Analyzing Direct Marketing Campaign Performance Using Weight of Evidence coding and Information value through SAS Enterprise Miner Incremental Response Modeling Node

Abstract:

Data Mining and predictive models are extensively used to find the optimal customer targets so as to maximize the return on investment. Direct marketing techniques target all the customers who are likely to buy regardless of the customer classification. In a real sense this mechanism couldn't classify the customers who are going to buy even without a marketing contact, there by resulting in a loss on investment. This paper focuses on Incremental Lift modeling approach using Weight of Evidence Coding and Information Value followed by Incremental Response and Outcome model Diagnostics. This model identifies the additional purchases that would not have taken place without a Marketing campaign. Modeling work was conducted using combined model. The research work is carried out on a Travel Center data, which identifies the increase in average response rate by 2.8% and the number of fuel gallons by 244 when compared with the results from the traditional campaign, which targeted everyone. This paper discusses in detail about the implementation of 'Incremental Response' node to direct the marketing campaigns and it's Incremental Revenue and Profit analysis.

Incremental Lift Modeling Approach

This paper presents the motivation behind using Incremental Response node to target the group of customers who are persuadable instead of spending resources on others. Complete work is carried out on a Travel Center data. Incremental Response node requires a Treatment variable, a response target that is a binary with 1 being a treatment group and 0 a control group and the Target response variable. Here, the Outcome target is the number of gallons the customer fuels in a specific marketing campaign.

In general, the fundamental approach of handling categorical variables in modeling is the dummy variable coding and this result in a situation called curse of dimensionality and may cause over fitting of data. So, the best way to work with a target categorical variable is to perform the WOE (weight of Evidence) coding. WOE and IV (Information Value) have been extremely useful in variable reduction and results in variables, which got high predictor importance.

Weight of Evidence & Information Value

The target variable is the customer response. let's code it as 'Y' and assume Y=1 as responded and Y=0 as not responded. The predictor variables are categorized into mutually exclusive bins, so the weight of evidence for each bin is going to be the log of the customers who responded to the ones who didn't respond.

Weight of Evidence can be calculated as

$$WOE_i = \log \frac{P(X = x_i|Y = 1)}{P(X = x_i|Y = 0)} \quad \text{for } i = 1,2,\ldots,I$$

Whereas in Incremental Response Modeling, we got Treatment and Control groups and the Net Weight of Evidence (NWOE) is calculated in the same way, but with the inclusion of both the Treatment and Control group responses.

$$NWOE = \log \frac{P(X = x_i|Y = 1)_T / P(X = x_i|Y = 0)_T}{P(X = x_i|Y = 1)_C / P(X = x_i|Y = 0)_C}$$

Weight of Evidence analyzes the predictive power of a variable with respect to the targeted outcome but the Information value gives the overall predictive ability of the variables being considered and this helps in comparing the predictive importance of one with that of the other variables.

$$IV = \sum_i (P(X = x_i|Y = 1) - P(X = x_i|Y = 0)) \cdot WOE_i$$

Information value for Incremental Response Modeling is going to include the Treatment and Control groups in its calculation to determine the predictive importance of variables.

Initial Exploratory Analysis

The data, which is being used, is of a Travel center and comes from different tables of SQL Server databases, which are very huge containing 40 to 50 tables in each database. All the data is extracted with respect to a specific campaign, which is being implemented in a conventional method of A/B Testing. SAS codes are written to blend the data as required and a final data set is of 31 variables and 4,117 records. The figure 1 shows the variable summary giving the roles and measurement levels of all the variables involved.

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	1
INPUT	BINARY	8
INPUT	INTERVAL	8
INPUT	NOMINAL	7
INPUT	ORDINAL	1
REJECTED	NOMINAL	3
TARGET	BINARY	1
TARGET	INTERVAL	1
TREATMENT	BINARY	1

Figure 1. V ariable Summary

The main purpose of the Incremental Response modeling is to know the customers who cause to achieve the True Lift for the Marketing Campaign. This paper explains in detail about the Combined True Lift Modeling Implementation using SAS Enterprise Miner.

Two Sample TTEST

The TTest assesses whether the means of two groups are statistically different from each other. In our context, we can determine whether the average response rate and average number of gallons are different between Treatment and Control groups.

