To: Prof. John Sparks
From: 658795799
Date: 03/30/2015
Re: Purchased at least $500 in home repairs from Mr. Handyman

This memo is in response to your request to build a model for converting prospects to purchase at least $500 in home repairs from Mr. Handyman.

Specifically, you requested a model and analysis that would address four questions:

1. Can we build a model to cut our mailing quantities by 25% and still get most of our responses
2. What are the variables in the model
3. Which ones are the most impactful and how do they impact the prediction of response
4. If I want to cut my mail quantity by a different percent, what would you suggest and what is the effect on the proportion of responses

Results show that the top 25% of prospects contain 83% of the total purchases. The variables used in the model are listed in Table1 at the bottom of this page. The most impactful variables were: whether the prospect was a member of the Young urban have-nots segment and number of motorcycle/scooter policies the prospect purchased. An alternative cut-point exists at 39% of the prospect universe which contains 57% of the purchases. This cut-point was identified using lift analysis as discussed on Page 3.

Cutting mail quantities by 25%. Regression modeling was used to identify the characteristics that were most strongly associated with purchase of at least $500 in home repairs from Mr. Handyman. The results of that modeling showed that the top 75% of prospects contained 83% of the total purchases. By abstaining from contacting the bottom one-quarter of the prospect file you can increase your response rate from 6.1% to 6.8%.

Variables in the Model. The variables in the model are shown below in Table1.

Table1: Variables in the model

Variable	Change in probability of responses
Prospects belonging to the segment of Young urban have-nots	7.565%
Number of motorcycle/scooter policies the prospect purchased	1.428%
Percentage of prospects who are Protestant in the neighborhood	-0.218%
Percentage of prospects who are married in the neighborhood	0.092%
Percentage of prospects who belong to Social class B1 in the neighborhood	-0.088%
Percentage of prospects who have Household with children in the neighborhood	-0.084%

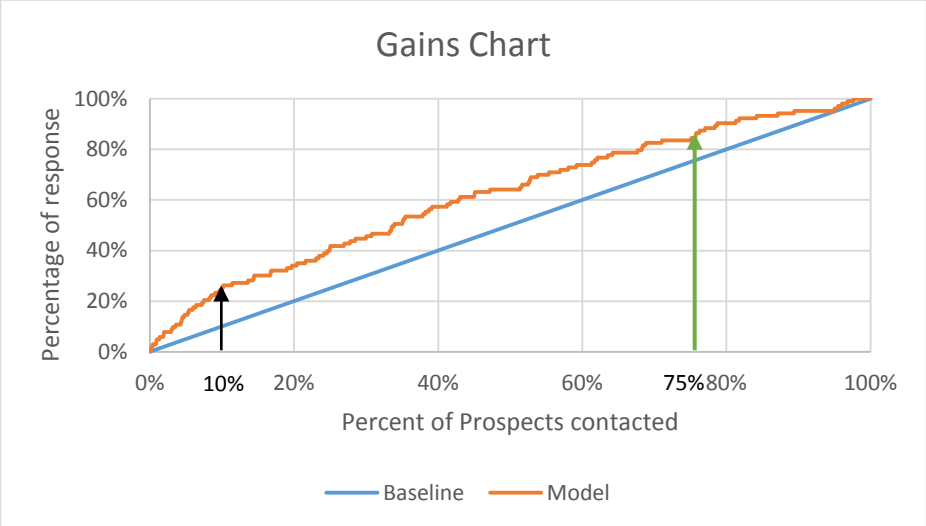
Variables in the Model and Impact. The variables used in the model for home repairs are shown in the table below. You can see that the slope for the prospects belonging to the segment of Young urban have-nots was 7.565%, meaning that if the prospect was a member of Young urban have-nots, then the probability of purchase increases by 7.565%. Similarly for the number of motorcycle/scooter policies the slope was 1.428%, meaning that an increase of 1 motorcycle/scooter policy held by the prospect increases the probability of response by 1.428%. For every additional percent of Household with children in the prospect's neighborhood the probability of purchase decreases by 0.084%. The slopes for all the variables are shown in the table.

