Re: Invested $1,000 in a certificate of deposit at Chase

This memo is in response to the request to build a model for converting prospects to invest $1,000 in a certificate of deposit at Chase Bank.

Specifically, the model and analysis address the following 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 75% of prospects contain 91% of the total sales. The variables used in the model are listed in table 1 at the bottom of this page. The most impactful variables were: If the prospect was a member of the young all-American family segment, a member of customers living well segment and the number of private 3rd party insurances. All of these had a positive association with purchase. An alternative cut-point exists at 40% of the prospect universe which contains 67% of the sales. This cut-point was identified using marginal 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 investment of $1,000 in a certificate of deposit at Chase. The results of that modeling showed that the top 75% of prospects contained 91% of the total sales. By abstaining from contacting the bottom quarter of the prospect file you can increase your response rate from 6% to 7.5%.

Variables in the Model. The variables in the model are shown below in table 1.

Table 1

Variable	**Impact on Probability of Sale**
Young all-American family	3.7583%
Number of private third party insurance	3.0142%
Customers living well	1.0844%
High level education	1.0743%
Driven Growers	1.0228%
Social class B2	0.6536%
Roman Catholic	0.3514%
Number of car policies	0.0004%

Variables in the Model and Impact. The variables used in the model for mobile home insurance are shown in the table below. You can see that the slope for young all-American family segment is 3.758%. So, if the prospect is a member of young all-American family segment, then the probability of a sale increases by 3.758%. Similarly, the slope for Number of private third-party insurance, 3.014%, indicates that with an increase in the prospect's private third-party insurances by $1, the probability of a sale will also increase by 3.014%. The impact of High-Level Education in the Neighborhood is 1.07%. This means that with every 1% increase in High Level Education people in the prospect's neighborhood, the probability of a sale increases by 1.07%. The slopes for all the variables are shown in Table 2 below.

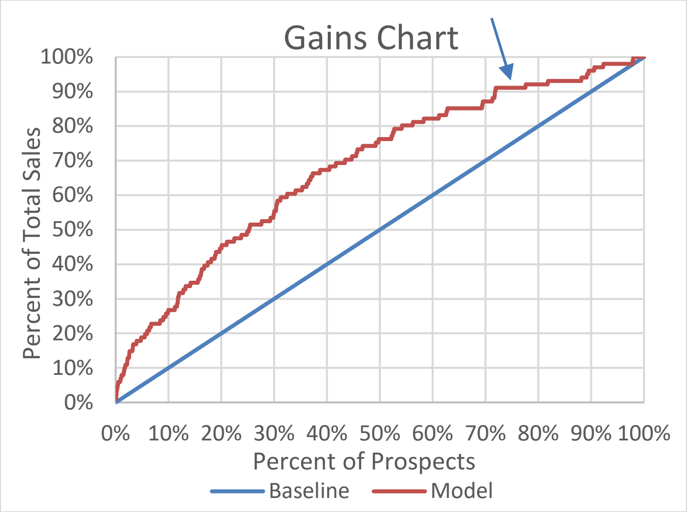
Table 2

Variable	Impact on Probability of Sale
Young all-American family	3.7583%
Number of private third party insurance	3.0142%
Customers living well	1.0844%
High level education	1.0743%
Driven Growers	1.0228%
Social class B2	0.6536%
Roman Catholic	0.3514%
Number of car policies	0.0004%

Gains Chart. The Gains Chart from the model is shown below. The blue line labeled 'Baseline' shows the percent of 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 the model is used to select prospects. You can see that the top 10% of prospects account for 27% 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 91% of all sales using the model for investment in certificate of deposit.

Figure 1



Alternative Cut-Point

The table below shows the increase in the percent of total sales for every 5% increase in the size of prospects file contact, after sorting from highest to lowest risk based on the probability of sale. You can see that after 40% of prospects contacted that less than 5% of the sales are gained. Therefore, if you only contacted prospects above 40% of total prospects then you would yield 49% of the total number of sales.

Table 3

Cumulative Percent of Contacts	Cumulative Percent of Sales	Difference of Contacts	Difference of Sales
5%	19%		
10%	27%	5%	8%
15%	35%	5%	8%
20%	45%	5%	10%
25%	50%	5%	5%
30%	55%	5%	6%
35%	61%	5%	6%
40%	**67%**	**5%**	**6%**
45%	71%	5%	4%
50%	76%	5%	5%
55%	80%	5%	4%
60%	82%	5%	2%
65%	85%	5%	3%
70%	87%	5%	2%
75%	91%	5%	4%
80%	92%	5%	1%
85%	93%	5%	1%
90%	96%	5%	3%
95%	98%	5%	2%

In summary, results show that the top 75% of prospects contain 91% of the total sales. The variables used in the model are listed in table 1 on page 1. The most impactful variables were: If the prospect was a member of the young all-American family Segment, a member of Customers living well Segment and the number of private 3rd party insurances. All of these had a positive association with purchase. An alternative cut-point exists at 40% of the prospect universe which contains 67% of the sales. This cut-point was identified using marginal analysis as shown in Table 3 above.

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 invest $1,000 in a certificate of deposit at Chase. The results of that model are shown below. Details regarding the steps preceding the actual model construction follow.