H0: Average Response rate of Treatment group is equal to Control group

H1: Average Response rate of Treatment group is not equal to Control group

SAS Enterprise Guide is used to run the TTest with default options at a 5% significant level. Response is being assigned as Analysis variable and Promotion being the classification variable. TTest provides two different methods to determine the significance based on the equality of variances. Figure 2 shows the Equality of Variances, P value is greater than the significance level, which means that we don't reject the Null Hypothesis and we go with Equal Variances.

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2141	1974	1.09	0.0525

Figure 2. F Test

Now that, we determined to go with Equal Variances (Pooled Method), Figure 3 shows the tvalue and significance values. Figure 3 show that P-value is greater than the significance level, which means that we don't reject the null hypothesis. This says that Average response rate is not significantly different between the Treatment and Control groups.

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4115	1.11	0.2675
Satterthwaite	Unequal	4109	1.11	0.2666

Figure 3. TTest Procedure

Let's do the TTest for Average gallons, which determines whether average gallons are significantly different between Treatment and Control groups.

H0: Average gallons of Treatment group is equal to Control group

H1: Average gallons of Treatment group is not equal to Control group

As we did before, we choose the method based on F-test, which is significant here. So, we choose unequal variances. Figure 4 shows the results from the T test. Here, the P-value using Satterthwaite method is less than the significance level, which is 0.05. This leads to rejection of Null Hypothesis, which concludes Average gallons of Treatment, and Control groups are significantly different.

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4115	4.47	<.0001
Satterthwaite	Unequal	4007.4	4.46	<.0001

Figure 4. T Test Procedure

Imputation

The data being extracted got missing values for few variables and are imputed using Impute node with the default options of Mean for the Interval variables and Count for the class variables and Unique type of Indicator variables are created. The numbers of missing values are relatively small for all the variables except for two variables, one being the 'Billing card', which identifies the type of card being used for purchase and the other one, which deals with the discount offer being provided to a customer.

Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	
Tree Imputation	
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	Unique
Source	Imputed Variables
Role	Input

Figure 5. Imputations Property Panel

In any modeling project, it is essential to proceed for an honest assessment, which improves the credibility of the model being built. So, the data is partitioned to training and validation sets with a ratio of 50% each. The default properties are being used in the properties panel. At this point, the data is in shape as required and is connected to the Incremental Response Node. Changing the properties customizes the model and a combined model is generated.

