FasterAnalytics for Consumer Goods – A Case Study

Introduction

New England Catalog is a well-established catalog retailer and mails a catalog to each of over 5 million customers 13 to 18 times per year. New England Catalog observed that while up to 64% of the list responds over the year, of those who respond there is a wide variation in the frequency and value of each response. This would indicate that it should be possible to develop a more sophisticated targeting policy that can recommend "mail-to" customers and the likely monetary value of their response.

New England Catalog purchased both the FasterAnalytics software suite as well as consulting services from DecisionQ.

Project Goals

The goals of the DecisionQ FasterAnalytics project were the following:

1. Leverage a rich history in the New England Catalog transactional data to save costs by reducing the number of catalogs distributed while maintaining revenues.

2. Visualize the relationships in the data in an interactive browser so that the executives at New England Catalog could understand how decisions were made for each individual and have the confidence to deploy a state-of-the-art solution.

3. Use the model to produce a ranked list of all prospective customers for mailing, predicting which of the proposed candidates will respond and what the expected profit value of that response would be.

Sequence of Events

New England Catalog purchased a bundle of software and services from the DecisionQ. This bundle included an annual software license for data analysis, data modeling and professional services. This bundle was designed to empower our customer to develop a predictive modeling program that optimized the direct marketing policy over the next year.

The suite is a unique solution because of the patent-pending technology of DecisionQ Corporation. Other predictive modeling solutions offer traditional approaches to prediction such as logistic regression, decision tree induction, and neural networks based on technologies such as SAS Enterprise Miner and IBM Intelligent Miner. DecisionQ Corporation provides a product that is optimized for Customer Targeting based on Bayesian Networks. Bayesian Networks enable decision stakeholders a graphical view of how decisions are made while providing the highest quality predictions and decisions.

Every four weeks (thirteen times per year) New England Catalog produces and sends a catalog costing $1.50 each to over 5 million potential customers. Over the year, approximately sixty-four percent of the customer list order from the catalog. The new higher level of sophistication and quality provided by DecisionQ FasterAnalytics will allow New England Catalog to deploy an optimized mailing policy that meets their goal of reducing targeting cost while maintaining revenue.

Methodology

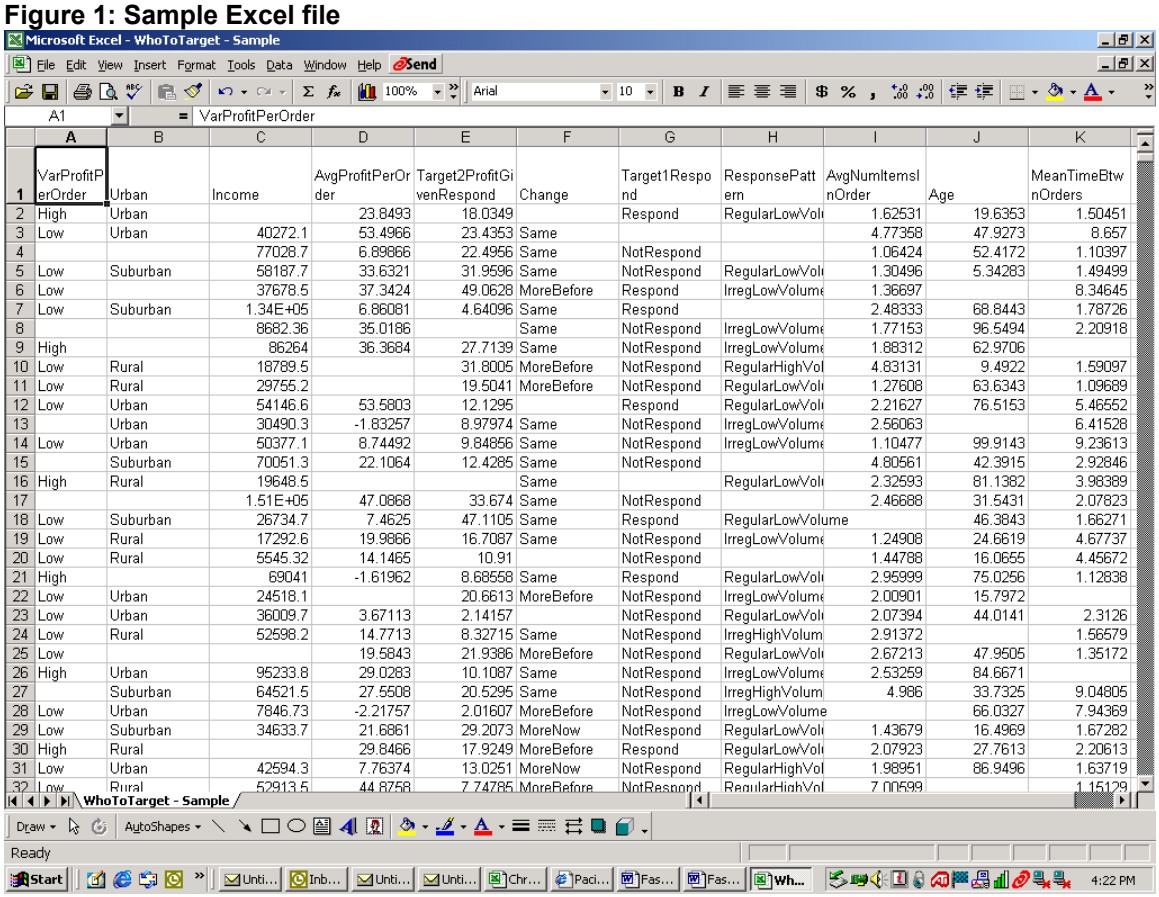
The stated key to developing a more sophisticated mailing policy for New England Catalog was to be able to predict which customers would respond to targeting and how profitable that response would be. Some customers respond regularly with large orders while other customers respond

