To: The Client

From: 661073066, 675930999, and 658972668.

Date: 11th April, 2017

Re: Predictive model for finding prospective buyers of John Deere riding mower.

This memo is in response to your request to build a model for converting prospects for John Deere riding mowers into customers.

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

1. Can we build a model to cut our prospects by 25% and still get most of our purchases?
2. What are the variables in the model?
3. Which ones are the most impactful and how do they impact the prediction of purchases?
4. What is an alternative cut-point for the prospect universe, and what is the effect on the proportion of purchases of riding mower?

Results show that cutting down 25% of the prospects will still get 85% of the total purchases. The variables used in the model are listed in Table 1 at the bottom of this page. The most impactful variables were: the number of private third party insurance policies the prospect owns, if the customer belonged to the subtype ‘Suburban Youth’, if the main type of customers is ‘successful hedonists’ and if the customer belonged to the subtype ‘Large Family, employed child’. An alternative cut-point exists at 54% of the prospect universe which contains 72% of the sales. This cut-point was identified using lift analysis as discussed on page 3.

**Top 75% of Prospects.**

Regression modeling was used to identify the characteristics that were most strongly associated with the purchase of John Deere riding mower**.** The results of that modeling showed that the top 75% of prospects contained 85% of the total sales. By abstaining from contacting the bottom quarter of the prospect file you can increase your response rate from 5.0% to 6.7%.

**Variables in the Model.**

The variables in the model are shown below in Table 1.

**Table 1**

|  |  |
| --- | --- |
| Description | Change in probability of response |
| Number of private third party insurance policies owned by prospect | 45% |
| Prospects in category ‘Suburban youth’ | 6% |
| Customers in category ‘successful hedonists’ | 5% |
| Prospects in category ‘Large family, employed child’ | 4% |
| Percent of people in the prospects neighborhood who own 2 cars | 0.1% |
| Average size of household of the prospect | 0.076% |
| Percent of people in prospects neighborhood who belong to social class C | 0.042% |
| Percent of protestants in the prospects neighborhood | 0.041% |

**Variables in the Model and Impact.**

The variables used in the model forJohn Deere riding mower prospect purchases are shown in the Table 2 below. The slope for the Number of private third party insurance owned by prospect was 45%, meaning that for every additional private third party insurance policy owned by the prospect, the probability of purchase increased by 45%, analogous to the variable ‘Average size of household’. The slope if the prospect belongs to the category ‘Suburban youth’ was 6%, meaning if prospect belongs to category ‘Suburban youth’ the probability of purchase increased by 6% analogous to the variables ‘successful hedonists’ and ‘Large Family, employed child’. The slopes for all the variables are shown below in the Table 2.

**Table 2**

|  |  |
| --- | --- |
| Description | Change in probability of response |
| Number of private third party insurance policies owned by prospect | 45% |
| Prospects in category ‘Suburban youth’ | 6% |
| Customers in category ‘successful hedonists’ | 5% |
| Prospects in category ‘Large family, employed child’ | 4% |
| Percent of people in the prospects neighborhood who own 2 cars | 0.1% |
| Average size of household of the prospect | 0.076% |
| Percent of people in prospects neighborhood who belong to social class C | 0.042% |
| Percent of protestants in the prospects neighborhood | 0.041% |

**Gains Chart.**

The Gains Chart from the model is shown in below Graph 1. The blue line labeled ‘Baseline’ shows the percent of John Deere riding mowersales when the 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 sales; 20% of randomly selected prospects would account for 20% of sales, and so on.

The line labeled as ‘Model’ shows analogous results when prospects were selected based on the model. You can see that the top 10% of prospects account for 18% of all sales using the model. The spot on the gains chart marked with an arrow shows that the top 75% of prospects account for 85% of all sales using the model for purchaser of John Deere riding mower.

**Graph 1**

**Alternative Cut-Point**

An alternative cut point has been identified by calculating the maximum lift. Lift can be defined as the difference in the response if prospects are selected using the model as opposed to selecting them on a random basis. The point with the maximum lift has been identified as an alternative cut-point, which is 54% of the prospects where we would get 72% of purchases.

In summary, we can conclude that using the model to cut our mails by 25% would still get 85% of our total purchases. The variables used in the model are shown in Table 3. The most impactful variables were: Number of private third party insurance owned by prospect, if the prospect belonged to the category ‘suburban youth’, if the customer belongs to category ‘successful hedonists’ and if the prospect belonged to category ‘Large family, employed child’. An alternate cut-point of the mails in prospect universe is identified at 54% which would get 72% of the total purchases.

