Cortex Competition (Round 2)

Optimizing Operation Surplus on Fundraising Campaign for Non- Profit Organization

Presented to Prof. Cristina Anton
in the context of MBAN 5210- Predictive Modelling II

Group 7:

Ahmed Saqib
Jun-Jie Chen
Xingkai Wu
Yujie Qian
Vaidehi Karvekar

November 15th, 2021

Table of Contents

Introduction

Predictive modeling is often used by for-profit organizations, but this report shows how even non-profit organizations benefit from using the same. Many non-profits have untapped potential and sleeping dogs in their database. The cost of contacting everyone on the database is sometimes remarkably high. Non-profits can use predictive modeling to identify persuadable, sure things, lost causes, and sleeping dogs to reduce contact cost and maximize donations.

The aim of using predictive modelling here is to help Non-profit organization to identify prospective donors based on historical data collected and stored by organization throughout the years. By using modelling techniques, organization will be able to forecast future donors and donation amounts.

The dataset used to develop two-stage model contains socioeconomic and past donation behaviour about 1 million users. The more information about dataset variables can be found in Appendix A.

For building the model, first the variables that affected past donations will be identified. The first stage would be to predict donation probability for each donor and the second stage would be predicting donation amounts. Both models will be combined with contacting costs to identify the operating surplus. This will help organization in gain deeper understanding about expenses and target operating surplus.

Data Preparation and Assumption

The CORTEX members dataset contains data for the past 12 years for each of the one million members. Upon inspection, it was found that 663,666 members had not donated in the first 10 years. As a result, the *Frequency, Recency, Seniority*, and columns related to gift amounts had missing values. Therefore, we first carried out missing value imputation for these members We assumed the following values for each of the variables for members who **did not donate in the first 10 years**.

Frequency:

- Is the number of times a member had donated; we have assumed that value of 0 in place of a missing value.

TotalGift, MaxGift, MinGift:

- Since these values are calculated based on the donations members had made, we assumed a value of 0 in place of a missing value.

***Seniority*:**

- *Seniority* is defined as the number of years since the first gift. Since Seniority is a categorical variable, we had to make sure any value of seniority showing up in the score dataset was also seen in the training (hist2) dataset. We found that seniority had a value of 0 in the score dataset for members who had first donated in 2020. The same members had a missing value in the hist2 dataset**.**
- Therefore, for consistency purpose, if *GaveLastYear* was 1 in hist2 dataset, we assigned missing seniority a value of 0 in hist2. However, if the member never donated in the first 11 years, missing seniority was assigned a large value of 99 in all 3 datasets so that we could group such members together.

Recency:

- *Recency* is the number of years since the last gift. Like *Seniority*, we found that *recency* had a value of 0 in the score dataset for members who had first donated in 2020. However, it was missing for such members in hist2.
- Therefore, using the same logic as we did for seniority, if *GaveLastYear* was 1 in hist2 dataset, we assigned missing *recency* a value of 0 in hist2; if the member never donated in the first 11 years, missing *seniority* was assigned a large value of 99.

Log- Transformation:

- Moreover, we found that the continuous variables had a large variability (a small percentage of the data contained significant outliers). Therefore, to reduce the variability of these continuous variables, we used the log of these continuous variables instead of the original variables.
- We have taken the log of the *Salary, Referrals, Total/Max/Min Gift* and *AmtLastYear*.

Data Splitting and Oversampling:

- we have assumed a 70:30 split for training and validation data. Lastly, we had oversampled the *GaveThisYear* = 1 event in the training dataset since it had a 14.95% response rate in the dataset. Therefore, we assumed this was less and oversampled it to increase the response rate to 40%. We made the correction on the predictions so that the results would take into account the actual response rate.

Predictors Selection and Model Building Process

Before implementing the two-stage modeling approach, a screening process was used to select the most useful predictors in order to achieve a more efficient forecasting process and better prediction ability.

Due to the drawback of the excessive number of levels for each categorical variable, Greenacre method was used for controlling the impact of the degree of freedom. As a result, the categorical variables of *seniorlist, seniority, nbactivities* originally represented by 11 levels collapsed to 5, 6, 7 clusters, respectively. With fewer levels in the categorical variables, these predictors have a stronger association with the target variables in the two stages based on the smaller log p-value in Appendix B - D.