less frequently with different order sizes. The ideal solution for optimizing a single campaign would be to rank the customers by the expected profitability for a particular campaign and then to mail to those customers with a positive expected profitability. This measure of profitability would account for the cost of the mailing campaign and all other costs. FasterAnalytics was designed to provide this ideal solution.

The Data

The data offered was not in a single location and did not have a common standard of reference organization or quality. DecisionQ Corporation began by working with New England Catalog to translate their legacy system transactional data into a data warehouse that was better suited for understanding customer behavior. This analytical data warehouse was designed specifically to track purchase history. DecisionQ customer engineers designed and implemented variables from the raw data that are useful for this type of analysis. In a large number of cases, customers already have data in a format that is ready for use with the FasterAnalytics product. Such formats would include Excel (Figure 1), Access, or any another ODBC-compliant database.

Figure 1: Sample Excel file



	A	B	C	D	E	F	G	H	I	J	K
1	VarProfitPerOrder	Urban	Income	AvgProfitPerOrder	Target2ProfitGivenRespond	Change	Target1Respond	ResponsePattern	AvgNumItemsInOrder	Age	MeanTimeBtwnOrders
2	High	Urban		23.8493	18.0349		Respond	RegularLowVol	1.62531	19.6353	1.50451
3	Low	Urban	40272.1	53.4966	23.4353	Same			4.77358	47.9273	8.657
4			77028.7	6.89866	22.4956	Same	NotRespond		1.06424	52.4172	1.10397
5	Low	Suburban	58187.7	33.6321	31.9596	Same	NotRespond	RegularLowVol	1.30496	5.34283	1.49499
6	Low		37678.5	37.3424	49.0628	MoreBefore	Respond	IrregLowVolume	1.36697		8.34645
7	Low	Suburban	1.34E+05	6.86081	4.64096	Same	Respond		2.48333	68.8443	1.78726
8			8682.36	35.0186		Same	NotRespond	IrregLowVolume	1.77153	96.5494	2.20918
9	High		86264	36.3684	27.7139	Same	NotRespond	IrregLowVolume	1.88312	62.9706	
10	Low	Rural	18789.5		31.8005	MoreBefore	NotRespond	RegularHighVol	4.83131	9.4922	1.59097
11	Low	Rural	29755.2		19.5041	MoreBefore	NotRespond	RegularLowVol	1.27608	63.6343	1.09689
12	Low	Urban	54146.6	53.5803	12.1295		Respond	RegularLowVol	2.21627	76.5153	5.46552
13		Urban	30490.3	-1.83257	8.97974	Same	NotRespond	IrregLowVolume	2.56063		6.41528
14	Low	Urban	50377.1	8.74492	9.84856	Same	NotRespond	IrregLowVolume	1.10477	99.9143	9.23613
15		Suburban	70051.3	22.1064	12.4285	Same	NotRespond		4.80561	42.3915	2.92846
16	High	Rural	19648.5			Same		RegularLowVol	2.32593	81.1382	3.98389
17			1.51E+05	47.0868	33.674	Same	NotRespond		2.46688	31.5431	2.07823
18	Low	Suburban	26734.7	7.4625	47.1105	Same	Respond	RegularLowVolume		46.3843	1.66271
19	Low	Rural	17292.6	19.9866	16.7087	Same	NotRespond	IrregLowVolume	1.24908	24.6619	4.67737
20	Low	Rural	5545.32	14.1465	10.91		NotRespond		1.44788	16.0655	4.45672
21	High		69041	-1.61962	8.68558	Same	Respond	RegularLowVol	2.95999	75.0256	1.12838
22	Low	Urban	24518.1		20.6613	MoreBefore	NotRespond	IrregLowVolume	2.00901	15.7972	
23	Low	Urban	36009.7	3.67113	2.14157		NotRespond	RegularLowVol	2.07394	44.0141	2.3126
24	Low	Rural	52598.2	14.7713	8.32715	Same	NotRespond	IrregHighVolum	2.91372		1.56579
25	Low			19.5843	21.9386	MoreBefore	NotRespond	RegularLowVol	2.67213	47.9505	1.35172
26	High	Urban	95233.8	29.0283	10.1087	Same	NotRespond	IrregLowVolume	2.53259	84.6671	
27		Suburban	64521.5	27.5508	20.5295	Same	NotRespond	IrregHighVolum	4.986	33.7325	9.04805
28	Low	Urban	7846.73	-2.21757	2.01607	MoreBefore	NotRespond	IrregLowVolume		66.0327	7.94369
29	Low	Suburban	34633.7	21.6861	29.2073	MoreNow	NotRespond	RegularLowVol	1.43679	16.4969	1.67282
30	High	Rural		29.8466	17.9249	MoreBefore	Respond	RegularLowVol	2.07923	27.7613	2.20613
31	Low	Urban	42594.3	7.76374	13.0251	MoreNow	NotRespond	RegularHighVol	1.98951	86.9496	1.63719
32	Low	Rural	52913.5	44.8758	7.74785	MoreBefore	NotRespond	RegularHighVol	7.00599		1.15129

Next, the customer engineers implemented a statistical process for training and validating a model that would predict the target variables. This included both qualitative and quantitative evaluations of the models. The data was divided randomly into three equal parts for training, testing, and validating models. A Bayesian Network was trained and tested using the first two data sets. This process was repeated until the customer engineers were satisfied with the performance of the final model. They then tested the model again on the third data set to ensure that the performance was consistent with that on the second data set.

Executives were provided with an interactive browser for the model that allowed them to see insights into the process of how their customers were behaving. They entered information for

specific people and then watched the flow of the information that results in a unique decision for that specific customer. This provided a qualitative validation of what was learned in the model.

Creation of Analysis Variables

Two types of variables were created, target variables and analysis variables. Target variables are the factors that one wishes to predict and analysis variables are summaries of the rich history of static and transactional customer data. The target variables in the case are clear.

Target Variables	
Target1Respond	Will a customer respond to the next promotion
Target2ProfitGivenResponded	Profit earned from a response

When designing analysis variables there are two main considerations:

- First, we can think of the set of variables as covering concepts that we believe are important to predicting the target variables, e.g. response pattern, recent behavior, recent vs. past behavior, urgency, loyalty, and demographics. The initial set of variables was designed to cover all of these concepts. Balance was important as to not over emphasize any one type or category of variable in particular.
- Second, when there is not much history for a customer, demographic variables can be used to infer future behavior. Thus, we can use demographics to help segment new members until transactions become significant.