Table1: Repeated for convenience

Variable	Change in probability of responses
Prospects belonging to the segment of Young urban have-nots	7.565%
Number of motorcycle/scooter policies the prospect purchased	1.428%
Percentage of prospects who are Protestant in the neighborhood	-0.218%
Percentage of prospects who are married in the neighborhood	0.092%
Percentage of prospects who belong to Social class B1 in the neighborhood	-0.088%
Percentage of prospects who have Household with children in the neighborhood	-0.084%

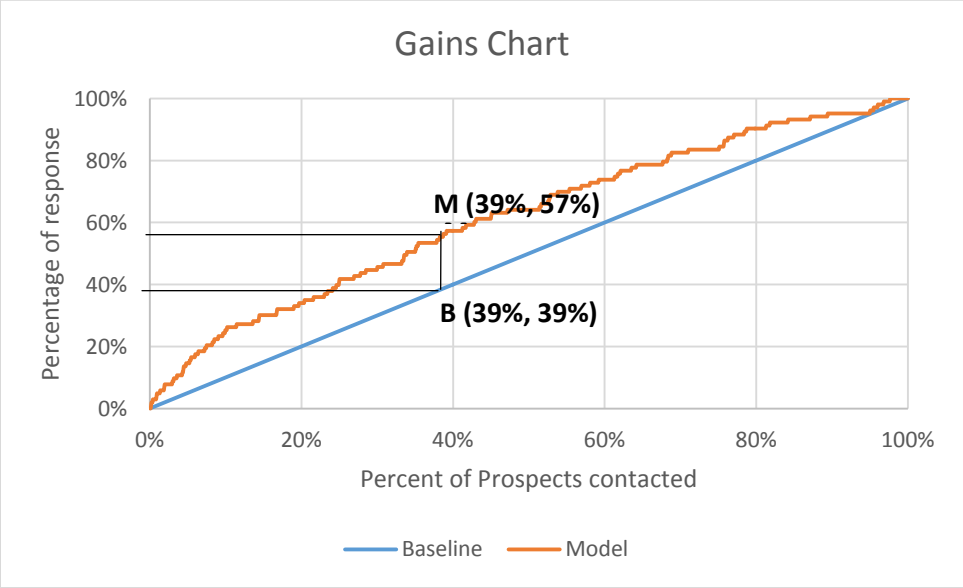
Gains Chart. The Gains Chart from the model is shown below. The blue line labeled 'Baseline' shows the percent of prospects purchased at least $500 home repairs from Mr. Handyman, if prospects were selected on a random basis. That is, we would expect that a random selection of 10% of all prospects to contain 10% of purchases; 20% of randomly selected prospects would account for 20% of purchases, etc. The line labeled as 'Model' shows analogous results if the model is used to select prospects. You can see that the top 10% of prospects account for 24% of all purchases using the model. The spot on the gains chart marked with an arrow in green color shows that the top 75% of prospects accounts for 83% prospects of all purchases of at least $500 in home repairs from Mr. Handyman.

Figure1: Gains Chart



Alternative Cut-Point. Alternative cut-point is determined using Lift analysis. Lift is the difference between model value and baseline value. This gives the increase in the percentage of prospects who have purchased at least $500 in home repairs compared to the baseline. The alternative cut-point exists at the point where we find the maximum difference between model and baseline values. In the figure below, cut point of the baseline is represented as B and cut-point of the model is represented as M. The difference between M and B gives the lift%. The maximum lift occurs at 39% of the mail file which contains 57% of the sales. Therefore, 39% was selected as alternative cut point.

Figure2: Gains Chart with representation of Lift



In summary, we will have 83% of prospects who purchased at least $500 in home repairs from Mr. Handyman by using the model and cutting down the mail quantities by 25%. The most impactful variables identified by regression model were: whether the prospect was a member of the Young urban have-nots segment and number of motorcycle/scooter policies the prospect purchased. An alternative cut point was determined using Lift analysis. Results show that alternative cut-point exists at 39% of the mail file which contains 57% of the sales.

Technical Appendix

This technical appendix provides details as to how the data was prepared for modeling and the construction of the model itself.

Logistic Regression

A logistic regression model was built to identify the characteristics of prospects more likely to purchase home repairs from Mr. Handyman. The results of that model are shown below. Details regarding the steps preceding the actual model construction follow.

