FACTOR AND CLUSTER ANALYSIS IN CONTEXT OF BANK'S PROPENSITY SCORE MATCHING

Sergej Sirota – Hana Řezanková

Abstract

Propensity score matching is a statistical matching technique, where the probability of an event occurrence is estimated by the score, which is mostly calculated using logistic regression (propensity model). It is classification task with a known number of groups. The aim of this contribution is to improve propensity model by using factor and cluster analyses. Due to cluster analysis and logistic regression reason, we apply a method of the significant variables selection based on the correlation comparison between explanatory variables and the target variable. We also use factor analysis with the Varimax rotation to create new variables to reduce the data set and include these factor variables in the process of matching. After the data set reduction, we apply the *k*-means and TwoStep clustering methods to add new information about the structure of selected objects for propensity score matching. The cluster quality (objects clustering) is compared by the silhouette coefficient. The effect of factor and cluster analyses application is measured by the total success rate of the created propensity model.

Key words: factor analysis, cluster analysis, propensity score matching, silhouette coefficient, success rate

JEL Code: C25, C38, D12

Introduction

In recent years, classification and predictive statistics models have been increasingly used in the banking world with the growth of available data to improve business demands, such as increasing sales rates or decreasing client's churn rate. Each business request needs a separate model. In general, these models could be described in the process of propensity score matching (PSM). It is a statistical matching technique, where the probability of phenomenon (which is defined by the business side) is estimated by the score, which lies in the interval $\langle 0,1 \rangle$. A higher score indicates a higher probability of event occurrence. In practice, the score separates clients into 10 groups with the same score intervals, where clients with the score value of at least 0.7

are selected for business demands (8th–10th group) and, of course, for model evaluation (measured by success rate – it expresses the proportion of clients who really reached observed phenomenon over the total number of selected clients based on the value of the score).

The most widely used statistical method is logistic regression for estimating bank's propensity model. The modeling base is very demanding of its size because we can use thousands of explanatory variables that represent products, transactional and socio-demographic features for hundreds of thousands of clients. More about PSM can be found in (Linoff and Berry, 2011, Huber et al., 2013).

Similarly, as in the paper (Sirota and Řezanková, 2017), the aim of this contribution is to improve the success rate of propensity model for a consumer loan. For this model, it is supposed that the target variable Y represents information, whether a client bought ($Y = 1$) or didn't buy ($Y = 0$) a product. In comparison with the paper mentioned above, this contribution focuses on different approaches of applications of cluster analysis, and moreover on the application of factor analysis. All calculations are performed in statistical program *IBM SPSS Modeler*.

Because of the large modeling base, we apply a method of the significant variables selection based on the correlation comparison between explanatory variables and the target variable. In addition, for some approaches, we use factor analysis to reduce the data set by creating new variables and use these variables as a new input in the process of PSM.

After data set reduction, we apply the *k*-means and TwoStep clustering methods to add new information about the structure of selected objects (clients). The cluster quality is measured by the silhouette coefficient. The effect of new approaches with the application of cluster (or factor and cluster) analyses is measured by the total success rate of created new propensity models against the total success rate of the current propensity model.

1 Approaches introduced in the literature

In the text below, we mention selected articles describing possible uses of factor and cluster analyses for a better predictive capability of the model, which is estimated by logistic regression.

1.1 Application of factor analysis

One of the possible data set reduction is an application of factor analysis. It describes variability among p observed (correlated) variables by f new unobserved factors, where $f \leq p$. The

advantages of the factors are that they are not correlated and retain useful information of original variables. For these reasons, factors can be used as explanatory variables in the model estimated by logistic regression. A detailed description of factor analysis is given e.g. in (Rummel, 1988).

Han et al. (2008) used factors in logistic regression to overcome multiple co-linearity among explanatory variables and meanwhile retain useful information of original variables. Their goal was to improve the predictive capability of the model for forecast firm failure during the financial crisis. The modeling base had 32 variables for 72 firms. The original variables represented financial information (liquidity, profitability, leverage, activity) and the target variable *Y* represented if selected company fell into crisis or not ($Y = 0/1$).

In the first round of calculations, they used factor analysis with the Varimax orthogonal rotation (based on the correlation matrix). The result was a selection of 12 factors that explained 91% of information of original variables. Due to the small number of variables (32), 12 factors had a logical explanation based on the rotated component matrix.

In the final step of calculations, they put all factors as explanatory variables in the following logistic model. The rate of model accuracy achieved 93%, which could be considered a great result. Similar usage can be found in (Kitikidou and Arambatzis, 2012, Manly and Alberto, 2016).