**Table 3**

|  |  |
| --- | --- |
| Description | Change in probability of response |
| Number of private third party insurance policies owned by prospect | 45% |
| Prospects in category ‘Suburban youth’ | 6% |
| Customers in category ‘successful hedonists’ | 5% |
| Prospects in category ‘Large family, employed child’ | 4% |
| Percent of people in the prospects neighborhood who own 2 cars | 0.1% |
| Average size of household of the prospect | 0.076% |
| Percent of people in prospects neighborhood who belong to social class C | 0.042% |
| Percent of protestants in the prospects neighborhood | 0.041% |

**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 John Deere riding mower. The results of that model are shown below in Table 4. Details regarding the steps preceding the actual model construction follow.

**Table 4**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| Intercept | 1 | -4.52 | 0.37 | 152.15 | <.0001 |
| mostyp\_19 | 1 | 0.63 | 0.21 | 8.61 | 0.0033 |
| mostyp\_34 | 1 | -1.49 | 0.72 | 4.23 | 0.0399 |
| MOSHOO\_1 | 1 | -1.77 | 0.75 | 5.62 | 0.0178 |
| MAUT2\_2 | 1 | 0.02 | 0.00 | 15.56 | <.0001 |
| MGEMOM2 | 1 | 0.02 | 0.00 | 18.15 | <.0001 |
| MSKC2 | 1 | 0.01 | 0.00 | 5.09 | 0.0240 |
| MGODPR2 | 1 | 0.01 | 0.00 | 13.42 | 0.0002 |
| AWAPAR | 1 | 3.60 | 1.71 | 4.43 | 0.0353 |

The Table 4 above shows that all the variables included in the model are statistically significant based on chi-square test. P values of all the variables included in the model are less than α of 0.05 which is one of the commonly accepted value of statistical significance. Mostyp\_19, which signifies the prospects in the category ‘Suburban youth’, has a p-value of 0.0033 which is less than 0.005 so, this variable is statistically significant. Similarly, mostyp\_34, which signifies the prospects in the category ‘successful hedonists’, has a p-value of 0.0399 which is less than the α-value, therefore, mostype\_34 is statistically significant.

**Data Preparation and Variable Selection**

The initial file contained 5,392 rows and 28 variables. A number of these variables required adjustment prior to building the model.

Ordinal Variables. The file contained 17 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 middle of the percentage range. Similarly, the file contained 3 variables representing a range of the amount a prospect spent on certain products. These were similarly transformed as shown in the Table 5 below.

**Table 5**

L3: Geo Demographic Format L4: Spend Format

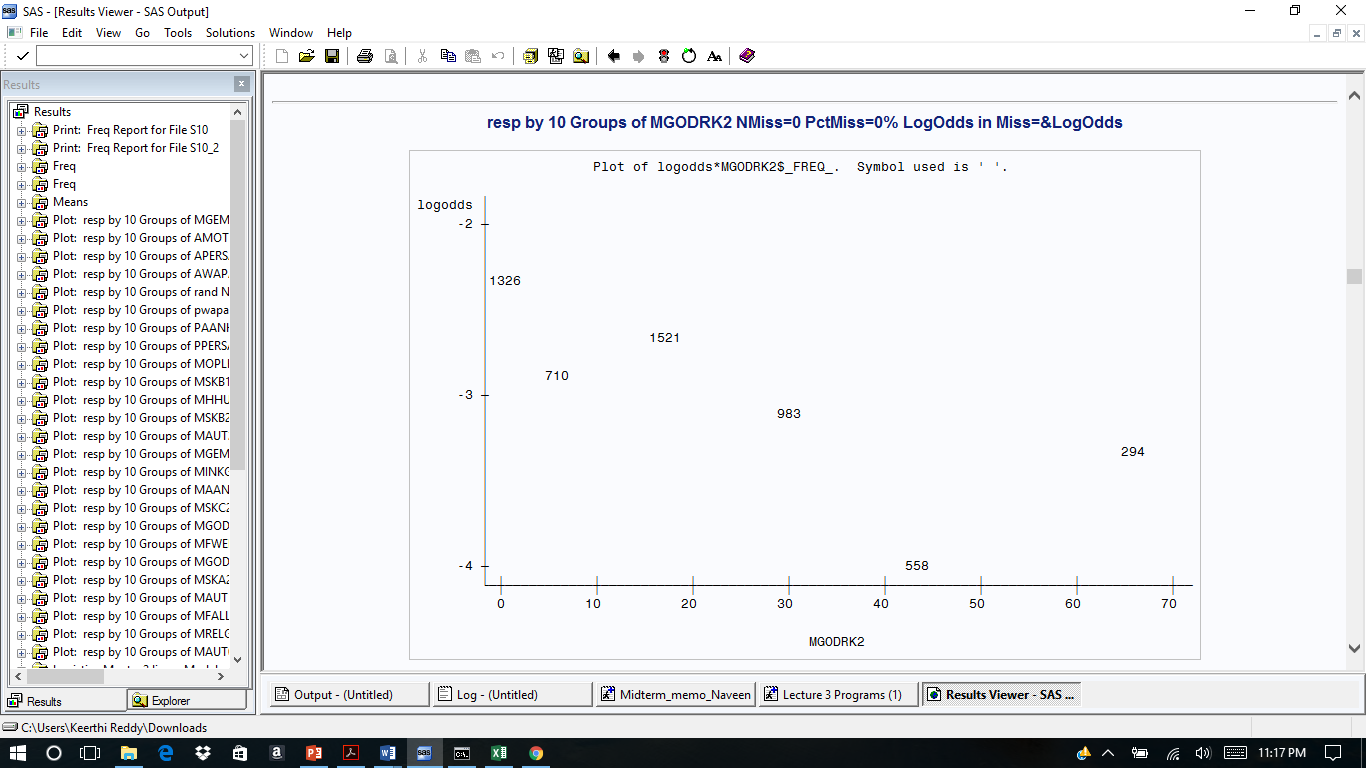
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Value | Original Value | New Value | Variable Value | Original Value | New Value |
| 0% | 0 | 0 | 0 | 0 | 0 |
| 1-10% | 1 | 5.5 | 1-49 | 1 | 25 |
| 11-23% | 2 | 17 | 50-99 | 2 | 75 |
| 24-36% | 3 | 30 | 100-199 | 3 | 150 |
| 37-49% | 4 | 43 | 200-499 | 4 | 350 |
| 50-62% | 5 | 56 | 500-999 | 5 | 750 |
| 63-75% | 6 | 69 | 1,000-4,999 | 6 | 3,000 |
| 76-88% | 7 | 82 | 5,000-9,999 | 7 | 7,500 |
| 89-99% | 8 | 94 | 10,000-19,999 | 8 | 15,000 |
| 100% | 9 | 100 | 20,000-40,000 | 9 | 30,000 |

