APPLICATION OF CLUSTERING METHODS IN BANK'S PROPENSITY MODEL

Sergej Sirota – Hana Řezanková

Abstract

Bank's propensity models are being developed for business support. They should help to choose clients with a higher potential (probability) to buy a product. Propensity models are mostly created by using logistic regression. It is a classification task for client's segmentation based on socio-demographics, products and transaction characteristics. A similar classification task can be solved by cluster analysis. Therefore, the aim of this contribution is to improve a propensity model for consumer loan by adding a new variable, which indicates the client's belonging to a certain cluster based on the application of some methods of cluster analysis. We apply two clustering methods. The first one is the *k*-means algorithm, which belongs to centroid-based clustering. The second one is TwoStep cluster analysis proposed for large data sets and hierarchical clustering use. The effect of adding a new variable is evaluated by comparing the total response rate in the current propensity model against the response rate in new propensity models with the "*cluster*" variable.

Key words: logistic regression, cluster analysis, propensity model

JEL Code: C25, C38, D12

Introduction

Bank's propensity models are created for client's segmentation, where logistic regression is mostly used for modeling. Cluster analysis could be also used for classification tasks. As Arpino and Cannas (2016) mentioned, using cluster analysis can improve the success of propensity models, where each model can be built for each cluster. Other authors, Rudolph et al. (2016), also used the combination of cluster analysis and propensity models, where objects in a data set could be clustered by propensity scores. Gawrysiak et al. (2001) mentioned, that using cluster analysis can lead to a more accurate estimation of the dependent variable by the regression function. It is assumed that in the first step, the objects in the data set are divided into clusters using cluster analysis. In the second step, after this division, values of the dependent variable are estimated in each created cluster. Estimated regression parameters can be different for each cluster. There are also other uses of cluster analysis, see e.g. (Kim et al., 2016, Qian and Wu, 2011, Svarc, 2016, Yang et al., 2016). Based on these examples, the idea of this contribution is about application of cluster analysis before creating a bank's propensity model by logistic regression.

The aim of this paper is to improve a propensity model for consumer loan by adding a new variable, which indicates that the client belongs to a certain cluster based on the application of some method of cluster analysis. We suppose a current propensity model for consumer loan in order to predict the probability of buying a product with a certain number of explanatory variables. The success of propensity models is measured by the response rate. In the text below, the principles of logistic regression and selected cluster methods, which are *k*-means and TwoStep, are described. This theoretical part is followed by the results of the analyses. In these analyses, „*cluster*" variables are created by using the above mentioned methods for clustering to the different number of clusters. After these, six new propensity models are built by logistic regression, where the inputs are the same explanatory variables as in the current propensity model and moreover, a new explanatory „cluster" variable (one for each model) is created by cluster analysis. The described process can be considered as combining cluster analysis with logistic regression. For the analyses performance, statistical program *IBM SPSS Modeler* is used.

1 Bank's propensity models

Classification and predictive statistic models have become popular in the banking world with the development of computing technologies. In banking, propensity models are used for client's classification according to their tendency to buy a banking product. One specific propensity model is developed for each bank product. Client's tendency is estimated by the value of the score, which indicates the probability of buying a banking product. The value of the score lies in the interval $\langle 0,1 \rangle$ and it is calculated by logistic regression. After the calculation, clients are divided into 10 groups (there is one division for one product) by values of the score. In the tenth group, there are clients with the highest values of the score in the interval $(0.9, 1\rangle$. The success of client's classification according to their tendency to buy bank's product can be measured by the response rate. It expresses the proportion of clients who really bought the product over the total number of selected clients based on the value of the score. More about modeling can be found in (Linoff and Berry, 2011).