Table2: Output for Logistic Regression

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.6456	0.1267	435.9710	<0.0001
mostyp_21	1	0.8041	0.2164	13.8047	0.0002
MRELGE2	1	0.0134	0.0035	14.8531	0.0001
MFWEKI2	1	-0.0252	0.0102	6.0423	0.0140
MSKB12	1	-0.0205	0.0049	17.2402	<0.0001
MGODPR2	1	-0.0420	0.0197	4.5282	0.0333
AMOTSC2	1	0.1108	0.0211	27.4847	<0.0001

In Table2, the p-value of the variable is represented in the column 'Pr > ChiSq'. Coefficients, Standard Error and Chi-Square values are also listed in the table. Variables are said to be statistically significant when the p-value is less than 0.05 significance level (alpha value). All the variables listed in 'Parameter' column are statistically significant as p-value is less than 0.05.

Data Preparation and Variable Selection

The initial file contained 5,399 rows and 28 variables. A number of these variables required adjustment prior to building the model.

Ordinal Variables. The file contained 15 geo-demographic ordinal variables. That is, a particular value for one of these variables represented a range of percentage of people of a certain type in the prospect's neighborhood. In order to use these variables in a linear model the original values were re-scaled to the mean value of the percentage range. Similarly, the file contained 3 variables regarding the

amount a prospect spent on certain products. These were similarly transformed as shown in the table below.

Table3: Ordinal variables conversion

L3: Geo-Demographic Ordinal Variables			L4: Spend variables		
Variable Value	**Original Value**	**New Value**	**Variable Value**	**Original Value**	**New Value**
0	0%	0	0	f 0	0
1	1-10 %	5.5	1	f 1-49	25
2	11-23 %	17	2	f 50-99	75
3	24-36 %	30	3	f 100-199	150
4	37-49 %	43	4	F 200-499	350
5	50-62 %	56	5	F 500-999	750
6	63-75 %	69	6	F 1,000-4,999	3,000
7	76-88 %	82	7	F 5,000-9,999	7,500
8	89-99 %	94	8	F 10,000-19,999	15,000
9	100%	100	9	F 20,000 - ?	30,000

For Example, if the value of the geo-demographical variable (L3) lies between 1% and 10%, then the value of that variable was changed to 5.5%.

Categorical Variables. Two categorical variables were included in the original file. These were MOSTYP and MOSHOO variables. These variables were converted into binary format so that they could be used in the model. The MOSHOO variable had 10 different values in the input file. Therefore the MOSHOO variable was converted into MOSHOO1 - MOSHOO10 binary variables. For example, if the value the MOSHOO variable is 5, then value of 1 was assigned to the MOSHOO5 variable and the remaining variables were assigned 0. Similarly, MOSTYP was converted to MOSTYP1 - MOSTYP41, as it had 41 different values in the input file.

Holdout Sample. The input dataset was divided into 70:30 proportion. 70% of the dataset was used as an analysis sample which was used to build regression model. 30% of the dataset was used as holdout sample. The main reason of the holdout sample was to implement the model on the sample that was not used during building the regression model. In the business memo, performance metrics of the holdout sample were used to observe the percentage of prospects purchased home repairs from Mr. Handyman either by contacting all prospects or a certain percent of prospects.

Non-Linear Relationships. The following graph shows the response rate for selected quantitative variables. The relationship between quantitative variables and the response variable were predicted as binary, linear or quadratic by observing the pattern generated on the graph. For certain variables that appeared to have a non-linear relationship with response, we built quadratic and linear models using logistic regression and compared their SC value. For example, Graph1 shows the pattern of response rate by the variable MAANTH. The pattern is potentially quadratic. Individual logistic regression models for a linear vs. quadratic relationship with response were calculated and the SC values examined, summarized in the table below.