Analysis Variables	
Variable	Concept
Urban	Urban Suburban or Rural in the customers zip code
Income	Income per capita in customers zip code
Age	Average age in the customers zip code
MeanTimeBtwnOrders	Average number of months between responses over the year
AvgNumItemsInOrder	Average number of items purchased in a response over the year
Change	Total items purchased Q3 Q4 minus total items purchased Q1 Q2
AvgProfitPerOrder	Average profit per response in the year
VarProfitPerOrder	Variance of profit per response in the year
ResponsePattern	Response pattern over the last year

Model Training

Once analysis variables are created and stored in a data warehouse with JDBC access methods, creating a Bayesian Network with DecisionQ FasterAnalytics is fast and easy. Numeric variables such as Income are automatically binned optimally for predicting the target variables. For example, AvgNumItemsInOrder was divided into three bins, 1 to 3, 3 to 6, and above 6 items. The patent-pending DecisionQ approach then searched an extremely large number of potential models of the relationships among the variables to find the model that was best for maximizing the profit of a campaign. The best model was output to a number of standard formats for viewing and making test predictions.

Analysis

Qualitative Validation

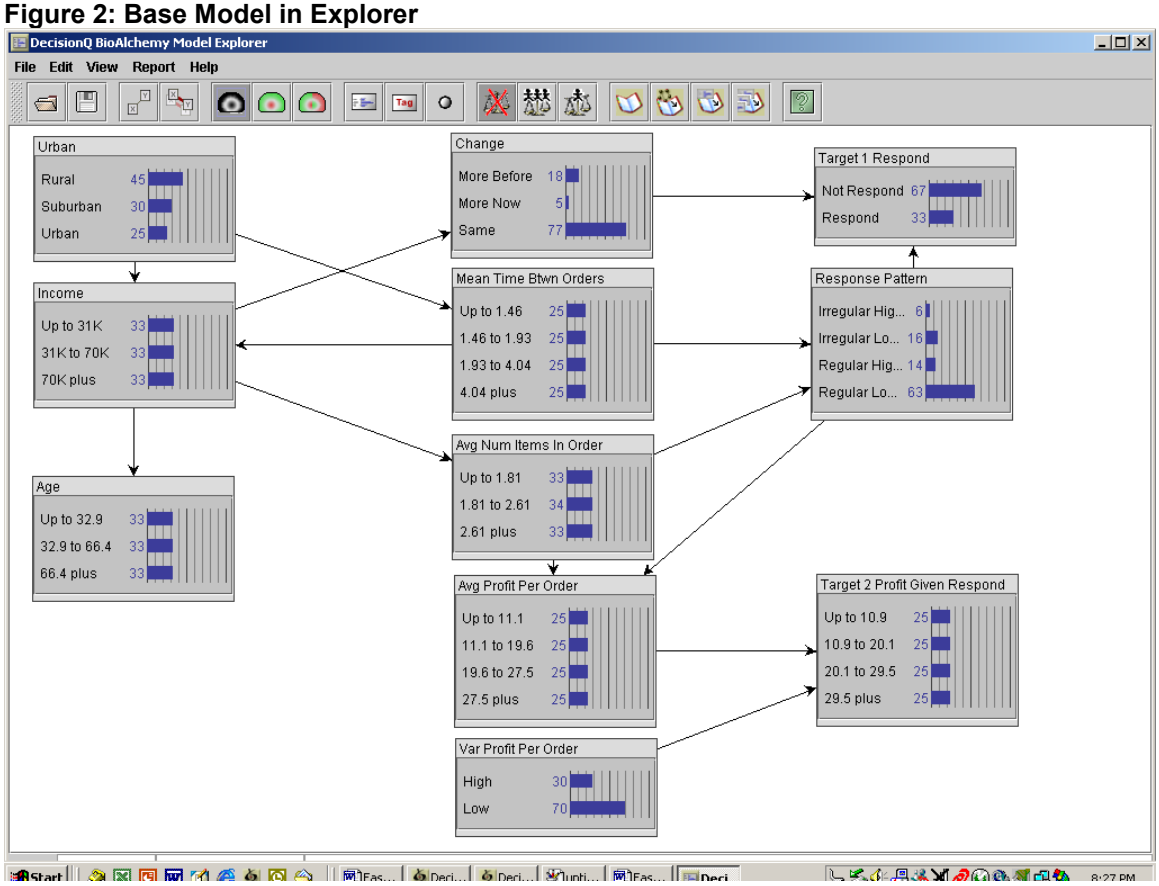
There were two ways to qualitatively validate a model from the perspective of Customer Targeting:

- Examine the relationships that were found in the data to ensure that they make sense.