1.2 Application of cluster analysis

Bank's classification task for client's segmentation in PSM is mostly solved by logistic regression. The success rate of propensity model can be improved by using cluster analysis, where relationships between objects are analyzed in a data set. Thanks to cluster analysis we can add new information about the structure of selected objects (clients). In general, the process of cluster analysis is about assigning objects into clusters, where objects in the same cluster are more similar to each other than objects in other clusters.

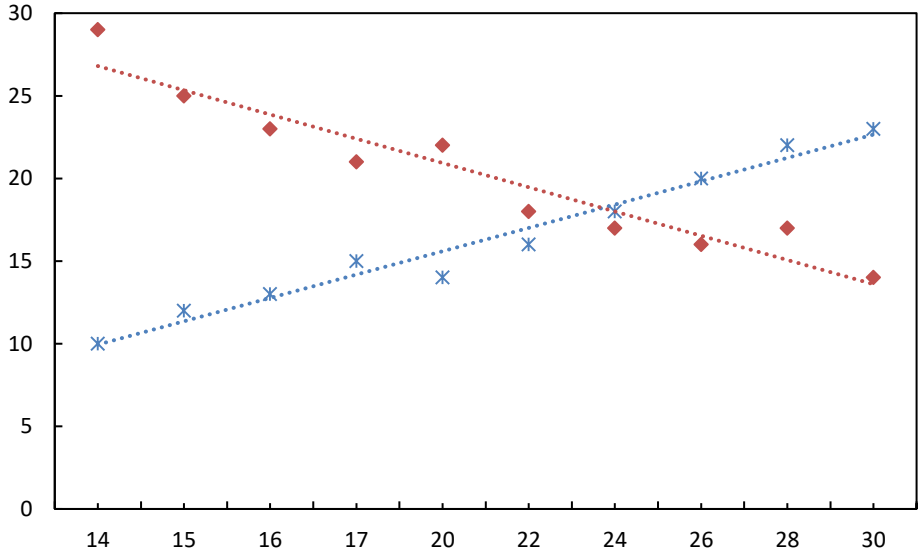
Li et al. (2016) used cluster-based logistic regression model. It could lead to better estimates compared to *the classic* logistic regression model. In the first step, the objects in the data set were assigned into clusters by cluster analysis. After that, values of the dependent variable were estimated in each cluster obtained by logistic regression (for example if we have two clusters, we need two logistic regression models). The first and the second step is iterative until the coefficients of determination are satisfactory.

Gawrysiak et al. (2001) used a similar approach for linear regression, where the determination coefficient was optimized for each cluster. The algorithm was used on a small

data set with consideration of two clusters. In both articles have been reached significantly better performance than using the regression model.

There are also other uses of regression with clustered data, see e.g. (Thoemmes and West, 2011, Leyrat et al., 2014, Arpino and Mealli, 2011, Li et al., 2013). Fig. 1 shows the basic idea of the regression clustering when the original data set has been split into two subsets (the objects are displayed as the blue and orange points) based on optimization of the determination coefficient for each linear regression. As an example, we can imagine a population of fish, when the dependency is investigated between the temperature of water (the x-axis) and fish speed (the y-axis). In population, there are two species of fish, so it seems to be logical to split the population by species. Regression clustering has better performance than the application of linear regression on the entire data set.

Fig. 1: The regression clustering principle



Source: own construction according to Gawrysiak et al. (2001)

More detailed description of logistic regression, the *k*-means and TwoStep clustering methods and cluster quality measuring can be found in (Sirota and Řezanková, 2017).

2 Approaches applied to improve the success rate of the propensity model

Due to large modeling base (154 113 × 2 312 matrix, where 154 113 is the number of clients and 2 312 is the number of variables), the first, we apply a method of the significant variables selection. Because of all variables are quantitative, the selection is based on the correlation

comparison between explanatory variables and the target variable (*Feature Selection node* in *IBM SPSS Modeler*). The best 100 quantitative explanatory variables were selected.

The second, factor and cluster analyses are applied on the whole modeling base, while for modeling purposes the base is divided into training and testing part in the 70:30 ratio.

The third part is about the success rate evaluation of the new propensity models estimated by logistic regression. The reference model (which is named *the current propensity model*) for the comparison is the model, which is based only on the application of logistic regression with 9 quantitative explanatory variables, that are a part of the initial variables selection. The success rate of the current model is 77.91% on the testing modeling part.

The proposed approaches can be divided into two parts. In the first part, cluster analysis and logistic regression are combined. In the second part, factor and cluster analyses and logistic regression are combined.

2.1 Cluster analysis combined with logistic regression

Based on cluster analysis we can use new information about the structure of the modeling base. We used the *k*-means (with the Euclidean distance) and TwoStep (with the log-likelihood distance) clustering methods with consideration of 5, 10 and 15 clusters in the modeling base. New *cluster* variables, which indicate the client's belonging to a certain cluster, were created based on different clustering methods a different number of clusters.