Graph1: Plot for variable MAANTH

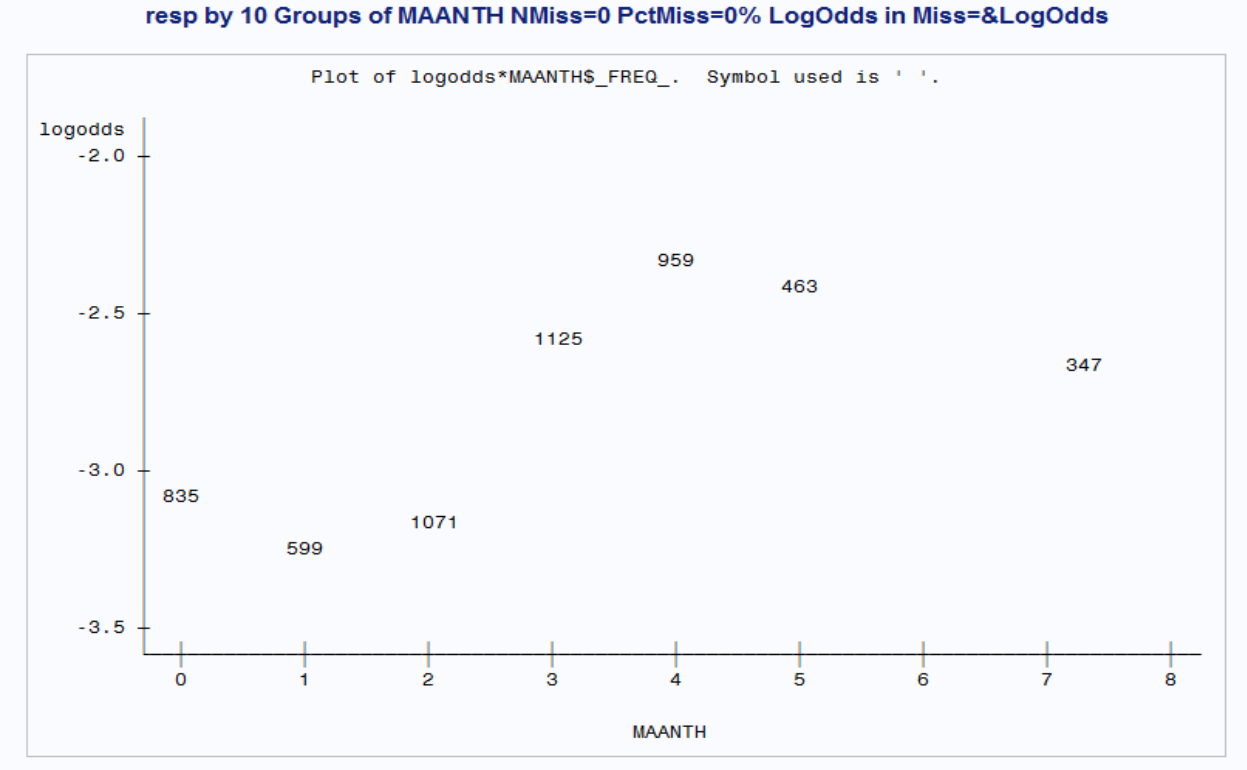


Table4: SC values for variable MAANTH

Variable	SC	
	Linear	Quadratic
MAANTH	2,447	2,445

The SC for the quadratic model was 2,445 vs. 2,447 for the linear for the variable MAANTH and therefore the quadratic form was used as a candidate independent variable for logistic regression.

Similarly, for variable MHHUUR2, logistic regression model for a linear vs. quadratic relationship with response were calculated and the SC values examined, summarized in the table below.