Categorical Variables. Two categorical variables were included in the original file. These were converted into a set of binary variables so that they could be used in the model. For example, in records where Moshoo was equal to 1, a new variable Moshoo\_1 is created, which equals 1 if the prospect is a member of the “Successful hedonists” segment and 0 if he is not. Another variable, Moshoo\_2, equals 1 if the prospect is a member of the “Driven Growers” segment and 0 If not. This transformation is done for all the possible values of Moshoo and Mostyp.

Holdout Sample. Out of the total data, 30% of the data was kept aside as a holdout sample. This holdout sample consisted of about 1663 records and was created by using a random number generator function. If the random number was greater than 0.7, the values from ‘resp’ were placed in ‘respholdout’ and ‘resp’ was replaced by ‘.’ Logistic regression was then performed on the remaining 70% of the data, then the trained Logistic regression model was used to score the response variable of holdout sample. The holdout sample was then sorted in descending order of the response variable to get most likely prospect of responding on the top of the list.

Non-Linear Relationships. Graphs showing the response rate for each of the quantitative variables were examined and those patterns that appeared to have a non-linear relationship with response were tested with the SC value to determine if quadratic is a better fit. For example, the Graph 2 as below shows the pattern of response rate by MGODRK2.

**Graph 2**



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.

Below Table 6 is, SC values tables for MGODRK2,

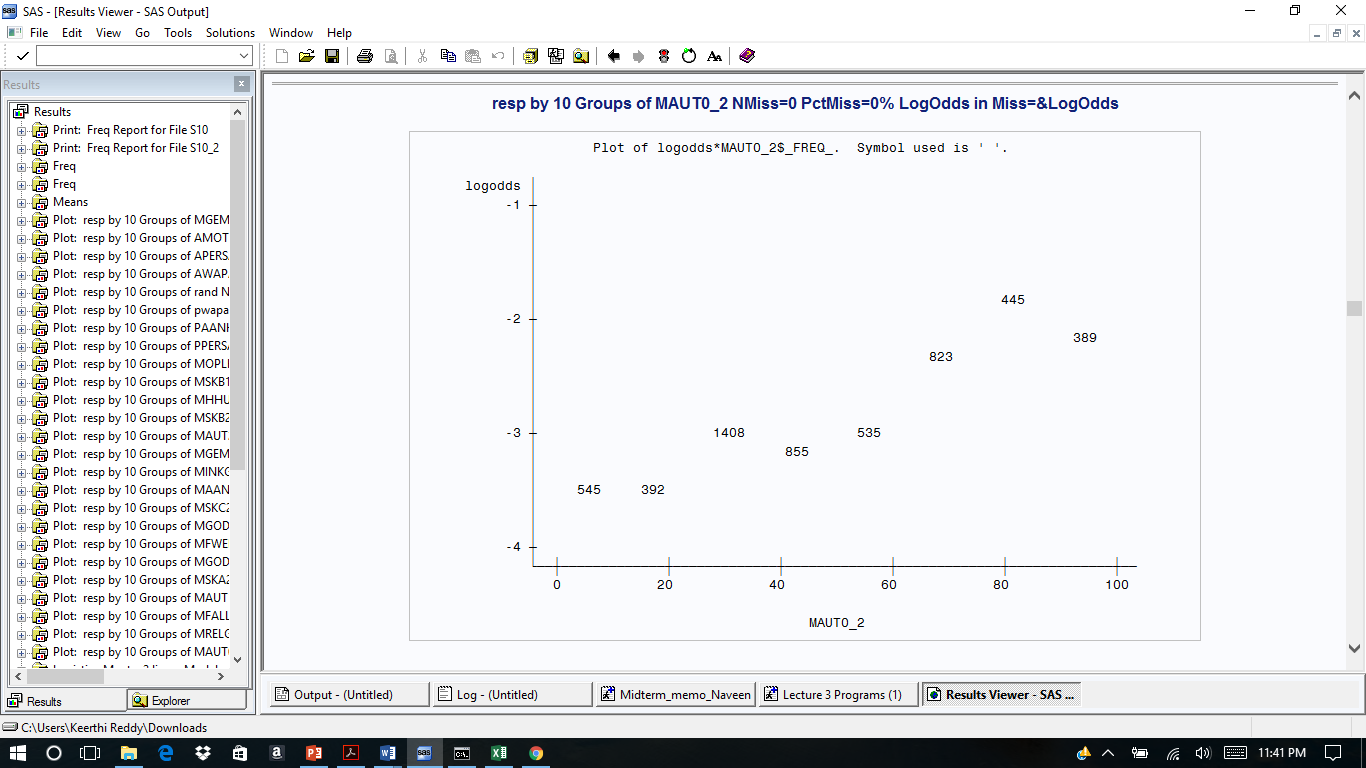
**Table 6**

|  |  |
| --- | --- |
|  | MGODRK2 |
| Linear | 1,644 |
| Quadratic | 1,652 |

The SC for the linear model was 1,644 vs. 1,652 for quadratic and therefore the linear form was used as a candidate independent variable for logistic regression.