To correct the nonlinearity between the continuous predictors with the target variables of the amount given this year and whether the individual gives this year in the two stages, log transformations were applied on the variables for the amount given last year, annual salary, number of referrals, total donations, minimum donation, and maximum donation.

Combined with age and transformed continuous variables, the original 7 clusters have been split into 5 groups according to the eigenvalue threshold of 0.7 in the variable clustering process. Based on one of the clustering performance criteria, which is the 1-R**2 ratio used here, the continuous variables were reduced to *logTotalGift, logSalary, logAmtLastYear, Age, logReferrals.*

After checking the irrelevant inputs and multi- collinearity by correlation metrics and Variance Inflation Index, all selected categorical and continuous variables were kept: *logSalary,*

logAmtLastYear, logTotalGift, Age, LogReferrals, NbActivities, SeniorList, Woman, Education, City, Contact, Seniority, GaveLastYear. The interactions between these predictors were detected by a forward selection, which was used to automatically enable the statistically significant variables to enter. Based on the SBC-reduction based p-value criteria, which balanced the goodness of fit and number of parameters in the model, 32 interactions were introduced into the model, which is listed in Appendix E.

Final Model

The final model is a combination of linear regression and logistic regression, where logistic regression model determines the probability of an individual donating (***GaveThisYear***), and linear regression determines the amount of an individual will donate (***AmtThisYear***).

The maximum likelihood analysis estimates in the final Logistic Regression model tells us that the more participations to annual meeting, and the newer donors, and more total amount donated, the more likely such candidate will donate after being contacted.

The final Linear Regression model suggests that on average, women is more likely to make more donation than men do; donors with university background who live in suburban area will donate more than in other areas (City, Downtown, Rural)

The two models will perform predictions on the given score datasets, one of which defined contact equals to One and another dataset defined contact equals to Zero. The product of probability and amount on each dataset will give the expected value of an individual would donate, and the difference between the two expect values gives the uplift after contacting. This technique allows marketing team to effectively contact individuals with highest lift and probability of donation after contacting.

In the final model, we want to identify those who are most persuadable, and avoid the sleeping dot, sure things, and lost causes (**See Appendix** F)

Persuadable: Those expected donation amount will be boosted after being contacted

- Uplift > 25, to make sure that associated cost can be covered and generate positive surplus.
- Probability of donation after being contacted > 0.4, to make sure those uplifts are realistic.
- (Probability of donation after being contacted - Probability of donation before being contacted) > 0.1, to make sure probability of making such uplift were persuaded.

Sleeping dogs: Those expected donation amount becomes less after being contacted

- Uplift < 0, namely, these individuals will make less donation after being contacted
- Expected donation amount > 25, to make sure those individuals were intended to make donation before being contacted.
- Probability of donation before being contacted > 0.4, to make sure their intentions are realistic

Sure- Things: Expected donation amount is high and about the same after being contacted

- Probability of donation before ***and*** after being contacted > 0.4, and the probability change are smaller than 0.1.
- Expected amount of donation before being contacted > 25

Loss Causes: expected donation amount is low and about the same after being contacted.

- Probability of donation before ***and*** after being contacted < 0.4, meaning that these individuals are not likely to donate before and after being contacted.

Final Model Performance

This part will introduce the performance of logistic and linear models individually and show the final decision we made in Cortex with a combination of two models.

For the logistic model, the graphs of ROC curves for the training and validation dataset can be found in Appendix G. The computed area under the curve or c-statistic for training dataset is 0.7297 and for validation dataset is 0.7326. The above 70 percent c-statistics show that our model has a goodness of fit. With a 0.29% difference between the performance from training and validation datasets, the potential problem for overfitting was fixed, so should in the unseen data to predict the probabilities of donation.

For the linear model, the graph in Appendix H shows that the root mean squared error is 219, Adjusted R-Square is 0.0153, meaning that the standard deviation of unexplained variance is 219 with 1.52% explained variability.

Just like mentioned above in our final model part, after combing the scored dataset of the logistic and linear models and calculating the ENC, EC and uplift, we chose 222,098 candidates, who had an uplift more than 25 and a probability of more than 40% to donate after contact, and the final expenses for contacting are $4,352,450. The operating surplus of "Real Life" dashboard is unknown for now.

Model Limitation: The model is currently adapted with linear regression with low R-squared and high RMSE, which does not explain variance well; in the future model development, decision tree regressor can be adapted to reduce the RMSE and boost model performance.

