Reports of Enlarged Session of the
Seminar of I. Vekua Institute
of Applied Mathematics
Volume 27, 2013

MATHEMATICAL MODELING OF MARKETING PROBLEM

Jokhadze V. Shugliashvili T.

Abstract. In this article we introduce how to use mathematical tools for solving marketing problems. We show the method that can be successfully used for solving problems related to sales and company management.

Keywords and phrases: Mathematical modeling, marketing.

AMS subject classification: 74K25, 74B20.

A power supply company faces increased rivalry on the market and with previous experience it loses 10% of consumers that alter to other providers. The company is operating on the monopolistic competitive market and needs price discrimination. To offer discount the company should figure out those consumers who are least loyal and soon would change to another supplier. The company has some qualitative data of each of their consumers, and also the same data about the last year consumers, including information of consumers who have canceled. One way of detaching current loyal and disloyal consumers is creating risk groups with corresponding probabilities of being a canceler. For this purposes it is necessary to find out the most important variables from the base of the qualitative data. With this chosen variables we make five risk group clusters for previous year's consumers and then divide the current consumer's of 10 000 in the groups that would have the same risks of canceling. The company would then be able to implement the price discrimination policy and "cream" consumers. In our model the variable we are interested to discuss, the dependent variable - "canceler" is a qualitative variable. Thus, we have to solve the problem of having dependent dummy variable. For this problem the Binomial logistic (Logit) model is the best solution. Binary Logistic Regression is a type of regression analysis where the dependent variable is a dummy variable. The "logit" model is built as follows:

$$\ln \frac{p}{1-p} = a + BX + e$$

Where: p is the probability that the event Y occurs, p(Y=1) ,
$\frac{p}{1-p}$ refers to the "odds ratio",
$\ln \frac{p}{1-p}$ is the log odds ratio, or "logit",
a is the coefficient on the constant term,
B is the coefficients matrix of the independent variables,
X is the independent variable matrix,
e is the error term.
The logistic regression model is a non-linear transformation of the linear regression. The "logistic" distribution is an S-shaped distribution function which is similar to the standard-normal distribution (which results in a probit regression model) but easier to work with in most of applications.Logit model showed that with 0,1 confidence level that the following 9 variables are significant:

Important Variables
Type of payment
Quantity of power consumed
Number of trading units in a building
Regional type
Family status
Car density in the region
Index of second hand cars
Insurance type of client
Pharmaceutical Typology: Informed and body conscious

Nine variables are still too many for building up strong clusters, thus additional measures should be taken. To decrease the number of variables in the model more insightful method of Principal Component Analysis is used. This method is for reduction of dimensionality. In other words, we reconstruct data of $x_1, x_2 ... x_n$ into another λ low dimensional points of R set as:

$$f_\lambda - \mu + \upsilon_q \lambda$$

The reconstruction error is minimized by the formula:

$$\min_{\mu, \lambda_{1...n}, \upsilon_q} \sum_{n-1}^{N} \parallel x_n - \mu - \upsilon_q \lambda_n \parallel$$

After making Principal Component Analysis the K-mean cluster is made. The clustering is done according the following formula:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x_j \in S_i} \parallel x_j - \mu_i \parallel^2$$

$x_1 x_n$ are observations, that are expressed in multidimensional real vector; k- is number of sets and $k \leq nS = S_1, S_2 S_k$ μ_i is mean in of points in the S_i set. This model refers to orthogonal linear transformation, which creates principal components. The principals are ordered from the most important to less important ones, the number of possible correlated variables is decreased and the analyzes is not complicated by least important variables. From PCA of 7500 data we receive that the first Principal Component explains 61,1% of variance. First three variables explain 75,35% of the variance. The final result is given on the following Table:

principal component	Percent of Variance Explained
1	0.6166
2	0.6953
3	0.7535
4	0.8080
5	0.8569
6	0.8993
7	0.9387
8	0.9737
9	1.0000

By data processing we build up five clusters, blue is is the first cluster, with red is denoted the second cluster, with green the third cluster, the fourth with grey and the fifth with brown.

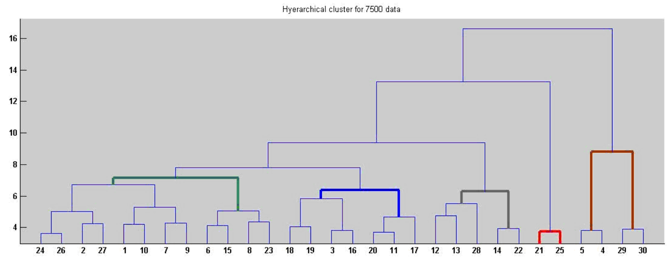

Fig. 1: datatest(2500) hyerarchical cluster and cluster quality 2500

For cluster of 7500 data we receive that silhouettes are almost 0,6 or more for every cluster and the cluster quality is high. The same cluster is built for 2500 data to test whether our model is applicable and gives the same result as already conducted trustful test does. For a sample of 2500, with the same method we get high cluster quality with Silhouettes higher than 0,6. With the same proportions the cluster for the data of 10 000 is built.