- Use the DecisionQ graphical interface to enter sample customers to understand how the model will predict new cases or potential customers.

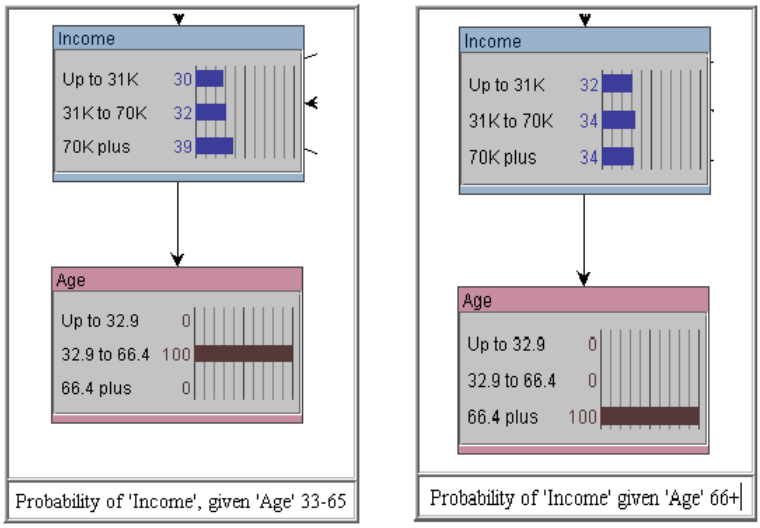
First, we examined relationships that were found in the data. Here is a screenshot of the model.

Figure 2: Base Model in Explorer



The arc (line) between the variables Age and Income indicated that knowing that a customer falls in a particular age group would influence the likely income of the customer. For example, if we knew that a customer had an age between 33 and 66, then there was a 39% chance that that the person earned over $70,000. If we knew that a customer was 66 or older, the likelihood decreased to a 34% chance that that person earned over $70,000. (See Figure 3 below)

