UNIVERSITY - PROJECT

Re: Sales Model for investing $10,000 in the Fidelity Magellan mutual fund

This memo is in response to your request to build a model for converting prospective investors into investing $10,000 in the Fidelity Magellan mutual fund.

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 used in the model?
3. Which of those variables are the most impactful and what is their relationship with the probability of making an investment?
4. What is an alternative cut-point for the prospect universe?

Results show that by cutting our mailing quantities by 25%, the remaining 75% of our prospects contain 93% 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 the amount of dollars spent on the car policies, the percent of Protestant in the prospect's neighborhood and the number of houses owned by the prospect. All of these had a positive association with investment. An alternative cut-point exists at 55% of the prospect universe which contains 83% 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 investment of $10,000 in the Fidelity Magellan mutual fund. The results of that modeling showed that the top 75% of prospects contained 93% of the total successful investments. By abstaining from contacting the bottom half of the prospect file you can increase your response rate from 5% to 7%.

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

VARIABLES IN THE MODEL	PROBABILITY OF RESPONSE	DESCRIPTION OF THE VARIABLE
MOSTYPE_34	-0.04264 %	Prospect belongs to a large family and has an employed child.
MOSTYPE_21	-0.04170 %	Prospect belongs to young urban community.
MAANTHUI	-0.01858 %	Number of houses owned by the prospect (1-10).
MGEMOMV	0.00970 %	Average household size of the prospect (1-6).
MAUT2	0.00056 %	Prospect owns two cars.
MGODPR	0.00049 %	Percentage of protestants near the prospect's neighborhood.
PPERSAUT	0.00002 %	Amount of Dollars spent by the prospect on car policies.

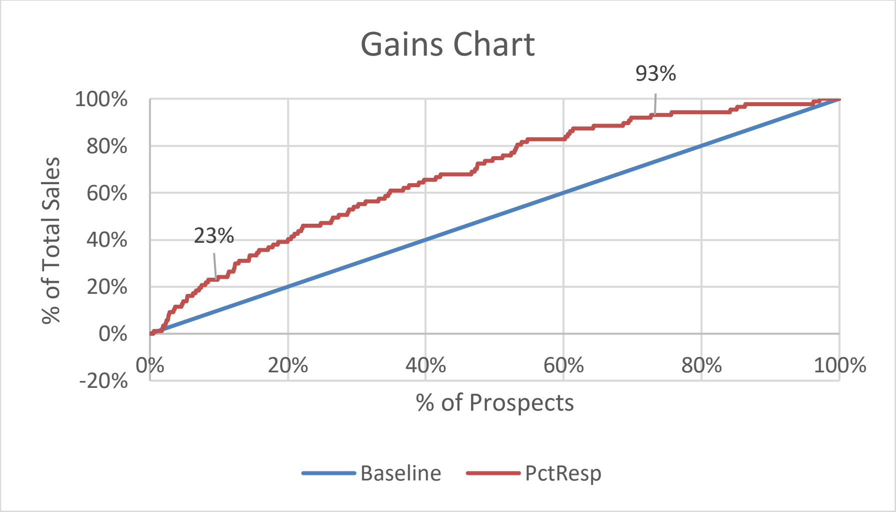
(Table-1)

Variables in the Model and Impact. The variables used in the model for converting prospective investors into successfully investing $10,000 in Fidelity Magellan mutual fund are shown in table 1. You can see that the slope for the percent of protestants in the prospect's neighborhood was 0.00049, meaning that for every additional percent of protestants in the prospect's neighborhood the probability of investment increased by 0.05%. Similarly for the amount spent on car insurance policies the slope was 0.00002, meaning that for every additional dollar spent on car policies, the probability of investment increased by 0.002%. For every additional house owned by the prospect the probability of investment decreased by 2%. The slopes for all the variables are shown in the table.

VARIABLES IN THE MODEL	PROBABILITY OF RESPONSE	DESCRIPTION OF THE VARIABLE
MOSTYPE_34	-0.04264 %	Prospect belongs to a large family and has an employed child.
MOSTYPE_21	-0.04170 %	Prospect belongs to young urban community.
MAANTHUI	-0.01858 %	Number of houses owned by the prospect (1-10).
MGEMOMV	0.00970 %	Average household size of the prospect (1-6).
MAUT2	0.00056 %	Prospect owns two cars.
MGODPR	0.00049 %	Percentage of protestants near the prospect's neighborhood.
PPERSAUT	0.00002 %	Amount of Dollars spent by the prospect on car policies.