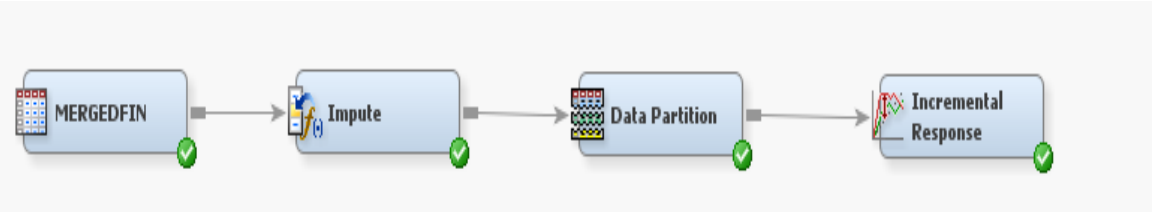
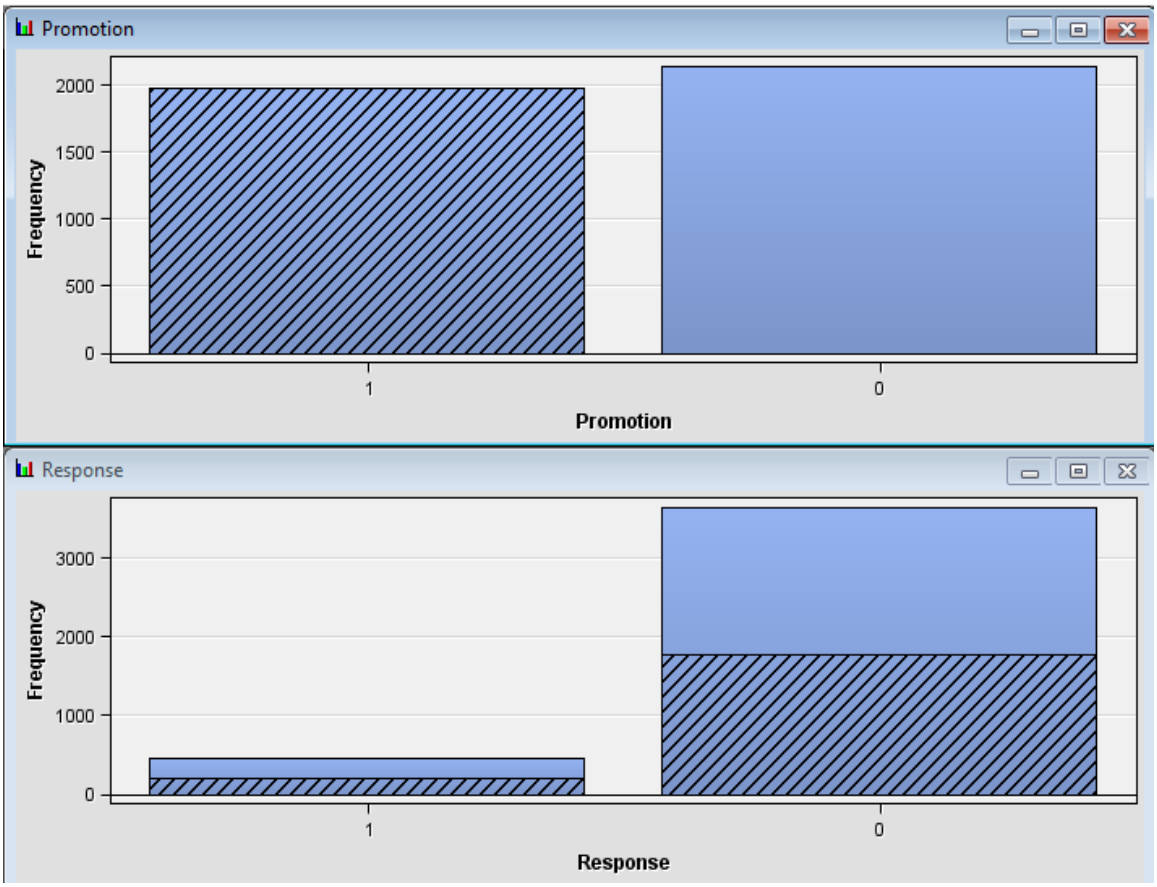


Figure 6. Display of SAS nodes connected

From the initial exploration of the binary response target variable and the treatment promotion variable, we get to see that percentage of response in both the treatment and the control group is about 12.54%.