1.1 Logistic regression

Logistic regression for prediction of probability π of buying a product is mostly used for developing bank's propensity models. As mentioned in e.g. (Agresti, 2002), in logistic regression the binary target variable Y is considered, where $Y = 1$ means, that client bought a product and $Y = 0$ means, that client didn't buy any products. Probability that $Y = 1$ can be denoted as $\mathrm{P}(Y = 1) = \pi$. In the classic linear regression, the infinite range of possible values for the target variable is considered. The problem is, that we need a finite range of possible values in the interval $\langle 0,1 \rangle$. It can be passed, if we consider non-linear relationship between a set of explanatory variables **X** and the target variable Y for each object (in our case for each client) in the well-known model of logistic regression:

$$P(Y = 1|\mathbf{X}) = \pi = \frac{\mathrm{e}^{\boldsymbol{\beta}\mathbf{X}}}{1 + \mathrm{e}^{\boldsymbol{\beta}\mathbf{X}}}, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of parameters of the regression function. Estimated parameters can be obtained by the maximum likelihood method.

1.2 Methods of cluster analysis

Relationships between objects in a data set can be analyzed by a cluster analysis. A general process of cluster analysis is about dividing objects into clusters in a way that objects in the same cluster are more similar to each other than objects in a different cluster. There are two general groups of clustering methods according to different ways of clustering. The first one is hierarchical clustering, where in each stage we obtain a different number of clusters. The second one is centroid-based clustering (a centroid is a vector of statistics, e.g. averages or medians, calculated for individual variables from values corresponding to the objects assigned to a certain cluster). For each of these methods, it is considered that one object can be only in one cluster. For analyses proposed in this paper, the *k*-means and TwoStep methods were chosen. For comparing the success of these clustering methods, the silhouette coefficient will be used. It measures clusters quality as mentioned e.g. in (Löster, 2016).

1.2.1 *K*-means method

The well-known method belonging to partition clustering is the *k*-means algorithm. It is determined for objects characterized by quantitative variables. The algorithm needs to define the number of clusters into which objects will be assigned. Finally, each object belongs only into one cluster. The algorithm is iterative; at every step the centroid is calculated for each cluster. The centroid represents the average values of variables for each given cluster. Object's partitioning into clusters is optimal, when the following function is minimized:

$$f_{\mathrm{KM}} = \sum_{h=1}^{k}\sum_{i=1}^{n} u_{ih} \left\| \mathbf{x}_{\boldsymbol{i}} - \bar{\mathbf{x}}_{\boldsymbol{h}} \right\|^2 , \tag{2}$$

where u_{ih} = 1, if the *i*-th object (*i* = 1, …, *n*, where *n* is the number of objects) belongs to the *h*-th cluster (*h* = 1, ..., *k*, where *k* is the number of clusters), otherwise u_{ih} = 0; $\bar{\mathbf{x}}_{\boldsymbol{h}}$ is a vector of average values of variables for each given *h*-th cluster, and $\| . \|$ is the Euclidean distance. The mentioned function (2) is minimized under following conditions:

$$\sum_{h=1}^{k} u_{ih} = 1 \text{ for } i = 1, 2, ..., n , \tag{3}$$

$$\sum_{i=1}^{n} u_{ih} > 0 \text{ for } h = 1, 2, ..., k . \tag{4}$$

1.2.2 TwoStep cluster analysis

For object clustering in a large data set TwoStep cluster analysis is a suitable method, which is a modification of hierarchical clustering methods. This method uses the BIRCH algorithm (Balanced Iterative Reducing and Clustering using Hierarchies). The idea is to create auxiliary clusters, which are clustered in a hierarchical way in the next step. This method uses either the Euclidean distance or the log-likelihood distance. At the beginning of the TwoStep cluster analysis, the number of clusters can be defined, as in the *k*-means algorithm. The range of clusters can be defined in *IBM SPSS Modeler*. More about the BIRCH algorithm can be found in (Zhang et al., 1996).

1.2.3 Silhouette coefficient

The silhouette coefficient combines the concepts of cluster cohesion and cluster separation. Values of the silhouette coefficient are in the interval $\langle -1, 1 \rangle$, where –1 is the worst and 1 is the best assignment of objects in clusters. The basis of this coefficient is the evaluation of assignment of individual objects to "its" cluster. This evaluation for the *i*-th object can be expressed as the value:

