To: The Client

From: 679850219, 672039068, 654324917

Date: April 3rd, 2017

Re: Blue Cross – Blue Shield health maintenance organization (HMO) model

This memo is in response to your request to build a model for converting prospects to sign up for Blue Cross – Blue Shield health maintenance organization.

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?

Yes, we can build a model to cut our mailing quantities by 25%, in the model discussed below, the top 75% of prospects contain 92% of the total sales. The variables used in the model are listed in table 1 at the bottom of this page. An alternative cut-point exists at 70% of the prospect universe which contains 91% of the total 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 signing up for the Blue Cross – Blue Shield health maintenance organization (HMO). The results of that modeling showed that the top 75% of prospects contained 92% of the total sales. By abstaining from contacting the bottom 25% of the prospect file you can increase your response rate from 5.94% to 7.28%.

**Variables in the Model.**

The variables used in the model for Blue Cross – Blue Shield health maintenance organization (HMO) and their corresponding slope estimates in descending order of magnitude are shown in the table below.

TABLE 1. Variables used in the model and their absolute slope estimates

|  |  |
| --- | --- |
| VARIABLE DESCRIPTION | ABSOLUTE SLOPE ESTIMATE (%) |
| Affluent Young Families Customer segment | 6.73 |
| Student in Apartments Customer segment | 6.579 |
| Dinki's (Double Income No Kids) Customer segment | 5.822 |
| Mixed Rural Customer segment | 5.233 |
| Large Family Farms Customer segment | 3.72 |
| Number of private third party insurance policies | 1.953 |
| Average age | .962 |
| Percent of people belonging to Social Class B1 | .045832 |
| Percent of Singles | .00087 |

**Variables in the Model and Impact.** The variables used in the model for Blue Cross – Blue Shield health maintenance organization (HMO) and their corresponding absolute slope estimates in descending order of magnitude are shown in the table 2 below.

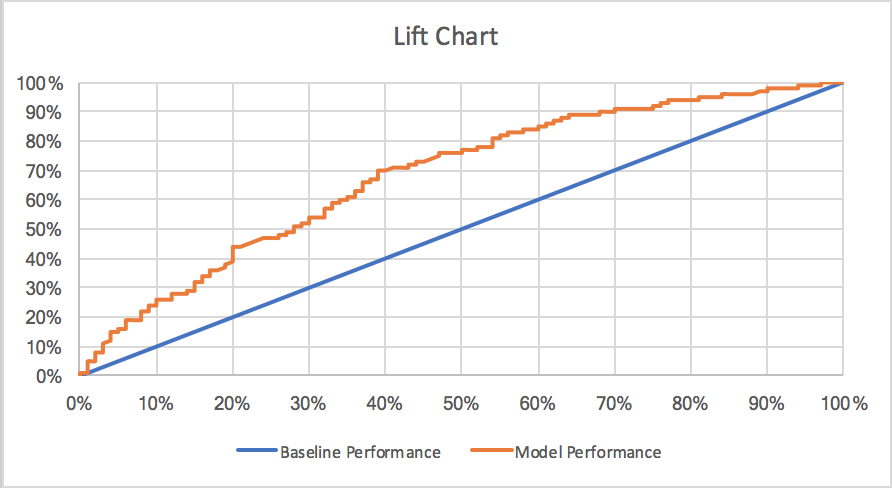
TABLE 2. Variables used in the model and their absolute slope estimates

|  |  |
| --- | --- |
| VARIABLE DESCRIPTION | ABSOLUTE SLOPE ESTIMATE (%) |
| Affluent Young Families Customer segment | 6.73 |
| Student in Apartments Customer segment | 6.579 |
| Dinki's (Double Income No Kids) Customer segment | 5.822 |
| Mixed Rural Customer segment | 5.233 |
| Large Family Farms Customer segment | 3.72 |
| Number of private third party insurance policies | 1.953 |
| Average age | .962 |
| Percent of people belonging to Social Class B1 | .045832 |
| Percent of Singles | .00087 |