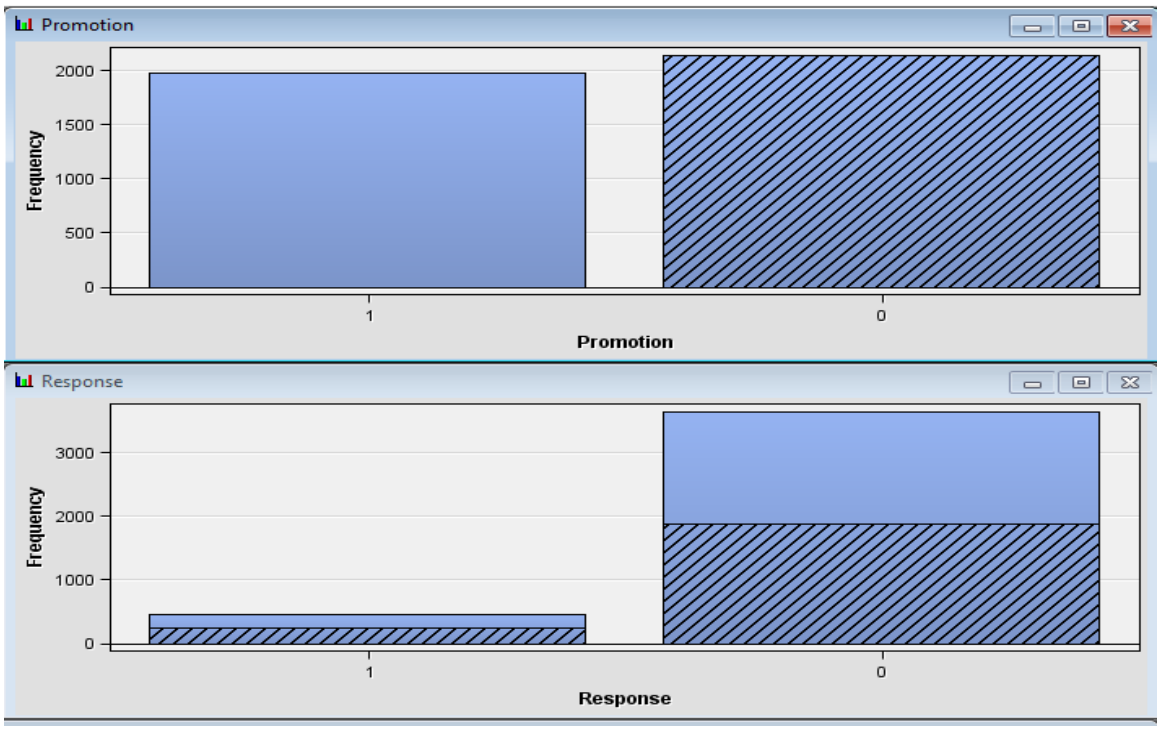


Figure 7. Variable Exploration

Response Outcome Summary

With the descriptive statistics performed, Incremental Response node is all set to go. After we ran this node, it gives all the necessary details for targeting specific customers. Looking over the Response outcome summary table, we got to see an average of around 130 gallons more in Treatment group than in Control with respect to Training data. The values of training and validation data clearly show that the fit of data is pretty good, as the data holds for validation.

Treatment variable:Promotion	Data Role	Number of observations	Number of Response	Rate of Response	Average gallons	Total gallons	Promotion
	0Train	1048	123	0.117366	1543.59	189861.50	
	1Train	1010	107	0.105941	1674.893	179213.61	
	0Validate	1094	127	0.116088	1469.105	186576.40	
	1Validate	965	102	0.105699	1577.671	160922.41	

Figure 8. Table Response Outcome Summary

Till this point, everything looks same as the traditional models. Let's jump in to discuss the real purpose of using this node.

Response Model Diagnostics

Figure 9 shows the incremental response for both the training and validation data. It seems the data holds out for validation, which is a good part. The pattern goes in a nicer fashion as we can observe; the first few deciles got higher values of response compared with the other ones. The predicted Increment in the first decile is about 11% compared to 10% in the observed increment. The difference between the predicted and observed increment should be small and it is of the same case here. The predicted increment comes through targeting just the persuadable and not counting on any other customers. So, this helped the company to design marketing campaign in such a way that it targets the first few deciles.

Percentile	Data Role	Predicted Treatment	Predicted Control	Predicted Increment ▼	Observed Control	Observed Treatment	Observed Increment
10	Train	0.378647	0.262198	0.11645	0.232	0.3375	0.1055
10	Validate	0.352662	0.244091	0.108571	0.226891	0.302326	0.075435
20	Train	0.189561	0.157995	0.031566	0.171171	0.180851	0.00968
20	Validate	0.168979	0.144836	0.024143	0.2	0.152941	-0.04706
30	Train	0.148505	0.143158	0.005346	0.105691	0.195122	0.089431
30	Validate	0.130861	0.13153	-.000669	0.178295	0.25	0.071705
40	Train	0.109121	0.118916	-0.00979	0.175926	0.154639	-0.02129
40	Validate	0.079233	0.092648	-0.01342	0.052174	0.044444	-0.00773

Figure 9. Incremental Response Modeling Statistics

As the model helps in predicting both the response and the outcome, it is of great importance to discuss the Incremental Outcome Model.

Incremental Outcome Model Diagnostics

When we look in to the difference between the average gallons made by the treatment and control groups, it is of 130 gallons, which means a gallons lift of 130 in overall training data. Now that, we build the Incremental model, the Figure10 shows the training and validation data average gallons. The first decile predicted increment gallons is 208 whereas the observed being 207, a slight difference. The point to note here is that, the predicted gallons are just from the specific persuadable customer group, which means that they account to True Lift.

The Travel center used to target all the set of customers and achieved a decent lift in gallons, but the present takes the lift in gallons to a higher level and there by the dollars. They targeted just the first few deciles to get more than the expected lift in gallons and dollars.

