAID Way

P H Anantha Desik and Samarendra Behera***

In view of the recent regulatory changes and volatile market conditions, the insurance and finance industry started focusing more on developing strategies to find new customer segments. Acquiring new customers is difficult especially in a fiercely competitive industry like insurance. Also, the dynamic nature of pricing for insurance products makes it even more challenging. In this context, it is essential for an insurer to have knowledge about customers, plan customer-centric offerings, attract more profitable customers and increase the bottom line. In the changing market and economic conditions, the insurance industry has started considering customer-centric rather than a product-centric view to well serve the customers. Information Technology (IT)-driven data analytics now has the capability to discover knowledge hidden inside very large amount of data to help in making business decisions which are customized to customer needs. This paper is an attempt to create business rules from customer lead data which will help in identifying customer segments for better marketing campaign and to acquire new customers, and also to explore answers for specific business problems like low lead conversion ratio, important attributes influencing lead conversion and right customer profile to optimize lead conversion.

Keywords: CHi-squared Automatic Interaction Detection (CHAID), Customer acquisition, Lead conversion in insurance

Introduction

In general, customer acquisition in insurance starts with identifying quality potential customers. This is sometimes accomplished by locating individuals and businesses that either express interest in or already use products similar to those produced by the business. From this initial list, these leads are then qualified a little further using various research methods to determine if there is any possibility of making a sale with a given lead. The prospective customer is then explored further after understanding the preference and behavior. This emphasizes the need for understanding customers and to be able to respond proactively to their needs. It also helps in designing customized products and services specific to customer needs.

Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was used for the implementation of this project. CRISP-DM breaks the process of data mining into six

* Head, Actuaries and Analytics, TCS ODC3, Synergy Park, Gachibowli, Hyderabad 19, Andhra Pradesh, India. E-mail: a.desik@tcs.com

** Lead, Analytics-Insurance, TCS ODC3, Synergy Park, Gachibowli, Hyderabad 19, Andhra Pradesh, India. E-mail: samarendra.b@tcs.com

major phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. The four important advantages of this methodology are: industry neutral, tool neutral, closely related to the knowledge discovery in databases process model, and anchoring the data mining process. We received 100K lead or prospect data which can be categorized into three areas as follows:

1. Insurance policy related data: Current product (yes/no), current product type, current coverage, new product type, new coverage;
2. Customer demographic variables: Age, gender, marital status, family members, education, occupation, job title, income; and
3. Decision variables: Status (converted/ not converted) rating.

Decision variable, 'status' having two values—converted and not converted—was taken as a dependent variable. The business understanding phase was iterative and several questions related to client business were clarified. This step was essential to validate the quality of data mining results. Successful completion of the business understanding phase leads to an understanding of the data and allows to put more business context to each variable. The data was prepared to apply the data mining algorithm, i.e., CHi-squared Automatic Interaction Detection (CHAID). Exploratory data analysis (univariate, frequency distribution, distribution of converted status in each class, etc.) was carried out for an in-depth understanding of the variables. Missing value imputation and outlier management were also carried out as required. To understand the interrelationship among the data points, correlation analysis was carried out. Several business-relevant derived variables were created. Also, class variables were created using dummy variables to make it analytic ready.

The data was partitioned into two samples, namely, training (50%) and testing (50%) using simple random sampling. The samples were also validated for proper representation of population both in terms of time and distribution of converted/non-converted. Ten different samples were created with replacement to validate the accuracy of the model.

Techniques/Tools Used

CHAID was used as the data mining technique. It is a technique based on multiway splitting to create discrete groups and understand their impact on the dependent variable. CHAID was preferred for analysis because of five major criteria:

1. A good proportion of input data was categorical;
2. Its efficiency in large datasets;
3. Its highly visual and ease of interpretation;
4. Ease of implementation/integration of business rules generated from CHAID in business; and
5. Input data quality can be handled efficiently.

Initially, all the variables were passed into CHAID algorithm for classification. CHAID model was iteratively developed on training data using pruning criteria like:

a. Any particular node should have a minimum of 3% observations.

b. Post pruning was applied to remove some insignificant variables in terms of segregation power.

Six different variables based on importance were identified to develop the final tree. SAS Enterprise Miner 5.2 was used to create the tree and also to create the data samples. The following criteria were applied to evaluate the model:

- Kolmogorov Smirnov (KS) statistics;
- Lift/gain chart; and
- Models were also evaluated for accuracy in all the test samples apart from the above mentioned statistics.

Finally, the champion model was selected out of various competing model. Final efficient segments in terms of conversion rate were obtained, and also the list of influential variables was obtained for scoring and implementation.