Approach 1

The first approach was described in (Sirota and Řezanková, 2017), where we put the new *cluster* variable and 9 explanatory variables from the current model into the new model estimated by logistic regression. For the reason that six new *cluster* variables were obtained, six models were estimated.

Approach 2

In the second approach, we divided the modeling base according to created clusters (we tried to divide the set of objects into 5, 10 and 15 clusters). For each cluster, the model by logistic regression with the same 9 explanatory variables as in the current model was estimated. Although the input variables for models were the same, this approach created 60 new models (based on 30 clusters obtained by two different clustering methods) because of different regression coefficients. The total success rate for the introduced approach was aggregated from all created clusters.

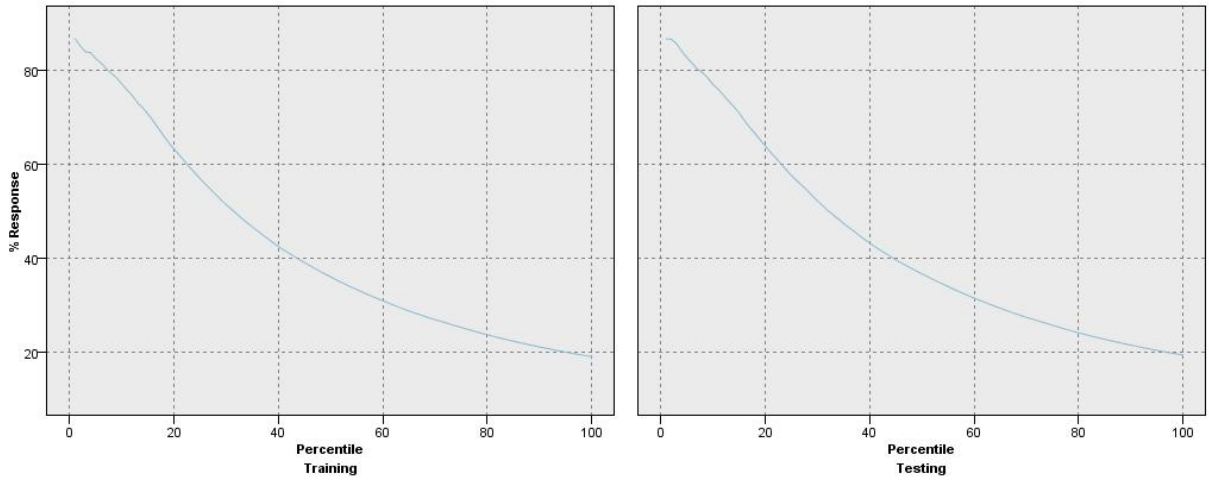
Approach 3

The third approach is very similar to the second with the difference that for each cluster a completely new model with different numbers of explanatory variables was estimated. Explanatory variables were selected from the above 100 variables based on these conditions:

- the significant Wald test of zero values of the regression coefficients (at the 5% significance level),
- the paired Pearson correlation coefficients of explanatory variables less than or equal to 0.8 in the absolute value,
- the course of the success rate curve on the training and testing parts according to Fig. 2.

Fig. 2 shows the desired course of the success rate curve of the bank's propensity model. This model classifies objects in the modeling base by the value of the score. We suppose a training part (the left graph in Fig. 2) and a testing part (the right graph in Fig. 2) for the model evaluation. The x-axis represents objects that are ranked from the highest to the lowest value of the score. The y-axis represents the success rate for each object (the success rate is denoted as the response rate). The course of the success rate is decreasing and it means, that objects with the higher value of score have the higher success rate.

Fig. 2: The course of the success rate curve



Source: own construction with *IBM SPSS Modeler*

This type of application is the most time-consuming of all described approaches. Approaches 2 and 3 were introduced in (Sirota, 2018).

2.2 Factor and cluster analyses combined with logistic regression

In PSM we applied factor analysis for the data set reduction. The analysis was based on the correlation matrix with the Varimax orthogonal rotation and the input variables were standardized. For the next step of calculations, we selected 12 factors that explained 89% of information of 100 original variables.

Approach 0

The method of using factor analysis is the same as in (Han et al., 2008) – we put all (and only) 12 created factors into the logistic regression model and calculate the total success rate.

Approach 1–3

These approaches are very similar to those described in Section 2.1, where we use factors as an input for the *k*-means and TwoStep cluster analysis. Variables selection for logistic regression is still from 100 variables.

3 Results and evaluation

Tabs. 1 and 2 show the success rates of new propensity models for all introduced approaches with exception of approach 0 for which the success rate was 79.2%. It also presents values of the silhouette coefficient, which measures the cluster quality. For each clustering method, green color shows the highest success rate and orange color shows the lowest success rate.