$$\psi_i = \frac{\mu_i - \eta_i}{\max\{\eta_i, \mu_i\}}, \tag{5}$$

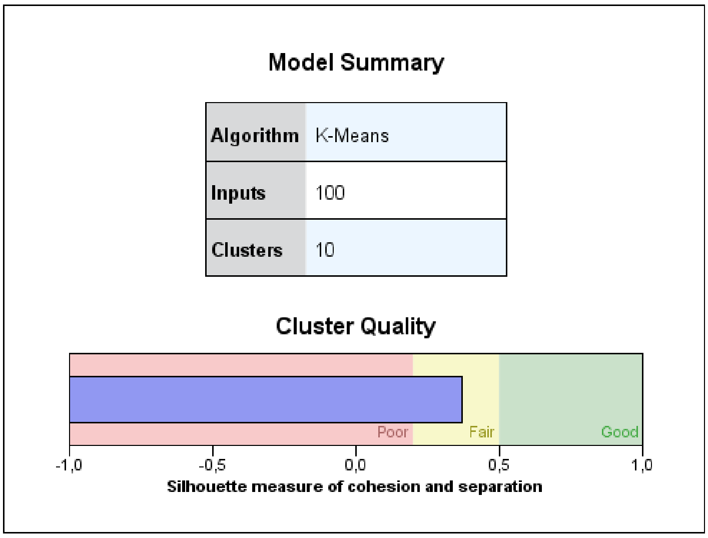
where $\eta_i = \dfrac{\sum\limits_{j \in C_h} d_{ij}}{n_h - 1}$, $\mu_i = \min_{g \neq h}\left(\dfrac{\sum\limits_{j \in C_g} d_{ij}}{n_g}\right)$, d_{ij} is the Euclidean distance between the i-th and j-th objects, and n_h (n_g) is the number of objects in the h-th (g-th) cluster.

The silhouette coefficient is the average value calculated over all objects in a data set, i.e.

$$\psi = \frac{\sum\limits_{i=1}^{n} \psi_i}{n}. \tag{6}$$

An example of graphical presentation of the value is shown in Fig. 1.

Fig. 1: Cluster quality measuring in *IBM SPSS Modeler*



Model Summary

Algorithm	K-Means
Inputs	100
Clusters	10

Cluster Quality

Source: own construction with *IBM SPSS Modeler*

2 Application of clustering methods in a propensity model

In general, a modeling base preparation for a propensity model is very demanding of its size, because it contains thousands of variables representing products and transactional characteristics for hundreds of thousands of clients. In our case, the data set of about 154 113 clients is analyzed. The target variable Y represents information, whether a client bought ($Y = 1$) or didn't buy ($Y = 0$) a consumer loan. Due to the cluster analysis purposes, the best 100 quantitative explanatory variables were selected for the modeling base. These 100 variables were chosen based on the correlation comparison between the target and explanatory variables. This type of variable selection is available in *IBM SPSS Modeler* in „*Feature Selection node*". The reference model for the evaluation of new propensity models will be a model based only on application of logistic regression with the same 100 explanatory variables. It will be named „the current propensity model“ in the following text.

For the creation of new propensity models, in the first step, the new “*cluster*” variables were created with using the *k*-means and TwoStep clustering methods (with the Euclidean distance) with a consideration of 5, 10 and 15 existing clusters (one new variable for each method and type). Cluster analysis was applied on the whole modeling base (154 113 objects and 100 variables).

After the first round of calculations, six new propensity models were built, where the score was calculated based on the application of logistic regression. Explanatory variables for each of the new propensity models were: one “*cluster*” variable from the first step and variables, which were used in the current propensity model. This modeling base was divided into training and testing part in the 70:30 ratio for modeling purposes, where the success of the current and each new propensity model was compared based on the response rate in the testing part. As mentioned above, the response rate is also influenced by the number of selected clients, so for calculations and comparisons, clients with the score value of at least 0.7 will be selected.

The current propensity model has the total response rate of 77.91 % on the selected testing modeling part.

Tab. 1 shows the response rate of new propensity models including „cluster“ variable obtained by the *k*-means algorithm with a consideration of 5, 10 and 15 clusters. It also presents values of the silhouette coefficient, which measures the cluster quality. The highest response rate is 79.25 % in the case with 10 clusters consideration, although the value of the silhouette coefficient is less than in other cases (0.4 vs. 0.5).