Findings

The final decision tree is depicted in Figure 1. Overall, six variables were used to define nine nodes. The six variables were found to be significant in the final model with job title, income, age, having current product (yes/no), current coverage and new coverage amount. Among these six variables, income was found to have a positive impact on the status, whereas all the others had a negative impact. More specifically, job title like business analyst, office clerk, personal assistant, programer, photographer and researcher were found to have a significant impact on the conversion status. Lift chart (Figure 2) showed us that approximately 88% of the converted customers lie in 40% of leads. Substantial increment in KS was noticed in the Nodes 4, 1, 3 and 6 in that order. These four nodes were used to create segments that were 'very likely' followed by Nodes 5 and 2 that are 'most likely' (Table 1). Figure 3 depicts the performance of the model in both training and test dataset. Performances in both the samples were found to be stable and hence the dependability on each rule was found to be trustable across various samples. Lead data was analyzed to seek answers to the four questions to help the client. The answers to each of them were as follows:

- Lead conversion ratio was low because 50% of the targeted customers were from least preferred segment making the overall conversion ratio to 38% from approximately 84%.
- Job title was found to be the most important variable influencing the conversion status, optimal customer profile for maximum conversion was found to be customers with job titles—business analyst, office clerk, personnel assistant, programer, photographer and researcher—with current coverage of more than or equal to 95 lakh, age between 40 and 55 years, and those who do not have an existing policy.

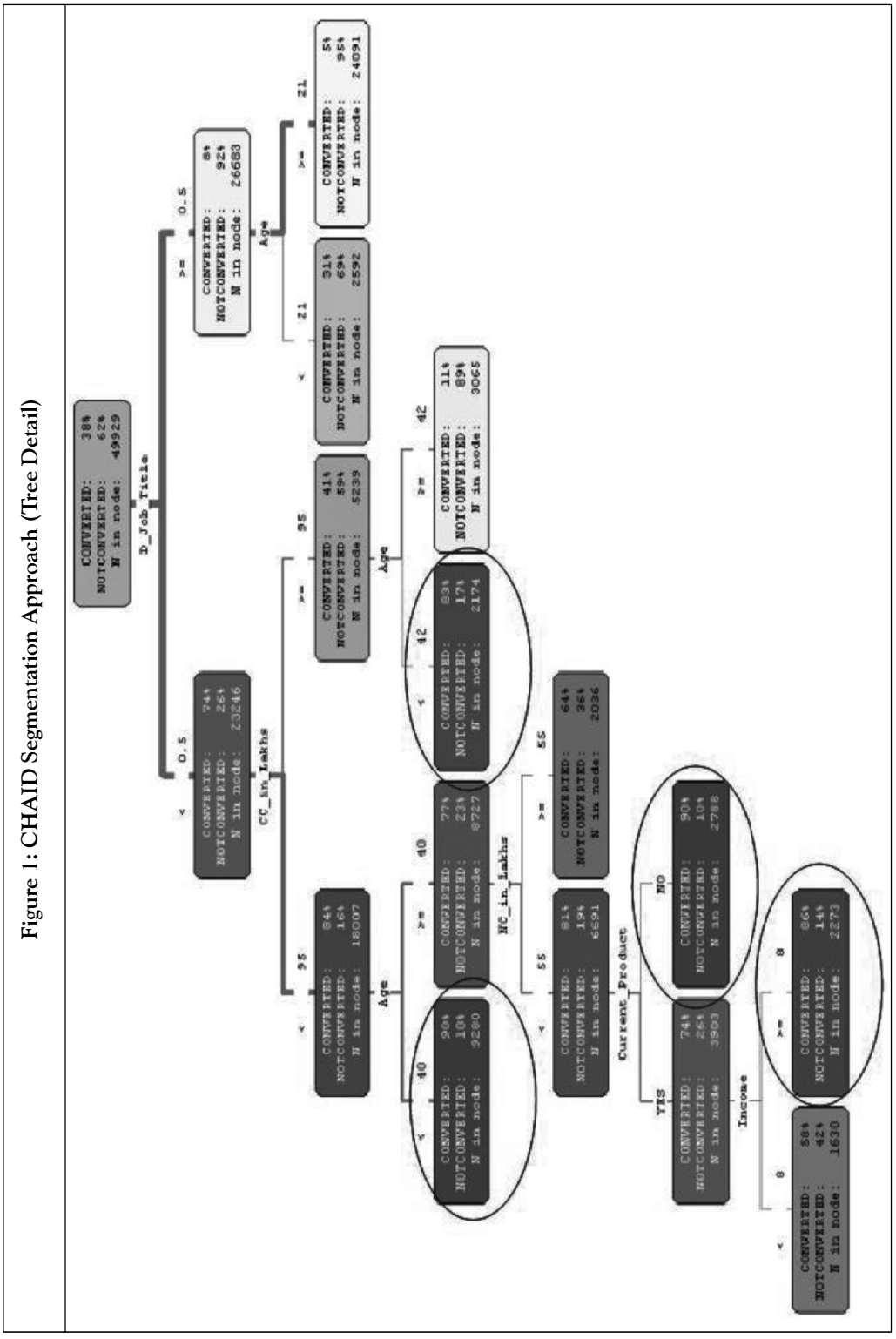


Figure 1: CHAID Segmentation Approach (Tree Detail)