You can see that the slope for the Affluent Young Families Customer segment is 6.73%, meaning that If the customer belonged to an affluent young family, then the probability of signing up increased by 6.73%. Similar interpretations can be drawn for variables belonging to Student in Apartments, Dinki's (Double Income No Kids), Mixed Rural and Large Family Farms customer segments. For the number of private third party insurance policies the slope was -0.01953, meaning that for every dollar increase in the amount spent for the private third party insurance policies, probability of signing up decreased by 1.953%. Likewise, the impact of the Average Age variable can be understood. The slope for the percent of people belonging to Social Class B1 in the customer’s neighborhood was -0.00045832, meaning that for every additional percent of people belonging to the Social Class B1 in the customer’s neighborhood, the probability of signing up decreased by 0.045832%. The impact of the Percent of Singles variable can be comprehended in the same way.

**Gains Chart**. The Gains Chart from the model is shown below. The blue line labeled ‘Baseline’ shows the percent of Blue Cross – Blue Shield HMO’s sales 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 sales; 20% of randomly selected prospects would account for 20% of sales, etc.

The line labeled as ‘Model’ shows analogous results if our model is used to select prospects. You can see that the top 10% of prospects account for 25% of all sales using the model. The spot on the gains chart marked with an arrow shows the top 70% of prospects accounts for 91% of all sales using the model for Blue Cross – Blue Shield health maintenance organization (HMO) for the alternative cut-point.

FIGURE 1: Gain Chart

**Alternative Cut-Point**

The lift is the difference between the baseline and model performance. It is the advantage of using the model over random selection, meaning that it describes the increase in the number of responses for every depth in the mail file as opposed to mailing randomly. The table 2 below shows the lift for different baseline and model performances. Therefore, we choose the alternative cut point to be the point where the lift is maximized. For the data provided, the maximum lift of 30% is at 70% where the total sales is 91%. Therefore, if you only mailed applicants below the 30% threshold of riskiness your percent of total sign ups would be reduced to 91% of the total number of sign ups.

TABLE 3: Lift Analysis

|  |  |  |
| --- | --- | --- |
| **BASELINE PERFORMANCE** | **MODEL PERFORMANCE** | **LIFT** |
| 0% | 0% | 0% |
| 3% | 10% | 7% |
| 8% | 20% | 12% |
| 15% | 30% | 15% |
| 20% | 40% | 21% |
| 28% | 50% | 22% |
| 34% | 60% | 26% |
| **39%** | **70%** | **30%** |
| 54% | 80% | 26% |
| 69% | 90% | 21% |

Results show that the top 75% of prospects contain 92% of the total sales. This means that if we cut out 25% of the mailing quantities, we would not lose more than 8-9% of the total sales. The most impactful variables were: whether the prospect was a member of the Affluent Young Families Customer segment or the Student in Apartments Customer segment or the Dinki’s (double income no kids) Customer segment or the Mixed Rural Customer segment or the Large Family Farms Customer segment. If we want to cut the mail quantity by a different percent, an alternative cut-point identified using lift analysis exists at 70% of the prospect universe which contains 91% 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 sign up for Blue Cross – Blue Shield health maintenance organization (HMO). The results of that model are shown below. Details regarding the steps preceding the actual model construction follow.

TABLE 3: Logistic Regression Results

|  |  |
| --- | --- |
| **VARIABLE** | **P-VALUES** |
| Affluent Young Families Customer segment | <0.0001 |
| Student in Apartments Customer segment | 0.0332 |
| Dinki's (Double Income No Kids) Customer segment | 0.0024 |
| Mixed Rural Customer segment | 0.0286 |
| Large Family Farms Customer segment | 0.0440 |
| Number of private third party insurance policies | 0.0003 |
| Average age | 0.0013 |
| Percent of people belonging to Social Class B1 | 0.001 |
| Percent of Singles | 0.0413 |

We comment on the statistical significance of the variables by looking at their p-values in table 3. The p-values of all the variables are less than 0.05, a standard measure of alpha. Since all the p-values are less than alpha, we say that all the variables are statistically significant.

**Data Preparation And Variable Selection**

The initial file contained 5,403 rows and 26 variables. A number of these variables required adjustment prior to building the model.