Other Information or Sources

- Can provide the information about within past 3-5 years how many times members were contacted and made donations after. This can better help in predicting sleeping dogs.
- As of now, the data is partially aggregated, if the detailed information in all donation events each donor made was provided, this will further help in modelling and trend analysis.

Future Considerations for The Organization

Update the database regularly, accurately, and include additional socio-economic information about members. Additional/Updated information recorded about members can be used to increase model efficiency and operating surplus by adding or removing variables. Increase engagement with persuadable donors and remove lost causes from database.

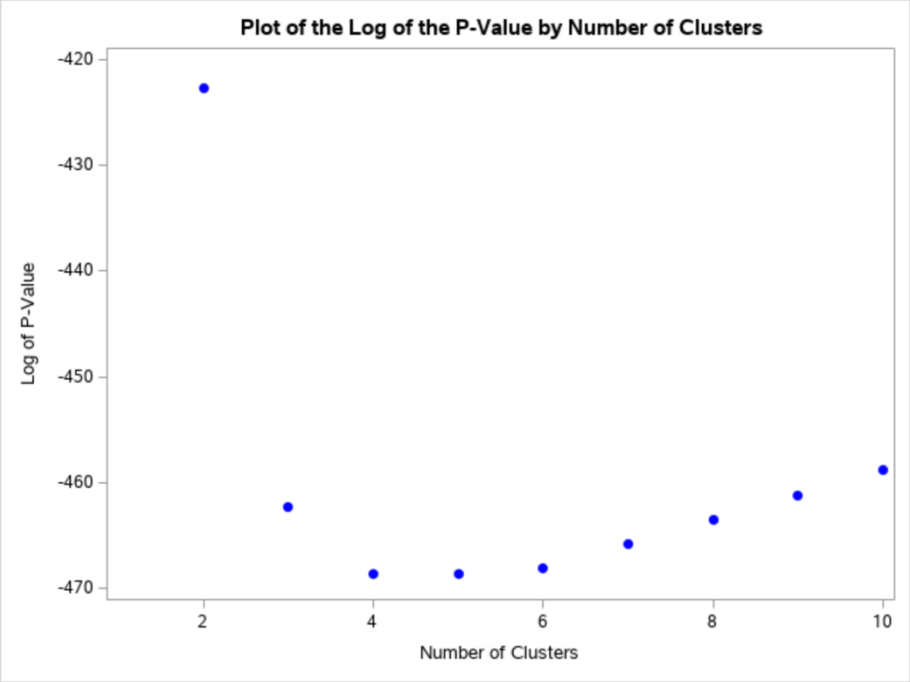
Conclusion Weight Choice

Model performance	Model Write- up
60%	40%

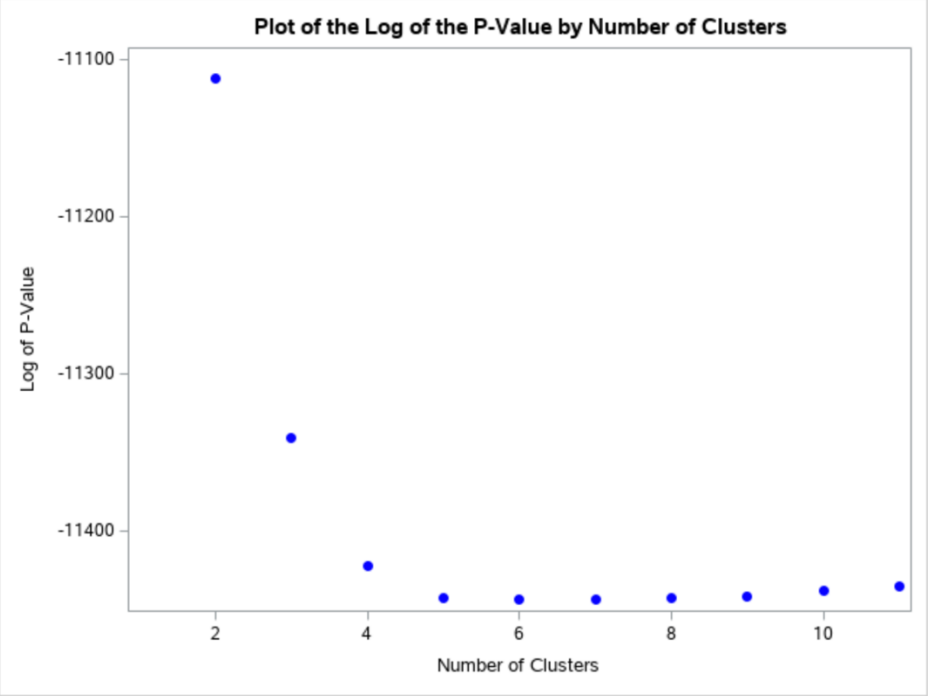
Appendix A Variables in Original Dataset