Figure 2: Lift and Gain Chart Across Deciles

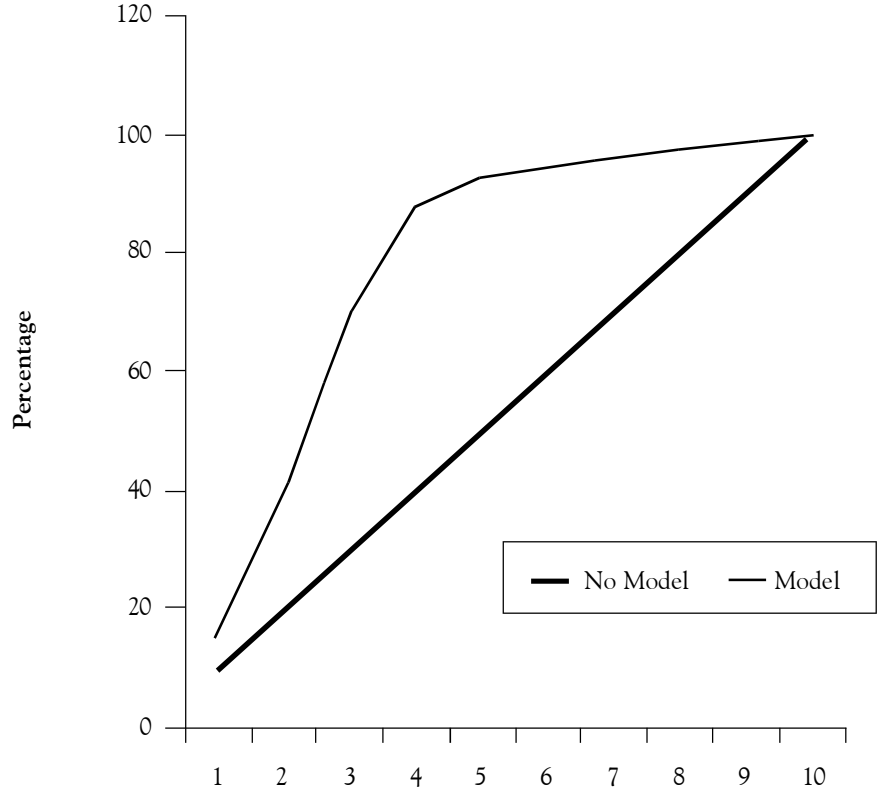
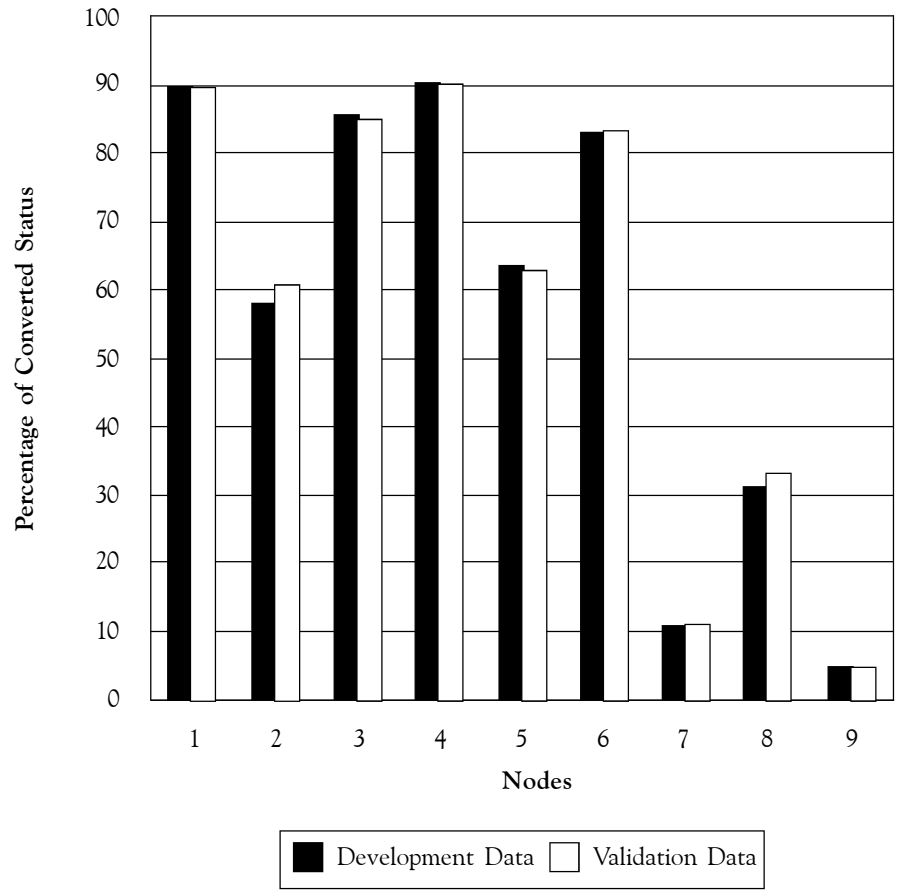