**Ordinal Variables.** The file contained 15 geo-demographic ordinal variables shown in Table 3. 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 as shown in Table 4. Similarly, the file contained 3 variables regarding the amount a prospect spent on certain products shown in Table 3. These were similarly transformed as shown in Table 4.

TABLE 3: Geo Demographic and Spend Variables

**L3: GEO\_DEMOGRAPHIC VARIABLES**

**L4: SPEND VARIABLES**

|  |
| --- |
| Percent of Roman Catholics |
| Percent of married people |
| Percent of people belonging to Social Class B2 |
| Percent of households with children |
| Percent of Rented houses |
| Percent of people belonging to Social Class A |
| Percent of people with 2 cars |
| Percent of Singles |
| Percent of people with 1 cars |
| Percent of people with the average income |
| Percent of people with a high level education |
| Percent of protestants |
| Percent of people with no cars |
| Percent of people belonging to Social Class B1 |
| Percent of people belonging to Social Class C |

|  |
| --- |
| Contribution of trailer policies |
| Contribution of private third party insurance |
| Contribution of car policies |

TABLE 4: Transformations for Geo Demographic and Spend Variables

|  |  |  |
| --- | --- | --- |
| **VARIABLE VALUE** | **ORIGINAL VALUE** | **NEW VALUE** |
| 0% | 0 | 0 |
| 1-10% | 1 | 5.5 |
| 11-23% | 2 | 17 |
| 24-36% | 3 | 30 |
| 37-49% | 4 | 43 |
| 50-62% | 5 | 56 |
| 63-75% | 6 | 69 |
| 76-88% | 7 | 82 |
| 89-99% | 8 | 94 |
| 100% | 9 | 100 |

|  |  |  |
| --- | --- | --- |
| **VARIABLE VALUE** | **ORIGINAL VALUE** | **NEW VALUE** |
| 0$ | 0 | 0 |
| 1-49$ | 1 | 25 |
| 50-99$ | 2 | 75 |
| 100-199$ | 3 | 150 |
| 200-499$ | 4 | 350 |
| 500-999$ | 5 | 750 |
| 1000-4999$ | 6 | 3000 |
| 5000-9999$ | 7 | 7,500 |
| 10,000-19,999$ | 8 | 15,000 |
| 20,000-? | 9 | 30,000 |