ID	Member number (Unique ID)
LastName	Last Name
FirstName	First Name
Woman	Sex (1=woman, 0=man)
Age	Age(year)
Salary	Annual salary in USD
Education	highest education level
City	Type of neighborhood
SeniorList	Seniority for being on the VIP list
NbActivities	Number of participations to annual meeting
Referrals	Number of years since last gift
Frequency	Number of donations
Seniority	Number of years since a member
TotalGift	Total donation since a member
MinGift	Minimum donation since a member
MaxGift	Maximum donation since on the VIP list
GaveLastYear	Whether or not the individual gave last year
AmtLastYear	Amount given last year
GaveThisYear	Whether or not the individual gave this year
AmtThisYear	Amount given this year
Contact	Whether member was contacted last year

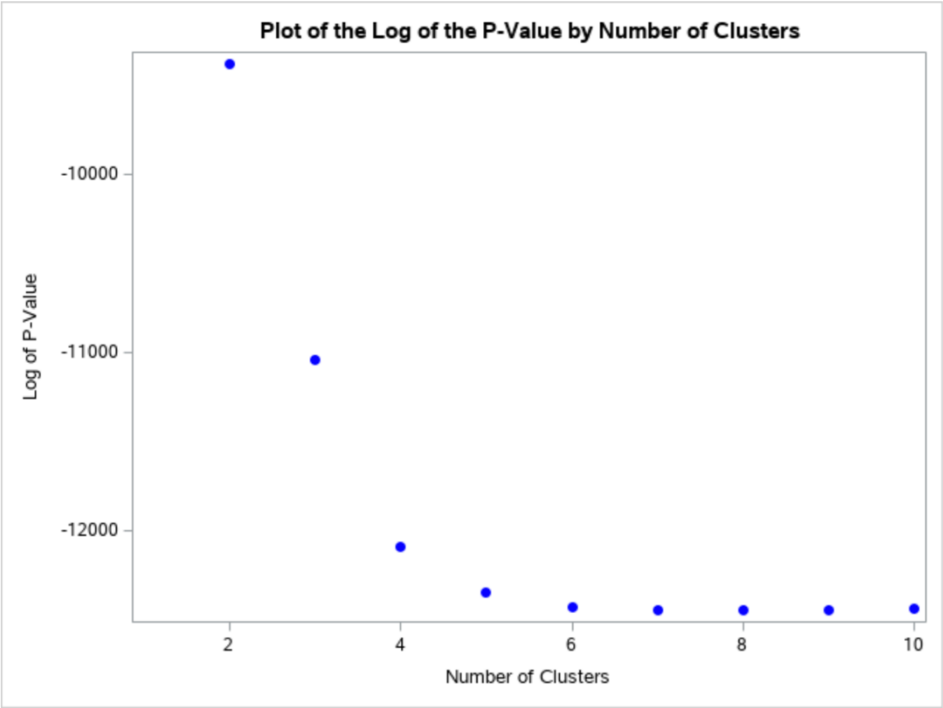
Appendix B Number of Clusters for *Seniorlist*