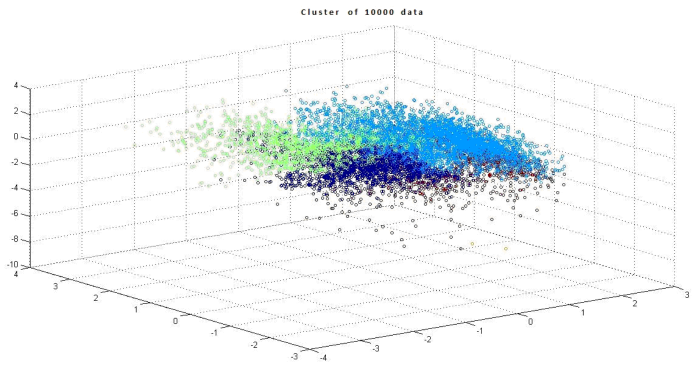

Fig. 2: Cluster for data of 10 000

The quality of the cluster is high and gives following 5 groups:

N	Percentage of consumers in the group from 2500 data	In 7500 data
Risk group 1	0,024	0,0861
Risk group 2	0,0012	0,0004
Risk group 3	0,104	0,0443
Risk group 4	0,79	0
Risk group 5	0,0808	0,0005

Cluster from last year data are built so that, each group has own canceling probability and we spread this probabilities the consumers to leave to the cluster of 10 000. From the first cluster consumers leave with the probability of 9.91% from the second with the probability of 33.33%, from the third with probability of 21.08%, from the fourth 21.08% and from the fifth, 25%. 22.763% of points (consumers) are dropped in the first cluster, in the second cluster 50.897% are of points, in the third risk group 14.9%, in the fourth group 0.026% of points, and, finally, 11.414%of points are dropped in the fifth risk group. We calculate total importance of the risk groups and find out that risk group 2 is most important one.

Risk groups	percent of consumers(10000 data)	probability of canceling	importance
1	22,8%	9,9%	174,660
2	50,9%	33,3%	1313,987
3	14,9%	21,1%	243,309
4	0,0%	9,4%	0,188
5	11,4%	25,0%	221,000

The company does not have to give big discounts to the loyal consumer groups, which are in the second and fourth risk groups, as the probability of them to leave is quiet low. From the data analysis we receive that cancelers will be most likely from the risk group 2, risk group 3 and risk group 5. We should focus on these groups and discount more to them to capture these groups and optimize the profit. While in the risk group 2 more than 50% of consumers are dropped and the probability them to cancel is high this risk group 2 is of high importance and should be paid most of the attention. The company could also discount with certain discount rates to less risky groups as well. Discount to the risk group 4 is not related to high costs while in this group there are only few consumers and the rate of canceling is low. The company should use sample price discrimination marketing strategy if it is more profitable. For instance, the company can use Simple average weighted price strategy. According to this strategy if the firm decides to give importance for P(a) for any risk group consumers and P(b) for another risk group consumers, it discounts according to this importance and for least risky consumer group leaves the previous prices (or almost the same price) and for the rest four groups offers special prices weighted according to risks of P(a) and P(b).

Acknowledgment. The authors are very thankful to Dr. Omar Purtukhia for the useful suggestions.

REFERENCES

1. Field A. Discovering Statistics Using SPSS. *Sage press*, 2009.
2. Lokach L., Maimon O. Data Mining with Decision Trees Theory and Application. *World Scientific*, 2008.
3. Agresti A. An Introduction to Categorical Data Analysis. *John Wiley and Sons INC publication*, 2007.
4. Bliefert F. Developing in Bank Failure Prediction Model for the European Banking Industry. *Aston Business school press*, 2010.

Received 18.05.2013; revised 27.09.2013; accepted 29.10.2013.

Authors' addresses:

V. Jokhadze
Friedrich-Schiller University of Jena
Germany
E-mail: jokhadze.valeriane@gmail.com

T. Shugliashvili
Iv. Javakhishvili Tbilisi State University
2, University St., Tbilisi 0186
Georgia
E-mail: teona.shugliashvili@gmail.com