**L3: GEO\_DEMOGRAPHIC FORMAT**

t

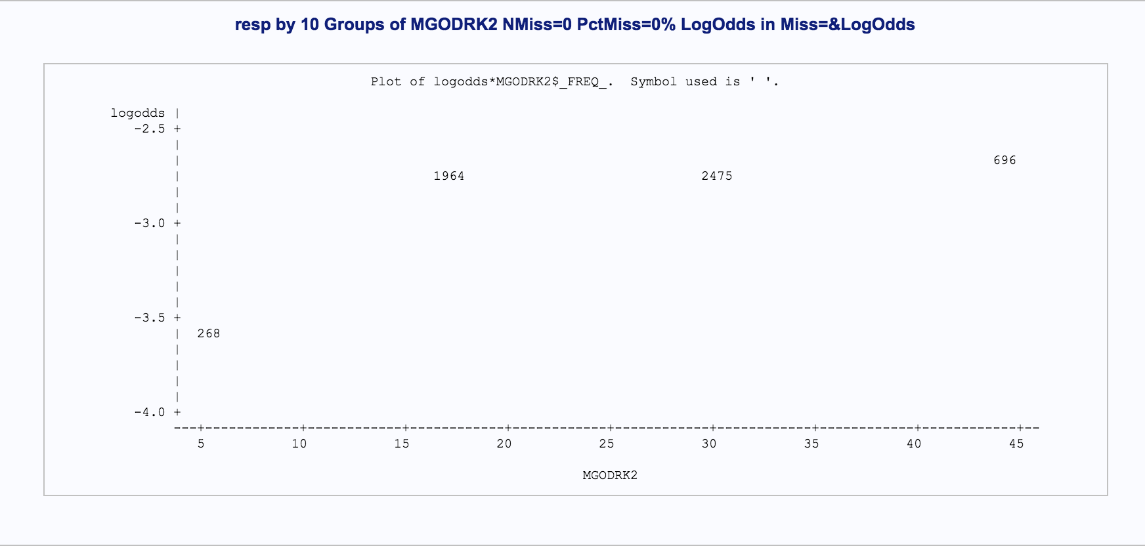
**L4: SPEND FORMAT**

**Categorical Variables.** Two categorical variables were included in the original file. The variables were: Customer main type and Customer subtype. We converted the categorical variables into binary variables so that they could be used in the model. For example, records where Customer main type was equal to 1 (Customer was a successful hedonist), a new binary variable was generated, which was equal to 1 for Customers belonging to the successful hedonist segment and was equal to 0 otherwise.

**Holdout Sample.** We assigned 30% of the data as a holdout sample to check how the model built on the analysis sample works on the unseen data. If the model does not work on the holdout sample, then the analysis sample was over-fit. That means the model found out relationships that exist in one sample but not in another. The performance on the holdout sample is an estimate of the performance of the model on unseen data in the future use. In the business memo, all the performance metrics that are mentioned were calculated based on results from the model that was built on the holdout sample.

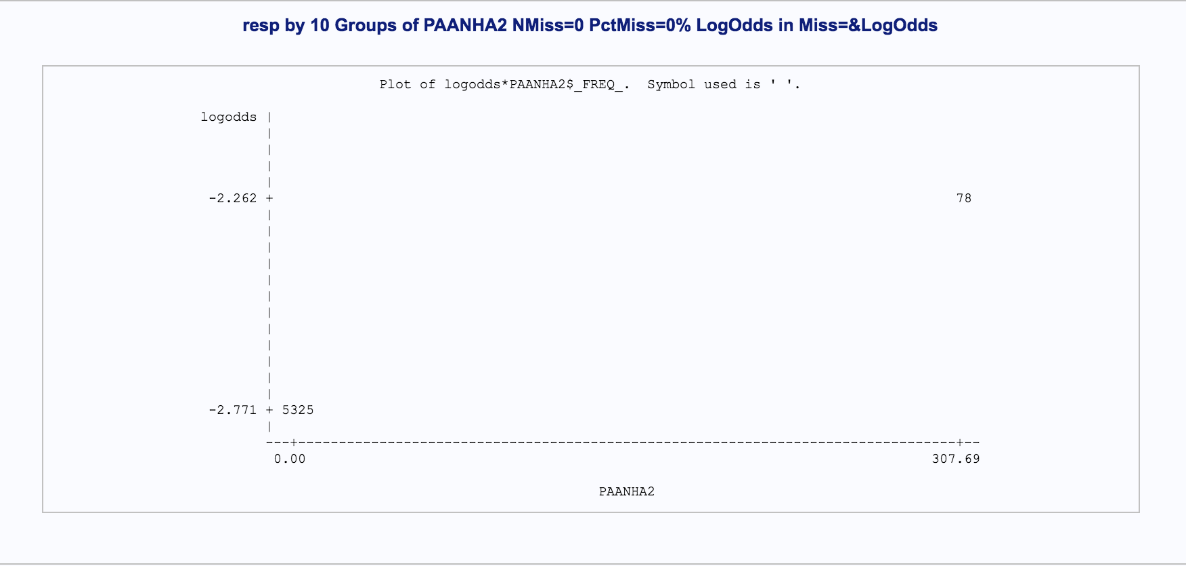
**Non-Linear Relationships.** Graphs showing the response rate for each of the quantitative variables were generated and analyzed. For those variables that appeared to have a non-linear relationship with response we transformed them to new variables exhibiting a quadratic or absolute relationship. For example, the graphs below show the pattern of response rate by the variables: Percent of Roman Catholics, Contribution of trailer policies and Number of third party insurance policies. The pattern is potentially quadratic. Individual logistic regression models for a linear vs. quadratic relationship with response were calculated and the likelihood ratio was examined.