Appendix C Number of Clusters for *Seniority*



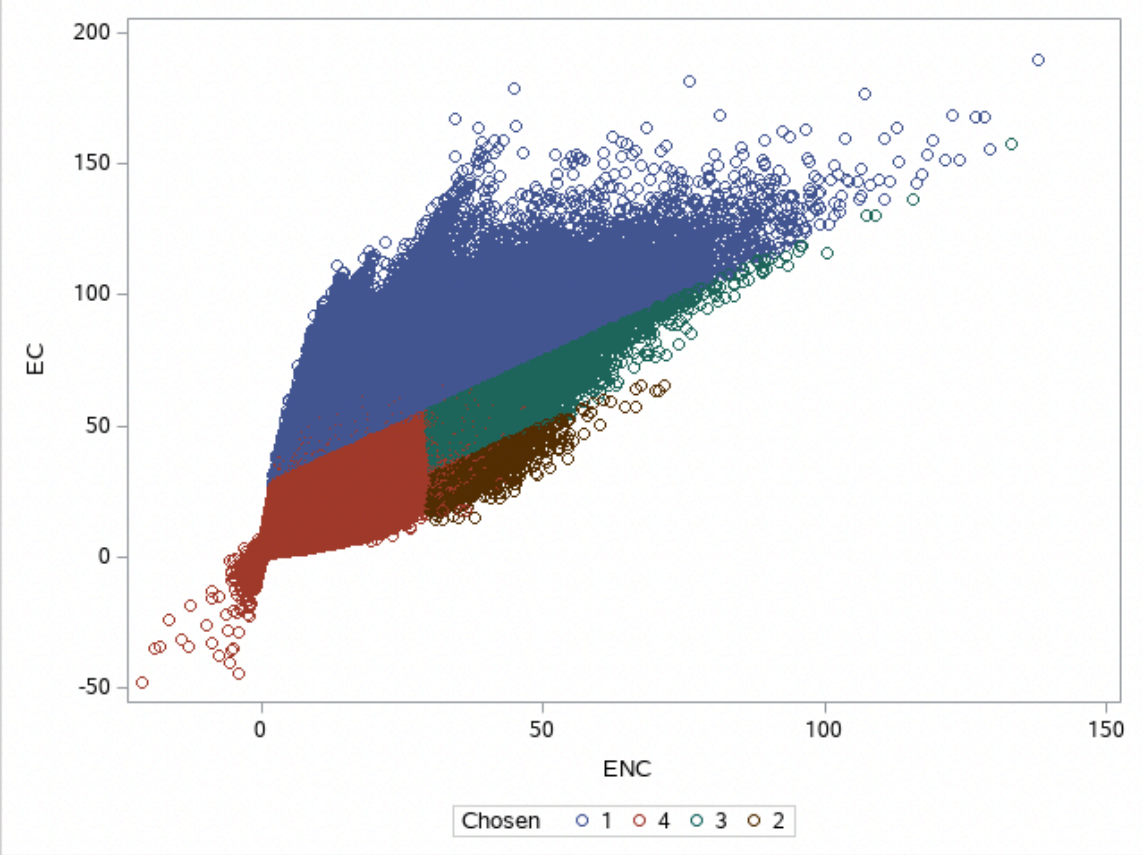
Appendix D Number of Clusters for *Nbactivities*



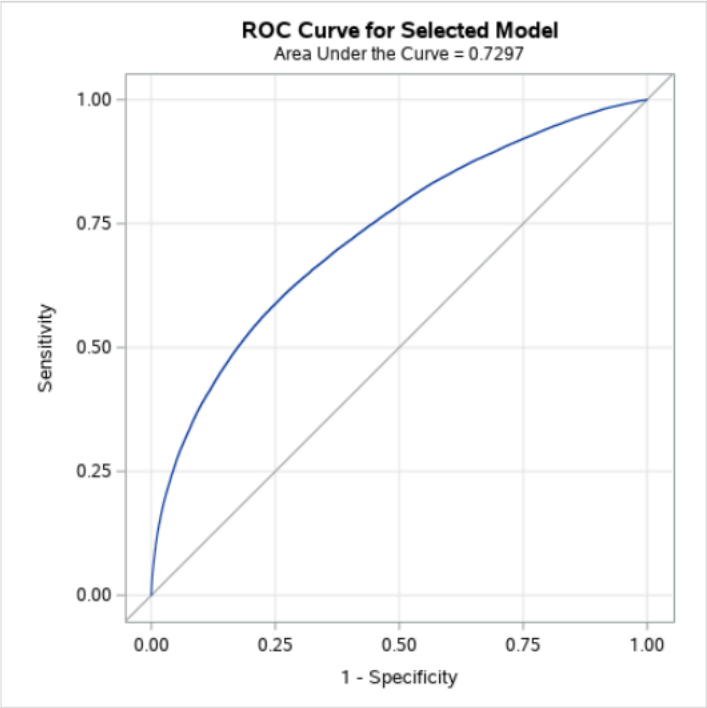
Appendix E Interaction included from Forward Selection