Table 1: Node Performance (KS Table)

Node	No. of Obs.	No. of Converted	% Converted	Cumu. Obs.	Cumm. Converted	Cumm. Non-Converted	% Cumm Obs.	% Cumm. Converted	% Cumm. Non-Converted	Cumm. Conversion Rate	KS
4	2,814	2,537	90.16	2,814	2,537	277	5.62	13.27	0.89	90.16	12.38
1	9,247	8,295	89.70	12,061	10,832	1,229	24.09	56.66	3.97	89.81	52.69
3	2,203	1,873	85.02	14,264	12,705	1,559	28.49	66.46	5.04	89.07	61.42
6	2,179	1,818	83.43	16,443	14,523	1,920	32.84	75.97	6.20	88.32	69.76
5	2,030	1,279	63.00	18,473	15,802	2,671	36.89	82.66	8.63	85.54	74.03
2	1,611	984	61.08	20,084	16,786	3,298	40.11	87.80	10.65	83.58	77.15
8	2,560	853	33.32	22,644	17,639	5,005	45.22	92.26	16.17	77.90	76.09
7	3,027	333	11.00	25,671	17,972	7,699	51.27	94.01	24.87	70.01	69.13
9	24,400	1,146	4.70	50,071	19,118	30,953	100.00	100.00	100.00	38.18	0

Figure 3: Percentage of Converted Across Development and Validation Data



- Least preferred customer profile was found to be customers with job titles—cab driver, crop farmer, dump truck driver, farmer horticultural, operator machine, and operator tractor—and age more than 21 years.
- Right customers were not targeted as approximately 49% of targeted customers were from the least preferred customer profile with a conversion rate of 4.7%.

Conclusion

Customer acquisition has always been a vital area for any insurer. Modern age Information Technology (IT)-driven analytic solutions are helping insurers to optimize this process by enabling faster, apt and optimal decision making to substantially increase the marketing Return on Investment (ROI). This helps the insurers to increase the bottom line and also to add to the top line by curtailing marketing effort on unwanted segments of customers. Acquiring new customers is always costlier than retaining the existing ones; so IT-driven

analysis has the potential to save a lot of money by targeting the optimal customer segments. An increasing number of insurers are now capitalizing on the IT-driven analytics to be in an advantageous position in a fiercely competitive regulatory insurance market. This paper establishes the benefits of applying analytics solutions for solving business problems related to insurance.■

Acknowledgment: *The authors thank TCS Insurance Management for providing an opportunity to bring out the researcher in us and supporting us.*

Bibliography

1. Bowman D and Das N (2004), "Linking Customer Management Effort to Customer Profitability in Business Markets", *Journal of Marketing Research*, Vol. 41, November, pp. 433-447.
2. Colombo R and Weina J (1999), "A Stochastic RFM Model", *Journal of Interactive Marketing*, Vol. 13, Summer, pp. 2-12.
3. Cooper F, Lee G and Giovanni G (2000), "Turning Data Mining into a Management Science Tool: New Algorithms and Empirical Results", *Management Science*, Vol. 46, No. 2, pp. 249-264.
4. Frédéric Tremblay (2010), "Embedded Value Calculation for a Life Insurance Company", available at http://wenku.baidu.com/view/c4ee0511cc7931b765ce1592.html
5. Gupta S and Donald R L (2003), "Customer as Assets", *Journal of Interactive Marketing*, Vol. 17, No. 1, pp. 9-24.
6. Thomas J S (2001), "A Methodology for Linking Customer Acquisition to Customer Retention", *Journal of Marketing Research*, Vol. 38, No. 2, pp. 262-268.

Reference # 29J-2012-07-01-01