(Table-2)

Gains Chart. The Gains Chart from the model is shown below. The blue line labeled 'Baseline' shows the percent of successful investments 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 based on the model. You can see that the top 10% of prospects account for 23% of all successful investments using the model. Similarly, it also shows that the top 75% of prospects accounts for 93% of all successful investments using the model for successful investment of $10,000 in Fidelity Magellan mutual fund.



(Figure-1)

Alternative Cut-Point

Taking the concept of maximum lift in consideration, you can see that after mailing 55% of the prospective investors contains 83% of investors successfully agreeing to invest $10,000 in Fidelity Magellan mutual fund, however as we mail more prospects the resulting response rate is not significant at every increment and is extra customer acquisition cost to the firm. Therefore, according to the model, it is advised that we mail 55% of the prospective investors as it is possible to cut mailing quantities by 25% and still get most of our responses, this cut would increment the response rate by 2% providing the maximum response rate of 8% for the least number of prospects contacted and save the firm almost 45% of customer acquisition cost opposed to contacting all prospective investors. The key variables for the segregation of potential prospects are the amount of dollars spent on the car policies, the percent of Protestant in the prospect's neighborhood and the number of houses owned by the prospect.

Cumulative	*Cumulative*	*Percent of*	*Pct of*	
Prospects	*Responses*	*Prospect*	*Response*	*Lift*
886	72	55%	83%	0.28

(Table-3)

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 $10,000 in the Fidelity Magellan mutual fund. The results of that model are shown below. Details regarding the steps preceding the actual model construction follow.

	ESTIMATE	STANDARD ERROR	Z. VALUE	PR(>|Z|)
(INTERCEPT)	-4.7445933317	4.655729e-01	-10.190872	2.178027e-24
PPERSA	0.0003333612	4.191616e-05	7.953046	1.819802e-15
MAANTH	-0.3107343159	1.194889e-01	-2.600528	9.308043e-03
MGODPR	0.0056807275	2.378546e-03	2.388320	1.692562e-02
MGEMOM	0.2172689005	5.250444e-02	4.138105	3.501859e-05
MOSTYP_34	-1.3401857171	6.433546e-01	-2.083121	3.724017e-02
MOSTYP_21	-1.0092513537	4.471093e-01	-2.257281	2.399052e-02
MAUT2	0.0122742210	5.302413e-03	2.314837	2.062186e-02

(Table-4)

This model contains the variables with p-value less than 0.05, this helps us segregate the variables which helps us reject the null hypothesis for the alternative hypothesis.

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 original file contained 16 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. To use these variables in a linear model the original values were re-scaled to the average of the percentage range. For example, the records for variable MGODRK originally represented a range of percentage of roman Catholics in the prospect's neighborhood, after the rescaling, average of that range was re-assigned to the concurrent variable value in the records. Next page has Table-4 demonstrating the rescaling for the associated variables.

L3: Geo Demographic Format

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

(Table–5)

Similarly, the file contained 6 variables regarding the amount a prospect spent on certain products. These were similarly transformed as shown in Table-5 below.

L4: Spend Format

VARIABLE VALUE	ORIGINAL VALUE	ASSIGNED VALUE
0	0	0
1	1 – 49	25
2	50 – 99	75
3	100 – 199	150
4	200 – 499	350
5	500 – 999	750
6	1,000 – 4,999	3,000
7	5,000 – 9,999	7,500
8	10,000 – 19,999	15,000
9	20,000 -?	30,000

(Table-6)

Categorical Variables. Two categorical variables were included in the original file. These were MOSTYPE and MOSHOOFD, they were later converted into binary columns so that they could be used in the model. For example, records containing MOSHOOFD were categorized into dummy variable named Moshoo_1, Moshoo_2, and so on until Moshoo_10 where one dummy variable was assigned the value of 1 if the prospect belonged to the concurrent segment and every other remaining dummy variable was assigned the value 0. For example, records where Moshoo_1 was equal to 1 represents that the prospect is a member of the successful hedonists segment and 0 when the records stated otherwise. Given below is the Table-6 stating the dummy variables and the segment they represent.