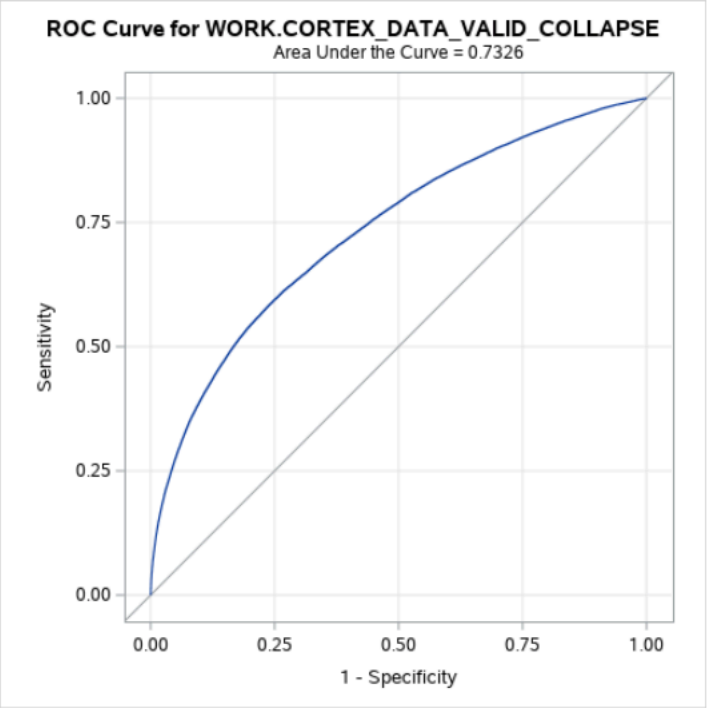
1	Woman*Contact
2	Age*Contact
3	seniority_1*Contact
4	Contact*NbActivities
5	City*Contact
6	SeniorList_1*Contact
7	SeniorLis*seniority_
8	logReferr*NbActiviti
9	City*SeniorList_1
10	seniority*NbActiviti
11	logSalary*Contact
12	Education*City
13	GaveLastYear
14	Education*Contact
15	logAmtLas*logReferra
16	logTotalGift*Contact
17	SeniorLis*NbActiviti
18	logTotalG*seniority_
19	logReferrals*Contact
20	logTotalG*logReferra
21	logReferr*SeniorList
22	logTotalGift*City
23	SeniorLis*GaveLastYe
24	City*NbActivities_1
25	Woman*seniority_1
26	GaveLastY*NbActiviti
27	Woman*NbActivities_1
28	City*seniority_1
29	Education*seniority_
30	Age*seniority_1
31	logSalary*seniority_
32	logAmtLastYear*City

Appendix F ENC, EC and Uplift



Appendix G ROC Curves for Training and Validation Dataset





Appendix H Linear Model Performance

Determine P-Value for Entry and Retention

The GLMSELECT Procedure
Selected Model

The selected model, based on Validation ASE, is the model at Step 53.

Effects:	Intercept Age*Woman logSalary*Woman Age*logSalary Age*Education Woman*City Woman*SeniorList_1 logSalary*SeniorList Education*SeniorList Woman*NbActivities_1 Age*NbActivities_1 logSalary*NbActiviti logReferrals*Woman Age*logReferrals logReferra*Education logTotalGift*Woman logSalary*logTotalGi logTotalGift*City logTotalG*SeniorList Age*Contact logSalary*Contact Education*Contact City*Contact logReferrals*Contact Age*GaveLastYear logSalary*GaveLastYe Education*GaveLastYe SeniorLis*GaveLastYe logReferr*GaveLastYe logTotalG*GaveLastYe logAmtLastYear*Woman Age*logAmtLastYear logSalary*logAmtLast logAmtLast*Education logAmtLastYear*City City*seniority_1 logReferr*seniority_ logTotalG*seniority_

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	124	83469157	673138	14.01	<.0001
Error	111968	5381401158	48062		
Corrected Total	112092	5464870315			

Root MSE	219.23037
Dependent Mean	62.75548
R-Square	0.0153
Adj R-Sq	0.0142
AIC	1320610
AICC	1320610
SBC	1209718
ASE (Train)	48008
ASE (Validate)	60678