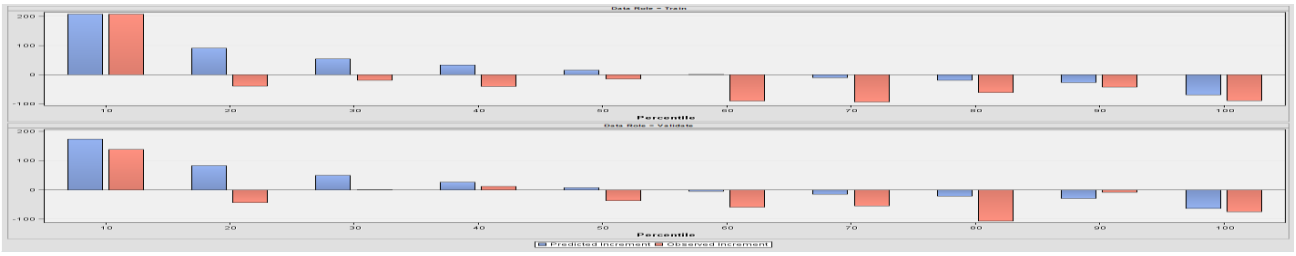


Figure 10. Incremental Model Outcome Diagnostics

The above two steps help in better prediction of the set of customers. We can even get to see the profitable deciles as a result of this node. Average Incremental Revenue shows that the first three deciles are profitable both in Training and validation datasets.

Percentile	Data Role	Profit Indicator	Average Revenue Increment
10	Train	Profitable	307.7214
20	Train	Profitable	59.59015
30	Train	Profitable	13.44971
10	Validate	Profitable	268.8275
20	Validate	Profitable	43.66524
30	Validate	Profitable	5.870103

Figure 11. Average Incremental Revenue

Penalized NIV

We looked through different aspects of the Incremental Response Node. As we discussed earlier about the NWOE and NIV values to determine the variable importance, it is now time to see the variables being selected by the model.

Honest assessment in predictive models is used to assess the stability and robustness of the data, which means that the predictive importance of the Training and Validation dataset shouldn't differ much. If the validation dataset doesn't hold with the training, the model is not robust. Variable, which lacks robustness, results in difference in NWOE of Training and Validation data set.

In order to minimize that, the NIV is adjusted with a penalty term, which takes in to account the difference of NWOE from training dataset and NWOE from validation dataset. This is termed as Penalized NIV

$$PNIV = NIV - penalty$$

Penalized NIV improves the predictive robustness of each variable with respect to the Target.

Figure 12 shows the Variables being selected by PNIV in our data. Seven variables are being selected as the most predictive variables sorted by Penalized NIV

Variable Name	Penalized Net Information Value ▼	Rank Percentile	Selection
IMP_merchandise_sale_net_val	57.92893	2.12766	Yes
IMP_gsow_monthly_potential	56.16455	4.255319	Yes
IMP_loyalty_points_redeemed	31.29528	6.382979	Yes
IMP_ln_flag	9.830942	8.510638	Yes
IMP_restaurant_sale_net_val	5.246825	10.6383	Yes
IMP_mobilenum_flag	2.068668	12.76596	Yes
IMP_tirecare_sale_tot_qty	2.005304	14.89362	Yes

Figure 12. Variables selected by the model

Results

There is an increase in the average response rate from 10% to 12.8% using this model and increase in the average number of fuel gallons from 150 to 394 gallons is also being observed

Conclusion

This Paper provides a different approach to Marketing Campaigns using Incremental Response Modeling and explains in detail on how the True Lift achieved with the help of this node is better than the traditional campaign Lift. It demonstrates the Weight of Evidence Coding and Information value, which helps in variable reduction and presents the high predictor variables. The traditional propensity model may be good at targeting customers, but it couldn't achieve the true responders of campaigns and this model helps the company to achieve that.

References

http://blogs.sas.com/content/subconsciousmusings/2013/07/12/how-incremental-response-modeling-can-help-you-reach-the-right-target-group-more-precisely/

http://support.sas.com/resources/papers/proceedings13/096-2013.pdf

http://www.tcasug.org/files/Incremental_Response_Modeling_20080914.pdf

http://www.socialresearchmethods.net/kb/stat_t.php

Acknowledgement

Dr. Goutam Chakraborty, Professor (Marketing) at Oklahoma State University and Management Consultant

Contact

Sravan Vadigepalli

Email: Sravan.vadigepalli@gmail.com