L2: Customer segment format for MOSHOOFD

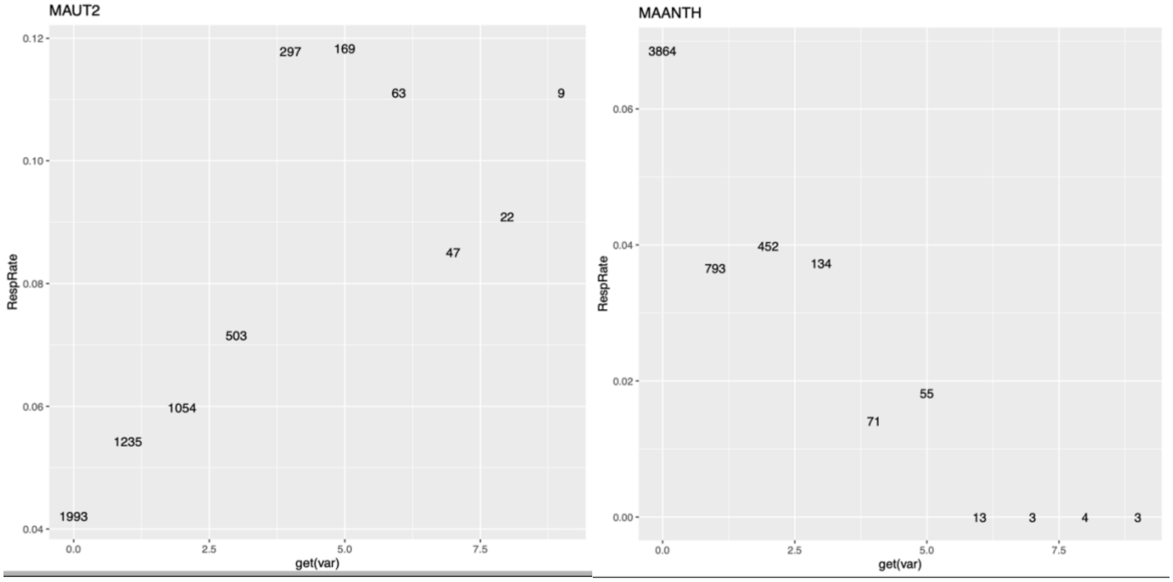
DUMMY VARIABLE NAME	SEGMENT REPRESENTED
MOSHOO_1	Successful hedonists
MOSHOO_2	Driven growers
MOSHOO_3	Average family
MOSHOO_4	Career loners
MOSHOO_5	Living well
MOSHOO_6	Cruising seniors
MOSHOO_7	Retired and religious
MOSHOO_8	Family with grown up
MOSHOO_9	Conservative families
MOSHOO_10	Farmers

(Table-7)

Similarly, the same process was repeated for the other categorical variable, records containing MOSTYPE were categorized into dummy variable named Mostyp_1, Mostyp_2, and so on until Moshoo_41 where one dummy variable was assigned the value of 1 if the prospect belonged to the concurrent segment and every other remaining dummy variable was assigned the value 0. This variable demonstrated the customer subtype and represented over 41 different subtypes as mentioned above.

Sampling the data. Once data was prepared and converted into binary columns, the next step was to segregate the data into two different datasets; training (holdout sample) and testing datasets, the split executed on the dataset was 70% for training and 30% for testing. After splitting, we trained a logistic regression on the training dataset and evaluated the model’s performance on testing dataset, after which we selected the variables with p value less than 0.05, since those are variables with most significance in our model. During the model preparation, it was made sure that there is no overfitting and dataset is pruned accordingly.

Non-Linear Relationships: For those variables that appeared to have a non-linear relationship with response were evaluated and plotted as a part of the data preparation. For example, the graph below shows the pattern of response rate with respect to variables MAANTH and MAUT2.



(Figure-2)

The pattern is potentially quadratic. Individual logistic regression models for a linear vs. quadratic relationship with response were calculated and the AIC values were examined and have been summarized in the table below.

	MAANTH (AIC)	*MAUT2 (AIC)*
Linear	1727	1735
Quadratic	1719	1732

(Table-8)

The AIC value for the linear model was 1727 vs. 1719 for quadratic in case of MAANTH and therefore the quadratic form was used as a candidate independent variable for logistic regression.

Logistic Regression Model. After splitting, we trained a logistic regression on the training dataset and evaluated the model's performance on testing dataset, after which 7 variables were with p value less than 0.05, since those are variables with most significance in our model. During the model preparation, it was made sure that there is no overfitting and dataset is pruned accordingly.

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 rescale the logistic regression output from log-odds of investments to probability of investment. The coefficients from that model were used to explain the probability of investment. All variables with high impact and less p-value along with their probability of investment (%) and description are presented in the table below.

VARIABLES IN THE MODEL	PROBABILITY OF RESPONSE	DESCRIPTION OF THE VARIABLE
MOSTYPE_34	-0.04264 %	Prospect belongs to a large family and has an employed child.
MOSTYPE_21	-0.04170 %	Prospect belongs to young urban community.
MAANTHUI	-0.01858 %	Number of houses owned by the prospect (1-10).
MGEMOMV	0.00970 %	Average household size of the prospect (1-6).
MAUT2	0.00056 %	Prospect owns two cars.
MGODPR	0.00049 %	Percentage of protestants near the prospect’s neighborhood.
PPERSAUT	0.00002 %	Amount of Dollars spent by the prospect on car policies.

(Table-9)