To: Prof. John Sparks

From: Mounica Sirineni

Date: April 27, 2015

Re: Model to Contact Sales Prospects for new heating system

This memo is in response to your request for an analysis regarding the following three questions:

1. Given our marketing budget of $25,000 which modeling technique should be used to maximize sales?
2. What is the overall performance of the models?
3. Provide a brief description of the modeling techniques.

By using Logistic Regression and contacting the prospects most likely to buy new heating system within the marketing budget of $25,000 we can contact 65 prospects who convert to customers and generate $33,000 in revenue. This represents an increase in revenue of $11,000 relative to using random selection to contact prospects. Overall logistic regression out-performs the other techniques used (Random Forest, MARS and CHAID) as seen on the gains chart on page 3. Descriptions of the four techniques are provided on page 3 and 4.

Analysis of Models for Marketing Budget

The constraints of the budget to maximize sales are as follows: The overall budget is $25,000 and the cost per contact is $30. Therefore we can contact 833 prospects ($25,000 /$30=833). An additional assumption of this analysis is that each sale of new heating system will yield $500 in revenue.

Within the constraint of the marketing budget and given the revenue assumption, which modeling technique should we use to contact the best sales prospects and generate the maximum amount of revenue? The table below summarizes the performance of the four modeling techniques vs. random selection. An analysis follows that also explains how the figures in the table were calculated.

Table 1: Revenue comparison (Modeling techniques vs Random selection)

Modeling Technique	**Revenue at cut-point**	**Increase in Revenue vs Random Selection**
Random Selection (Baseline)	$22,000	
Logistic Regression	$33,000	$ 11,000
Random Forest	$32,000	$ 10,000
MARS	$28,500	$ 6,500
CHAID	$26,500	$ 4,500

Table 2: Revenue comparison (Modeling techniques vs Random selection)

Modeling Technique	**Revenue at cut-point**	**Increase in Revenue vs Random Selection**
Random Selection (Baseline)	$22,000	
Logistic Regression	$33,000	$ 11,000
Random Forest	$32,000	$ 10,000
MARS	$28,500	$ 6,500
CHAID	$26,500	$ 4,500

Random Selection (or Baseline) If we select sales prospects for new heating system randomly, then we would expect that a random selection of 10% of all prospects to contain 10% of sales; 20% of randomly selected prospects would account for 20% of sales, etc. The marketing budget allows us to contact 50% of our file. There were 89 total sales on the file. 50% of 89 is 44 total sales prospects with an associated revenue of $22,000 (44 X $500).

Logistic Regression By comparison, using logistic regression we reach 66 sales prospects among the 833 prospects for a maximum revenue of $33,000 (66 X $500). This is an increase in revenue of $11,000 relative to using random selection ($33,000 - $22,000).

Random Forest If Random Forest is used, we would earn $1,000 less in revenue as compared to logistic regression. Specifically, this technique involves converting 64 prospects to customers for a revenue of $32,000 (64 X $500).

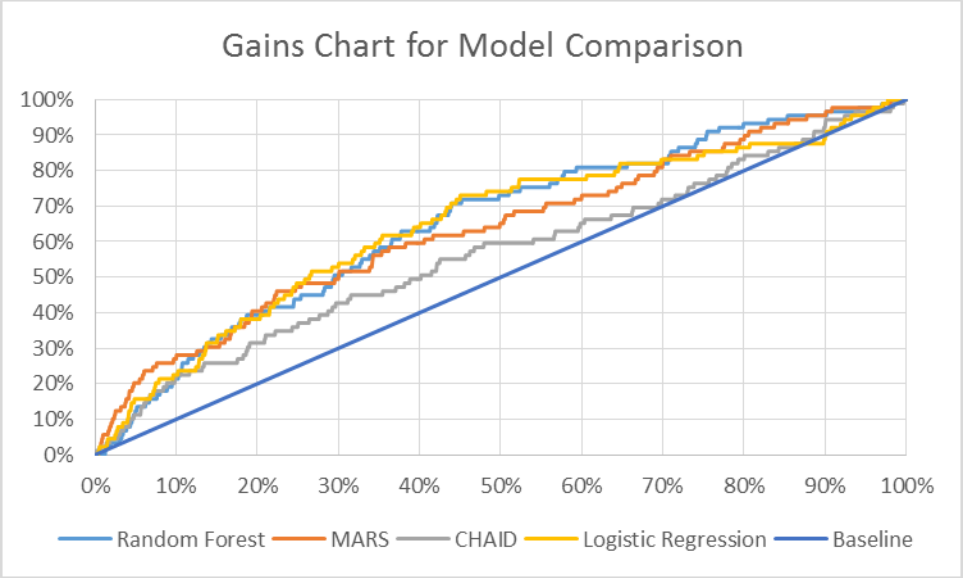
MARS If MARS is used, we would earn $4,500 less in revenue as compared to logistic regression. Specifically, this technique involves converting 57 prospects to customers for a revenue of $28,500 (57 X $500).

CHAID If CHAID is used, the revenue earned would decrease by $6,500 as compared to logistic regression. Specifically, this technique involves converting 53 prospects to customers for a revenue of $26,500 (53 X $500).

Overall Performance Comparison

The gains chart shown on page 3 provides a method to summarize the overall performance of the four models. Generally, a gains chart shows a comparison of the four models versus the baseline. The gap between the prediction line for the four models and the baseline indicates the gains that the company foresees from using the respective predictive models to prioritize its mailing list.

Figure 1: Gains Chart for Baseline vs Prediction Models



The lift curve for Logistic Regression is above the curve for Random Forest, MARS and CHAID for the range of 20% to 70% of total prospects. The lift curve for MARS is above the curve of the other models for the range spanning 0% to 15% of total prospects. By contrast, the lift curve for CHAID is below the curve of Logistic Regression, Random Forest and MARS for the entire range of prospects. Therefore, Logistic Regression generally has a stronger performance in predicting sales, except at the left-most and right-most part of the curve.

Modeling Techniques

The four techniques used in this analysis were Logistic Regression, Random Forest, CHAID and MARS. A description of each is provided below.

Logistic Regression: Logistic regression is a regression technique that measures the relationship between the binary dependent variable and independent variables. In this technique the log odds of the outcome is modeled as a linear combination of the predictor variables.

MARS: MARS is a regression technique with knots. It recognizes turning points that exist in the linear relationship. MARS technique does not assume any particular type or class of relationship (e.g., linear, logistic, etc.) between the predictor variables and the dependent variable.

CHAID: CHAID involves recursive partitioning of data into groups. In this method, the chi-square statistic is calculated between the dependent variable and each independent variable. The independent variable with the largest chi-square statistic is selected and the data are split into groups based on that independent variable. Each of these sub-groups is then examined and the process is repeated.

Random Forest: Random Forest is an ensemble of trees. Random forest technique is a modified tree learning algorithm that randomly removes small number of variables and builds the best tree without

those variables. It uses the average of the predicted rates for all the different trees to find the best estimate for the probability of response for each prospect.

Summary

The results of this analysis has shown that the maximum sales can be reached using Logistic Regression as opposed to Random Forest, CHAID and MARS and that logistic regression clearly out-performs a random selection of prospects. The revenue from using Logistic Regression for the sales model for new heating system is $33,000. Also, Logistic Regression has superior performance relative to Random Forest, CHAID and MARS across most of the range of percentage of total prospects contacted. A description of the four techniques was provided on page 3 and 4.

Mounica Sirineni