By comparing the success rate among approaches we have achieved the best results in approach 3, because we estimated a new model of logistic regression for each created cluster (explanatory variables were selected from the above 100 variables based on described conditions in Section 2.1). This approach is most similar to the idea of regression clustering.

Application of factor analysis as an input for cluster analysis reached better results with using the TwoStep algorithm (in all three approaches), while it did not reach better results for the *k*-means algorithm.

The values of the silhouette coefficient are in the range of 0.2 and 0.7, which can be considered as a good result (the higher values of the silhouette coefficient were obtained in case when both factor and cluster analysis were applied). Better results were achieved with using the *k*-means algorithm.

The highest success rate is 81.19% (the current propensity model has 77.91%). It is achieved in approach 3 with consideration of five clusters and the *k*-means algorithm without using factor analysis. Interestingly, approach 0 is more successful than the current model (79.2%).

Tab. 1: The success rates of new propensity models combined with cluster analysis

no. of clusters	silhouette coeff.		*k*-means algorithm			TwoStep algorithm		
			success rate (in %)			success rate (in %)		
	k-means	TwoStep	approach 1	approach 2	approach 3	approach 1	approach 2	approach 3
5	**0.5**	**0.3**	79.20	79.57	**81.19**	78.49	78.52	78.48
10	0.4	0.2	79.25	79.74	80.23	78.94	79.10	**77.99**
15	**0.5**	0.2	**79.16**	79.76	79.48	78.85	79.06	**79.54**

Source: own construction

Tab. 2: The success rates of new propensity models combined with factor and cluster analyses

no. of clusters	silhouette coeff.		*k*-means algorithm			TwoStep algorithm		
			success rate (in %)			success rate (in %)		
	k-means	TwoStep	approach 1	approach 2	approach 3	approach 1	approach 2	approach 3
5	**0.7**	**0.4**	78.02	78.51	**80.30**	**78.70**	79.36	79.40
10	0.6	0.4	**78.00**	78.96	79.20	78.86	79.64	79.70
15	**0.7**	0.3	78.43	78.67	78.90	78.89	79.69	**80.10**

Source: own construction

Conclusion

The paper introduced several approaches for increasing the success rate of the bank's propensity model. We used the combination of the application of factor or/and cluster analysis before the model estimation by logistic regression. These approaches were divided into two parts. The difference was based on the inputs for cluster analysis – in the first part it was 100 quantitative variables; in the second one, it was 12 factors. All introduced approaches have reached a higher success rate than the current propensity model. Overall, the highest success rates were achieved in the third approach (the highest success rate of the new propensity model was 81.19% with the *k*-means algorithm (without factor analysis) and consideration of five clusters in the modeling base –in comparison with the success rate of 77.91% in case of the current propensity model had). However, it should be noted, that the third approach was also the most time-consuming.

To confirm the above conclusions, it would be appropriate to apply these approaches on further modeling bases.

Acknowledgment

This work was supported by the University of Economics, Prague under Grant IGA F4/44/2018.

References

Arpino, B. & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770-1780.

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

Han, D., Ma, L., & Yu, C. (2008). Financial Prediction: Application of Logistic Regression with Factor Analysis. In *4th International Conference on Wireless Communications, Networking and Mobile Computing* (WiCOM'08). IEEE, pp. 1-4.

Huber, M., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, *175*(1), 1-21.

Kitikidou, K. & Arambatzis, N. (2013). Factor analysis and logistic regression for forest categorical and quantitative data. *Innova Ciência*, *5*(5), 2-6.

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, *32*(19), 3373-3387.

Li, J., Weng, J., Shao, C., & Guo, H. (2016). Cluster-based logistic regression model for holiday travel mode choice. *Procedia Engineering*, *137*, 729-737.

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

Leyrat, C., Caille, A., Donner, A., & Giraudeau, B. (2014). Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. *Statistics in Medicine*, *33*(20), 3556-3575.

Manly, B. F. & Alberto, J. A. N. (2016). *Multivariate statistical methods: a primer*. CRC Press.

Rummel, R. J. (1988). *Applied factor analysis*. Northwestern University Press.

Sirota, S. (2018). Use of selected clustering methods in bank's propensity model (in Czech). In *Scientific seminar of doctoral studies FIS 2018*. Prague: Oeconomica, pp. 122-127.

Sirota, S. & Řezanková, H. (2017). Application of clustering methods in bank's propensity model. In Löster T., Pavelka T. (Eds.), *11th International Days of Statistics and Economics*, Prague: Melandrium, pp. 1431-1439.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*(3), 514-543.

Contact

Ing. Sergej Sirota
University of Economics, Prague, Department of Statistics and Probability
Sq. W. Churchill 1938/4, 130 67 Prague 3, Czech Republic
xsirs00vse.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