Similar test was done on the MAUT0\_2 variable to find out the pattern. Graph 3 shows relationship of response rate for MAUT0\_2.

**Graph 3**



**Table 7**

|  |  |
| --- | --- |
|  | MAUT0\_2 |
| Linear | 1,693 |
| Quadratic | 1,701 |

From SC value the variable MAUT0\_2 as shown in Table 7, is found out to be having a linear form, hence a linear candidate independent variable is used for logistic regression.

Logistic Regression Model. After preparing all the variables for potential use in the model, all 28 variables were submitted to a logistic regression model using stepwise variable selection. This resulted in 8 statistically significant variables as shown in Table 8.

**Table 8**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| Intercept | 1 | -4.52 | 0.37 | 152.14 | <.0001 |
| mostyp\_19 | 1 | 0.63 | 0.21 | 8.61 | 0.0033 |
| mostyp\_34 | 1 | -1.49 | 0.72 | 4.22 | 0.0399 |
| MOSHOO\_1 | 1 | -1.77 | 0.75 | 5.62 | 0.0178 |
| MAUT2\_2 | 1 | 0.02 | 0.00 | 15.56 | <.0001 |
| MGEMOM2 | 1 | 0.02 | 0.00 | 18.15 | <.0001 |
| MSKC2 | 1 | 0.01 | 0.00 | 5.09 | 0.0240 |
| MGODPR2 | 1 | 0.01 | 0.00 | 13.42 | 0.0002 |
| AWAPAR | 1 | 3.59 | 1.70 | 4.43 | 0.0353 |

Linear Regression. After using logistic regression to select the variables for the model, the final set of independent variables were used to build a linear regression equation. This was done to explain the importance of a variable in a business setting, as explaining the importance of a variable using log-odds is complicated. The coefficients in this model represent the change in probability of response for one unit increase in the independent variables. For categorical variables, these coefficients represent the change in probability of response if a prospect belongs to that category. The results of linear regression are shown below in Table 9.

|  |  |
| --- | --- |
| **Variable** | **Parameter**  **Estimate** |
| AWAPAR | 0.45 |
| MOSHOO\_1 | -0.05 |
| MOSTYP\_19 | 0.07 |
| MOSTYP\_34 | -0.04 |
| MAUT2\_2 | 0.00 |
| MGEMOM2 | 0.00 |
| MSKC2 | 0.00 |
| MGODPR2 | 0.00 |

**Table 9**