FIGURE 2: Quadratic relationship of the variable Percent of Roman Catholics

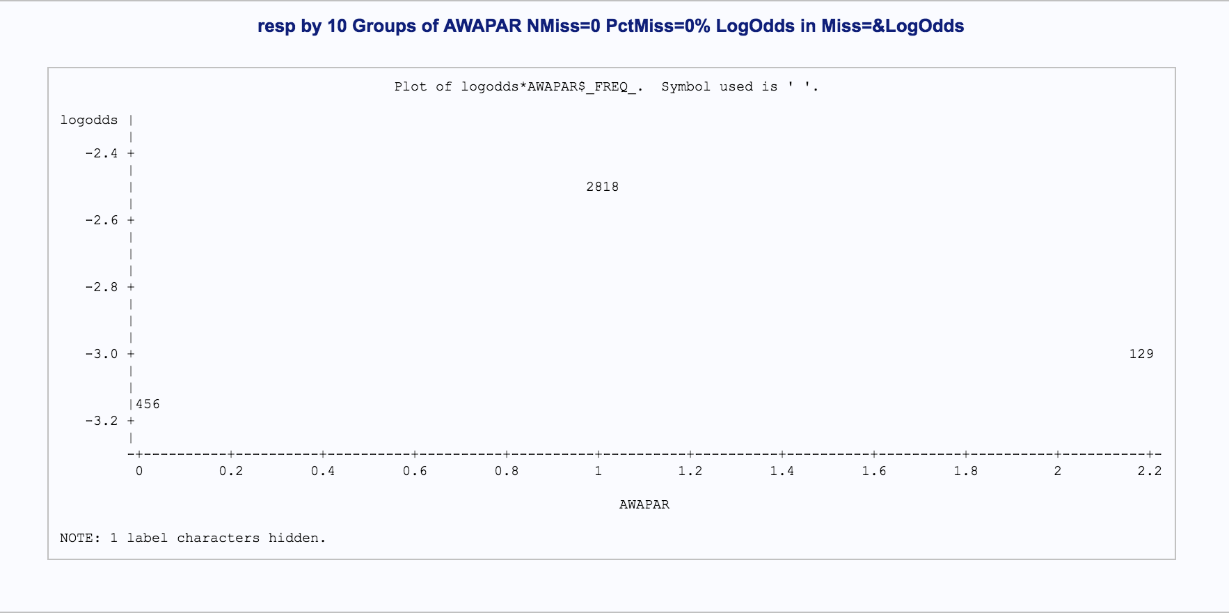


The Likelihood ratio for the linear model was 0.4604 vs. 1.6983 for the quadratic model and 1.3219 for the absolute model and therefore the quadratic form was used as a candidate independent variable for logistic regression for the variable MGODRK2.

FIGURE 3: Quadratic relationship of the variable Contribution of trailer policies



The Likelihood ratio for the linear model was 0.0823 vs. 1.2609 for the quadratic model and 0.7096 for the absolute model and therefore the quadratic form was used as a candidate independent variable for logistic regression for the variable PAANHA2.

FIGURE 4: Quadratic relationship of the variable Number of third party insurance policies 

The Likelihood ratio for the linear model was 12.0449 vs. 20.0969 for the quadratic model and 19.868 for the absolute model and therefore the quadratic form was used as a candidate independent variable for logistic regression for the variable AWAPAR.

In summary,

TABLE 5: Examining Non-Linear Relationships

|  |  |  |  |
| --- | --- | --- | --- |
| **VARIABLE NAME** | **% of ROMAN CATHOLICS** | **CONTRIBUTION OF TRAILER POLICIES** | **NUMBER OF THIRD PARTY INSURANCE POLICIES** |
| Likelihood Ratio for Linear Relationship | 0.4604 | 0.0823 | 12.0449 |
| Likelihood Ratio for Quadratic Relationship | 1.6983 | 1.2609 | 20.0969 |
| Likelihood Ratio for Absolute Relationship | 1.3219 | 0.7096 | 19.868 |

Logistic Regression Model. After preparing all the variables for potential use in the model, all 74 variables were submitted to a logistic regression model using stepwise variable selection. This resulted in 9 statistically significant variables as shown in table 3.

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 to find the slope estimates of every attribute and understand their impact in the model performance. The coefficients from that model were used to show the change in probability of signing up for the Blue Cross – Blue Shield health maintenance organization (HMO).