***Ivanova Y.V.*[1]*, Terentiev O.N.*[1]*, Korshevnyuk L.O.*[1]*, Prosyankina-Zharova T.I.*[2]**

[1] *Institute for Applied System Analysis of NTUU "KPI", Kyiv, Ukraine;* [2] *European University, Uman, Ukraine*

Using modified logistic regression to increase customer response rate to marketing campaigns

Introduction. While creating marketing campaigns, most companies use conventional predictive models, which target all customers who are likely to buy. This approach can lead to wasting money on customers who will buy regardless of the marketing contact [1]. Therefore, considered models which look for customers who are likely to buy or respond positively to marketing campaigns when they are targeted but are not likely to buy if they are not targeted.

Research methodology and result analysis. In incremental response modeling, one of the basic models is the difference score model, which measures the differences between predicted values from the model of the treatment group and from the model of the control group [1]. Suppose there are two data sets, D_T and D_C, which are the treatment group and the control group, respectively.

Denote a dependent variable y and explanatory variables x, and denote the number of observations in each group, n_t and n_c. Then,

$$D_T = \{x_i, y_i\}_{i=1}^{n_t} \tag{1}$$

$$D_C = \{x_i, y_i\}_{i=1}^{n_c} \tag{2}$$

$$D = D_T \cup D_C = \{x_i, y_i\}_{i=1}^{n}, n = n_t + n_c \tag{3}$$

Without loss of generality, a linear model is considered as follows:

$$Y = X\beta + \epsilon \tag{4}$$

Difference score from two separate models. For a more correct modeling two data sets are used – treatment and control. As a result, two models are built separately for each data set:

$$\hat{Y_T} = X_T \hat{\beta_T} \tag{5}$$

$$\hat{Y_C} = X_C \hat{\beta_C} \tag{6}$$

Then, both models are used to calculate predicted values from the entire data sets ($D = D_T \cup D_C$). The difference scores can be obtained from the predicted values as

$$\hat{DS_i} = (\hat{Y}_T - \hat{Y}_C)_i \tag{7}$$

where $i = 1, 2, ...n$.

Customers who have a positive value of $\hat{DS_i}$ are initially considered as the incremental responders as a result of the promotion campaign. However, to determine the final set of incremental responders, further analyses must be performed with the ranked difference scores in decreasing order.

Difference score from a combined model. This model uses a new indicator variable, T_i: $T_i = 1$ for D_T and $T_i = 0$ for D_C and he fit the model to the entire, combined data ($D = D_T \cup D_C$) :

$$Y = X\beta + T\gamma + XT\varphi + \epsilon \tag{8}$$

Then, the difference scores are obtained from the estimated equation:

$$\hat{Y_T} = X\hat{\beta} + \hat{\gamma} + X\hat{\varphi} \tag{9}$$

$$\hat{Y_C} = X\hat{\beta} \tag{10}$$

$$\hat{DS_i} = (\hat{Y}_T - \hat{Y}_C)_i = \hat{\gamma} + X\hat{\varphi} \tag{11}$$

where $i = 1, 2, ...n$.

This is a simplified expression of the difference scores in linear model, but in general they are calculated from the form of conditional expectation of Y:

$$\hat{DS_i} = (\hat{Y}_T - \hat{Y}_C)_i = E(Y|X, T = 1) - E(Y|X, T = 0) \tag{12}$$

where $i = 1, 2, ...n$.

One advantage of this model is that it shows some influential variables directly to the incremental response through the significant parameter estimates [2].

Conclusions. A program has been created using Matlab to build incremental response models for different types of models. Customer data was taken from a database, which contains information on 9000 clients, including their previous records and traits [3]. The incremental response model diagnostics plot in fig. 1 shows both the predicted and observed incremental response rate by decile. The top decile has the highest incremental response rate, with a predicted increment of 21,2 % and an observed increment of 35,2 %. This predicted increment is approximately three times higher than the average incremental response rate of the data (8,3 %).

In fig. 2 response rate for a combined model is shown. The average increment in response rate equals 8,9%. This model also shows predicted increment of 44,4% , which is more than was achieved with previous model.

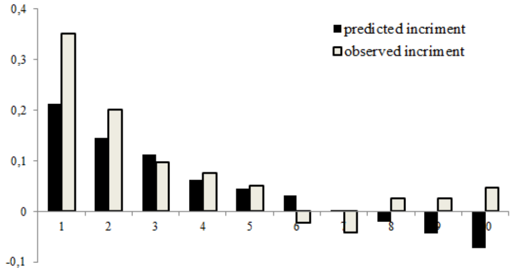

Figure 1. Incremental response model diagnostics plot for separate models. Y-axis shows deciles, while X-axis shows increment rate

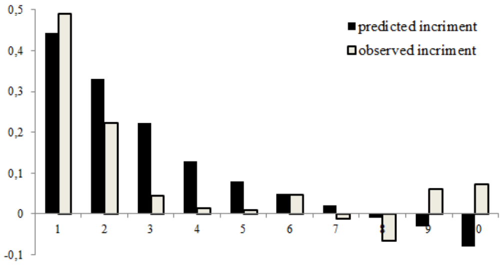

Figure 2. Incremental response model diagnostics plot for combined model. Y-axis shows deciles, while X-axis shows increment rate

Table 1. Example of wide table placed between text paragraphs

Model Name	SSE	Misc	Roc	Gini
Separate models	0.1878	0.2731	0.7106	0.4212
Combined model	0.1962	0.2837	0.6850	0.3701

For both models, all results are in the range of accepted values:

- *Roc* – the higher the value is, the better. $Roc = 1$ means perfect classifier, $Roc = 0,5$ means that the classifier is purely random.
- *Gini* - the higher the value is, the better. $Gini = 1$ means perfect classifier, $Gini = 0$ means that the classifier is purely random.
- *Misc* values range from 0 to 1, where 0 means a perfect classifier.

The first model shows better results in *Roc* and *Gini* as well as lower values of mean-square error and classification error rate. Combined model shows inferior results, which is caused by a large number of predictors which increase the model's complexity. This makes the model more suitable for use with stepwise regression. However, the combined model demonstrates a predicted increment rate much higher than average incremental response rate, which allows to find more potential customers.

References. **1.** Taiyeong L., Ruiwen Z., Xiangxiang M. and Laura R. Incremental Response Modeling Using SAS® Enterprise Miner // Conference SAS Global Forum 2013. - LA.: SAS Institute, 2013. - 13 p. **2.** Lo. V. The True Lift Model: A Novel Data Mining Approach to Response Modeling in Database Marketing, ACM SIGKDD Explorations Newsletter, 2002, 4. – 78–86 p. **3.** Data set and program https://sites.google.com/site/data4mining/datasets/sait_2016_thesis.