Tab. 1: The success of new propensity models combined with the *k*-means algorithm

number of clusters	silhouette coefficient	response rate (in %)
5 clusters	0.5	79.20
10 clusters	0.4	79.25
15 clusters	0.5	79.16

Source: own construction

In Tab. 2, the response rates of new propensity models are shown, which are combined with TwoStep cluster analysis. There is also 5, 10 and 15 clusters consideration in the data set. TwoStep cluster analysis was developed for large data set, but the response rate is worse than using the *k*-means algorithm in these three cases. Even though the response rate is lower, in comparison with the current propensity model, it is slightly higher.

Tab. 2: The success of new propensity models combined with TwoStep cluster analysis

number of clusters	silhouette coefficient	response rate (in %)
5 clusters	0.3	78.49
10 clusters	0.3	78.94
15 clusters	0.2	78.85

Source: own construction

Conclusion

The response rate in the new propensity models is slightly higher after including the „*cluster*“ explanatory variable into the logistic regression. This new variable indicates the client belonging to a certain cluster based on the application of the *k*-means algorithm or TwoStep cluster analysis. Three variants were considered for each cluster method, which varied only in the number of clusters in the data set (5, 10 and 15 clusters). Values of the silhouette coefficient, which measures the quality of the created clusters, were between 0.2 and 0.5 for the introduced types of clustering methods, so it can be considered as a good result. In the used data set, in general, the propensity models combined with the *k*-means algorithm had a higher response rate in comparison with TwoStep cluster analysis. The current propensity model had the response rate of 77.91 %. Based on the highest response rate, the best new propensity model was the model, where *k*-means clustering method was used with consideration of 10 clusters in the data set. The response rate was 79.25 % and the value of the silhouette coefficient was 0.4, which indicates good clusters quality. On the opposite end, the new propensity model with the lowest response rate was the model, where the “cluster” variable was created by TwoStep cluster analysis with the consideration of 5 clusters in the data set. The response rate was 78.49 % and the value of the silhouette coefficient was 0.3. Based on the achieved results, all six newly created propensity models could be put into real deployment, but due to only a slight improvement, there is no reason to replace the current propensity model.

Due to the provided analysis of the use of clustering methods in propensity models created by logistic regression, it is possible to explore further for enhancements.

Acknowledgment

This paper was processed with the contribution of long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). NYC: John Wiley & Sons.

Arpino, B. & Cannas, M. (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine, 35*(12), 2074-2091.

Gawrysiak, P., Okoniewski, M., & Rybinski, H. (2001). Regression – yet another clustering method. In *Intelligent Information Systems 2001*. Physica-Verlag HD, pp. 87-95.

Kim, J. K., Kwon, Y., & Paik, M. C. (2016). Calibrated propensity score method for survey nonresponse in cluster sampling. *Biometrika*, *103*(2), 461-473.

Linoff, G. S. & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. NYC: John Wiley & Sons.

Löster, T. (2016). Determining the optimal number of clusters in cluster analysis. In Löster T., Pavelka T. (Eds.), *10th International Days of Statistics and Economics*. Praha: Melandrium, pp. 1078-1090.

Qian, G. & Wu, Y. (2011). Estimation and selection in regression clustering. *European Journal of Pure and Applied Mathematics*, *4*(4), 455-466.

Rudolph, K. E., Colson, K., Stuart, E. A., & Ahern, J. (2016). Optimally combining propensity score subclasses. *Statistics in Medicine*, *35*(27), 4937-4947.

Svarc, M. (2016). Propensity score matching and alternative approach to individuals matching for counterfactual evaluation purposes. In Reiff, M., Gezik, P. (Eds.), Proceedings of the 2016 *International Scientific Conference on Quantitative Methods in Economics – Multiple Criteria Decision Making XVIII*, pp. 381-387.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., & Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, *72*(4), 1055-1065.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record*, *25*(2), 103-114.

Contact

Ing. Sergej Sirota
University of Economics, Prague, Department of Statistics and Probability
Sq. W. Churchill 1938/4, 130 67 Prague 3, Czech Republic
sirota@centrum.cz

prof. Ing. Hana Řezanková, CSc.
University of Economics, Prague, Department of Statistics and Probability
Sq. W. Churchill 1938/4, 130 67 Prague 3, Czech Republic
hana.rezankova@vse.cz