Figure 3: Expected Relationship Between Age and Income

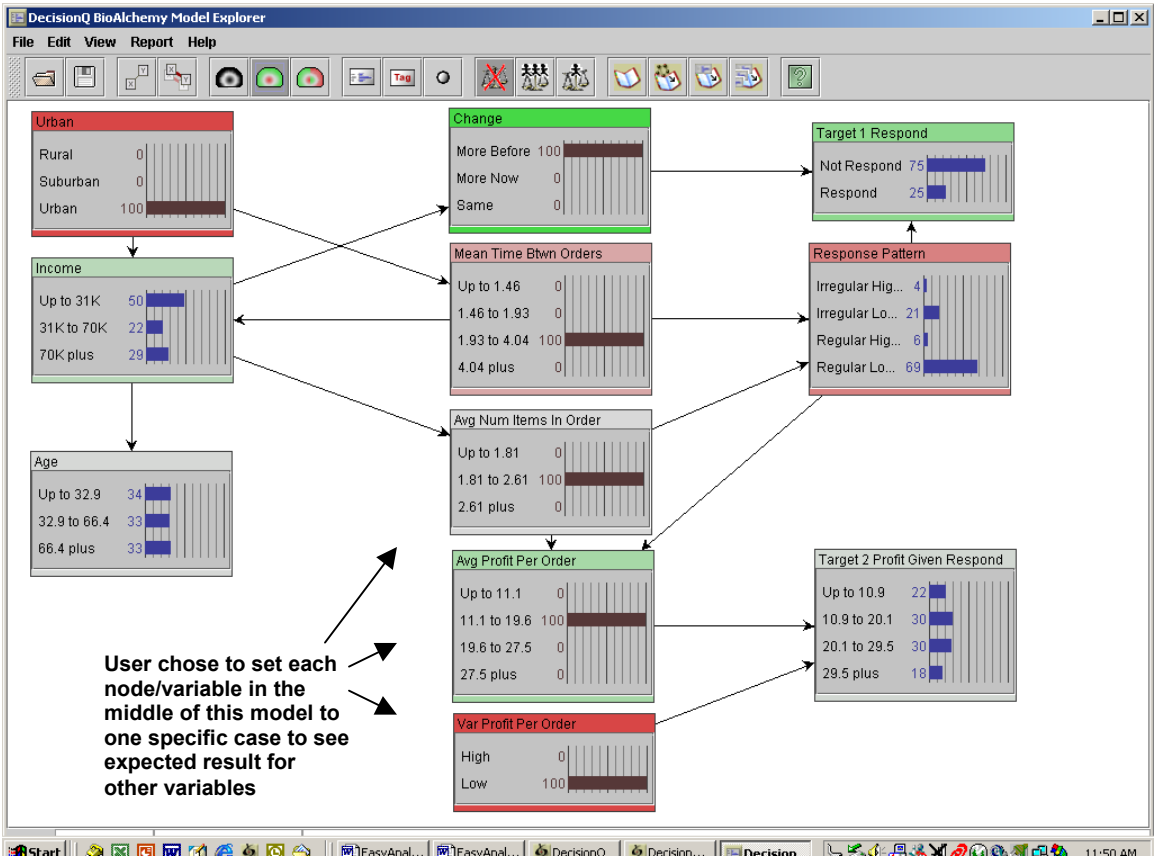


Thus, knowing age implied information about income. In a similar way, we were able to ascertain other relationships between variables in the model.

Second, we tested the model by entering in some customers to see how the model predicted outcomes. One example assumed a new customer with these characteristics: no transaction history, urban zip code, 42 years old. With this information entered into the model, the user could see that there is an inferred 42% chance that the person's income would be $70,000 or more. The model also inferred that there would be a 42% chance that the average time between orders would be less than 2 months. Ultimately, information flows through the model in a sensible manner and predicted that there was a 31% chance that the customer would respond to the campaign in question.

In a similar fashion, users could “toggle” the cases for other variables to understand the expected outcome of the rest of the network.

Figure 4



This interface allows a user to quickly try different cases to understand the flow of information that leads to the predictions of the target variables.

Quantitative Validation

It was clear from interacting with the model that the predictions made sense qualitatively. Before deploying the model, it was also important to understand how it would impact the profitability of the targeting campaigns.

Decision Analysis; Maximizing Expected Profit

The test dataset could also be used to understand how much profit would have been made if only customers with expected profit greater than zero were to have been mailed vs. the ‘mail to everyone’ mailing policy. For each customer the expected value of mailing for the campaign was computed as:

EV=p(responded)*[$1*p(btwnM5and5)+$10*p(btwn5and15)+$23*p(btwn15and30)+$45*p(above30)]-Mailing Cost

Customers were sorted by this value. Customers with a value below zero were identified as “do not mail to” customers assuming the current campaign was the last opportunity to reach the customer. This demonstrated the expected saving from mailing to only those members with a predicted expected value of responding that is greater than zero.