Graph2: Plot for variable MHHUUR2

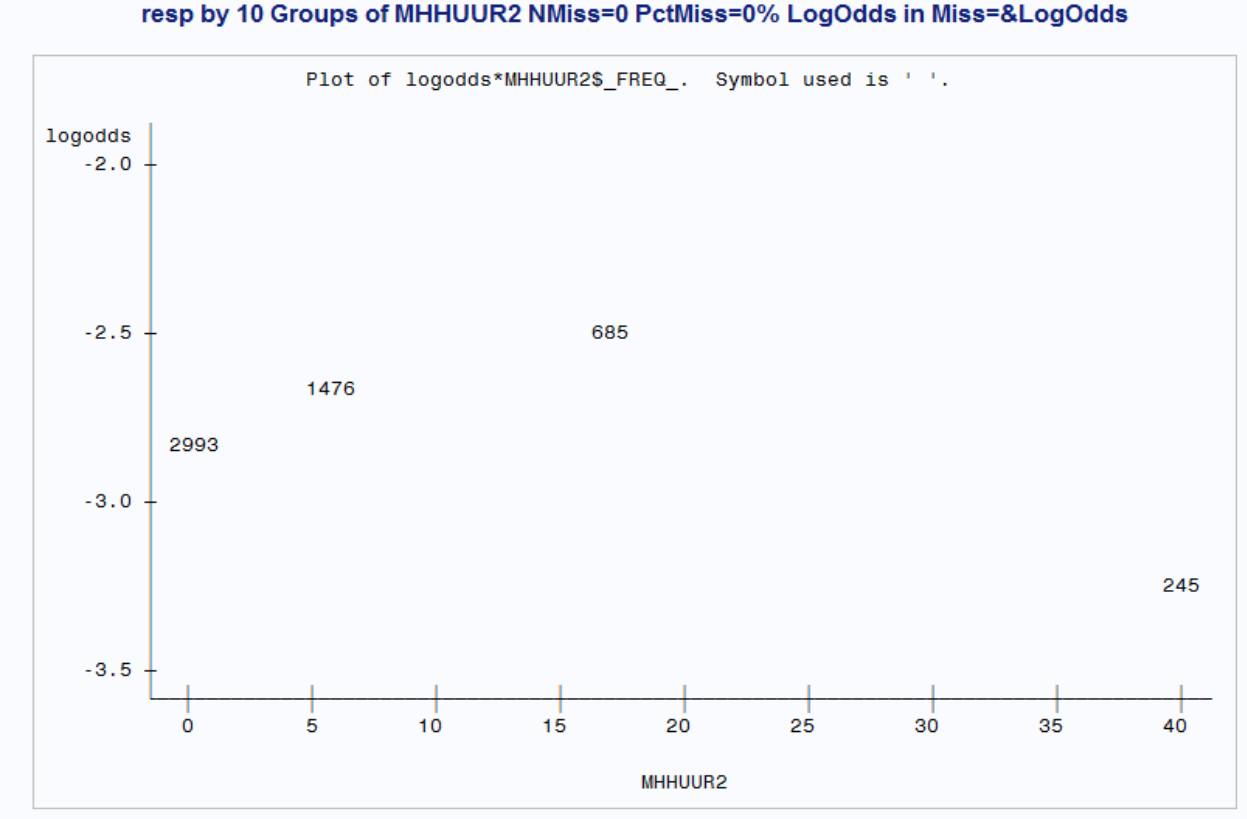


Table5: SC values for variable MHHUUR2

Variable	SC values	
	Linear	Quadratic
MHHUUR2	2,462	2,466

The SC for the linear model was 2,462 vs. 2,466 for the quadratic for the variable MHHUUR2 and therefore the linear form was used as a candidate independent variable for logistic regression.

Logistic Regression Model. After preparing all the variables for potential use in the model, 73 variables were submitted to a logistic regression model using stepwise variable selection. This resulted in 6 statistically significant variables as shown in Table7

Table7: Output for Logistic Regression

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.6456	0.1267	435.9710	<0.0001
mostyp_21	1	0.8041	0.2164	13.8047	0.0002
MRELGE2	1	0.0134	0.0035	14.8531	0.0001
MFWEKI2	1	-0.0252	0.0102	6.0423	0.0140
MSKB12	1	-0.0205	0.0049	17.2402	<0.0001
MGODPR2	1	-0.0420	0.0197	4.5282	0.0333
AMOTSC2	1	0.1108	0.0211	27.4847	<0.0001

Linear Regression. After using logistic regression to select the variables for the model, the final set of independent variables was used to build a linear regression equation. This was done in order to express the impact of the selected variables in terms of probability of response. The parameter estimate values for logistic regression would give the change in the log of odds of response which will be difficult to interpret compared with linear regression. Therefore for explanation linear regression was preferred over logistic regression. The coefficients from linear regression model were used to observe the change in the response variable with the increase in one unit of the selected variable. Results of the linear regression used in the business memo are shown in the below table.

Table8: Output for Linear Regression

Variable	**Change in probability of responses**
Prospects belonging to the segment of Young urban have-nots	7.565%
Number of motorcycle/scooter policies the prospect purchased	1.428%
Percentage of prospects who are Protestant in the neighborhood	-0.218%
Percentage of prospects who are married in the neighborhood	0.092%
Percentage of prospects who belong to Social class B1 in the neighborhood	-0.088%
Percentage of prospects who have Household with children in the neighborhood	-0.084%

- Sai Lahari Jalaparthi