Table 4

Variable	Description	Estimate	Standard Value	Z-value	P-value
MGODRK	Roman Catholic	0.01932	0.00662	2.91880	0.00351
APERSA	Number of car policies	0.64552	0.13924	4.63615	0.00000
MOSHOO_2	Driven Growers	0.53858	0.20973	2.56802	0.01023
MOSHOO_5	Customers living well	-0.90282	0.35437	-2.54769	0.01084
MSKB2	Social class B2	-0.05704	0.02097	-2.71959	0.00654
MOPLHO	High level education	-0.02422	0.00950	-2.55095	0.01074
AWAPAR	No. of private 3rd party insurance	0.78471	0.36191	2.16823	0.03014
MOSTYP_13	Young all-American family	0.79727	0.38342	2.07937	0.03758
(Intercept)	-	-3.57247	0.27523	-12.98014	0.00000

The p-values for each of the variables are provided in the far-right column above. They are all well below the 5% threshold, therefore they are all statistically significant, meaning that their impact on these variables contribute the most in explaining the variation in the dependent variable.

Data Preparation and Variable Selection

The initial file contained 5,464 rows and 28 variables. A number of these variables required adjustment prior to building the model.

Ordinal Variables. The file contained 18 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 regarding 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	1-10%	5.5%	1	$1-$49	$25
2	11-23%	17%	2	$50-$99	$75
3	24-36%	30%	3	$100-$199	$150
4	37-49%	43%	4	$200-$499	$350
5	50-62%	56%	5	$500-$999	$750
6	63-75%	69%	6	$1,000-$4,999	$3,000
7	76-88%	82%	7	$5,000-$9,999	$7,500
8	89-99%	94%	8	$10,000-$19,999	$15,000
9	100%	100%	9	$20,000+	$30,000

Categorical Variables. Two categorical variables were included in the original file- MOSHOO and MOSTYPE. These were converted to binary so that they could be used in the model. For example, records where MOSHOO was equal to 1 dummy variable MOSHOO_1 was set to 1 to indicate that the prospect belonged to the successful hedonist category and all the other MOSHOO dummy variables (MOSHOO_2, MOSHOO_3 etc.) for that prospect were set to 0. Similarly, records where MOSTYPE was equal to 13 dummy variable MOSTYP_13 was set to 1 to indicate that the prospect belonged to the Young all-American family category and all the other MOSTYP dummy variables for that prospect were set to 0. MOSHOO variable was converted to 10 dummy variables, and MOSTYP was converted to 41 dummy variables.

Holdout Sample. 30% of the records were used for the holdout sample. This was done to know how the model performs on records that were not used to build the original model. The test file is the best estimate for how the model will perform on the records that were not used to build the model. So, all the results, metrics and performance given in the business memo are based on the test file. If this data came from the training file, we would report findings from a potentially over-fit sample because the records from the train set were used to build the model, so it would perform at least a little bit better and either show very high accuracy or low loss.

Non-Linear Relationships. Graphs below show the response rate for each of the quantitative variables that were plotted and examined. For those variables that appeared to have a non-linear relationship with response we tested them for quadratic form. For example, the graph below shows the pattern of response rate by MAUT0 and MSKC. The pattern is potentially quadratic. Individual logistic regression models for a linear vs. quadratic relationship with response were calculated and the AIC values examined, summarized in the table below.

Figure 2

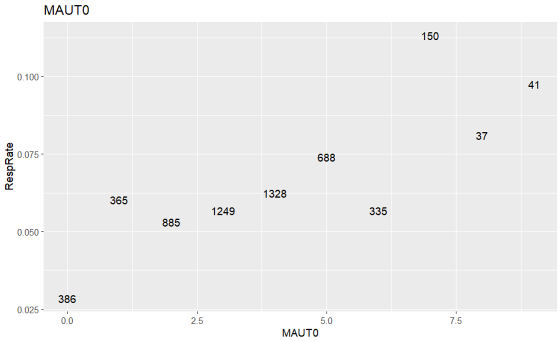


Figure 3

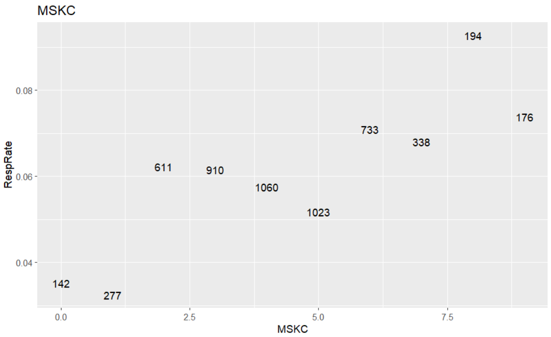


Table 6

Independent Variable	**Linear AIC**	**Quadratic AIC**
MAUT0	2,452	2,471
MSKC	2,482	2,484

For the memo, the AIC for the linear model was 2,452 vs. 2,471 for quadratic model, and therefore the linear form was used as a candidate independent variable for logistic regression. Similar results were seen for MSKC.

Logistic Regression Model. After preparing all the variables for potential use in the model, all 26 variables were submitted to a logistic regression model using stepwise variable selection. This resulted in 8 statistically significant variables as shown in table 4 above on page 4.

Linear Regression. After using logistic regression, the final set of independent variables was used to build a linear regression equation. This was done to produce slopes in terms of probability of sale as opposed to log odds.

Rohan Agrawal