Optimal Targeting Policy

The above result could be extended to create an optimal policy over the entire year. For certain customers who might never purchase more than twice a year, it was clear that mailing material twelve times a year was excessive. For the bottom fraction of the positive expected value customers, it would make sense to target less frequently. DecisionQ then designed experiments to find the optimal targeting frequency for segments in the bottom fraction of the sorted list.

Results

While the initial results of the mailing strategy were impressive when compared to traditional modeling techniques, the real strength of the FasterAnalytics approach was in the fundamental sophistication of our methodology. Its ability to make use of subtle nuances in the data, to learn over time, to learn quickly, and to always be transparent for review by people at all levels of understanding allowed for higher degrees of campaign optimization by the sophisticated business user.

Our final proposed model was tested on historical data where New England Catalog had mailed to all prospective customers. The customers who had responded were kept secret until after the test. Decision Q's FasterAnalytics model was applied to the whole list and it predicted which customers it expected to respond. Results of the actual mailing were then compared with the recommendations of the model. Savings of mailing costs could be calculated, with the gross figure being reduced by any "don't mail" customers who actually did respond. The net figure showed a savings of approximately $400,000 per mailing or almost $5 million per year over the 'mail to everyone' strategy.

The use of sophisticated analytics enabled a more detailed understanding of customer behavior. FasterAnalytics created a model that showed an approach for increasing target efficiency while reducing cost.

If you have any further questions or would like to schedule a more detailed demonstration in person or over the web, please contact us.

DecisionQ Corporation
3726 Connecticut Ave. NW, Suite 519
Washington, DC 20008
www.decisionq.com
Phone: 415-254-7996
Fax : 415-276-6356
Email: info@decisionq.com