Data Mining and Modeling Use Case in Banking Industry

Stefan M. Kostić, Miloš Đuričić, Mirjana I. Simić, *Member, IEEE*, Miroljub V. Kostić

***Abstract* — Banking industry is very competitive and the market itself has become very saturated in the recent years. Because of this, it is vital for every banking institution to use all available information in order to gain some sort of competitive advantage. In this paper we will demonstrate that it is possible to gain more insight about banking service clients by using data mining and analysis tools, more specifically by using K-Means clustering. Specifically, we will demonstrate that we can define a group of customers that are more prone to purchase housing loan.**

***Key Words* — banking, data mining, K-Means clustering, machine learning**

I. Introduction

IN retail banking, a service tailored to the preferences of each customer is not possible because of time and cost issues. This implies that some degree of standardization is necessary. The trade-off between a more standardized and a more individual service can be made easier by a classification of customers according to multi-dimensional intrinsic characteristics of personality [1]. Companies are trying to segment their customers by identifying groups of persons with need structures that are as homogeneous as possible within each group and significantly heterogeneous between groups [2]. This segmentation can be done by using the data mining technologies. By definition, data mining is the process of discovering patterns in large datasets involving methods at the intersection of machine learning, statistics, and database systems.

Additionally, there is a current need for better understanding of banking clients and their financial needs. This is important since a better client insight can be used as a tool for banking sector. Using that tool, banks can better leverage their business potential by increasing sales, lowering costs and mitigating risks. Intelligent insight into client behavior that belongs to different segments and portfolios can be achieved using mathematic and statistic tools that can be implemented using machine learning.

We can summarize the main aspects of this paper as the following:

- Analyze the real life banking data in order to provide better understanding of clients and their financial needs
- Analyze the business potential of this approach
- Propose future plan of development

There are a lot of new concepts that are beginning to be widely used not only in academic circles, but in the industry as well. Concepts like deep and machine learning, and data science are becoming a must in modern industries because of the rising demands from the client side. In this paper we will present proof of concept of a real life machine learning use case using real life data.

II. Data collection

In order to create a successful prediction model, the first step was to analyze wide spectrum of data that was stored in various databases and different sources. We analyzed several distinct groups of data on client level. Those groups were:

- Sociodemographic data,
- Loan history data,
- Client complaints data,
- Transactional data,
- Current account data,
- Mobile banking data and
- Client employer data.

In total we analyzed more than 50 initial variables. After collecting the initial variables, we created derived variables. The derived variables were necessary to capture the long and short term changes in the initial variables. For example, if we observe one initial variable e.g. client wage, we can calculate the derived variables as following:

1 Mean wage in 12 months period,
2 Mean wage in 3 months period,
3 Absolute change between mean wage in 12 months period and last month wage,
4 Relative change between mean wage in 12 months period and last month wage,
5 Number of wage payments per month, etc.

By using this approach, we created the final dataset that contained approximately 400 derived variables.

In addition to the variables, we generated one separate flag variable. Its values are defined as following: FLAG=1 if and only if client purchased housing loan in the next three months; FLAG=0 if and only if client did not purchase housing loan in the next three months. This FLAG variable will be used to validate our results. We will

Stefan M. Kostić is with the School of Electrical Engineering, University of Belgrade, Belgrade. E-mail: stefankostic08@gmail.com.
Miloš Đuričić is with the Erste Bank a.d. Novi Sad, E-mail: Milos.Djuricic@erstebank.rs
Mirjana I. Simić is with the School of Electrical Engineering, University of Belgrade, Belgrade. E-mail: mira@etf.rs.
Miroljub V. Kostić is with the Statistical Office of the Republic of Serbia, E-mail: miroljub.kostic@stat.gov.rs.

978-1-5386-7171-9/18/$31.00 ©2018 IEEE

demonstrate that we can define a specific segment of clients where the percentage of clients that have FLAG=1is significantly greater than the percentage of clients that have FLAG=1 in entire population.

III. VARIABLE SELECTION

Next important step in the process was the final variables selection. If we skip this step we may have a problem that included variables might be irrelevant to the clusters that we will create. Additionally, variables that are strongly associated might have disproportionate impact on the meaning of the resulting clusters.

It is important to emphasize that, while for regression type model irrelevant variables can be eliminated based on their association with the target variable, in cluster analysis we do not have available target variable. Thus, irrelevant variables should be eliminated before performing any cluster analysis, based on subject matter knowledge or some additional analyses. The subject matter knowledge was used in the initial process, and in the next paragraph we will present the analytical segment of variable selection.

Eliminating redundant variables is also important for successful clustering process. The algorithm that we used attempts to divide a set of variables into non-overlapping clusters so that each cluster can be interpreted as one-dimensional. This means that a single representative of each cluster can be chosen, and the other variables from the same segment can be discarded.

There are various methods for grouping variables in clusters. We used the method that created variable clusters by using the following rule: the variable cluster will be split if it has the largest second eigenvalue after performing eigenvalue decomposition and if that eigenvalue is greater than predefined cutoff value (in our use case cutoff was set to 0.7).

Next issue is to select one variable from a variable cluster. There are many ways to choose a representative from the variable group. For example, using the domain knowledge we can expertly define one of the variables in a cluster as a particularly relevant one. On the other hand, if the variables are being selected as potential inputs into a supervised technique such as regression, selection can be based on the variable in each cluster that is most highly correlated with the target. Finally, the $1-R^2$ ratio can be used:

$$1-R^2=\frac{1-R^2_{Own\_Cluster}}{1-R^2_{Next\_Closest\_Cluster}}, \qquad (1)$$

where R^2 is a statistical measure that's used to assess the goodness of fit of model.

Small values of this ratio indicate that the associated variable has a strong correlation with its own cluster and a weak correlation with the other clusters. In other words, the variable generating the smallest value $1-R^2$ in each cluster is preferred. In our model development process we used $1-R^2$ ratio as the method of choice.

Finally, by using the above mentioned method for variable selection that was implemented in SAS Enterprise Guide software, we created the final variable selection that contained 88 variables.

IV. DATA CLUSTERIZATION

A. Pre-processing

In order to use the variables in the final model we performed standardization of the variables using the formula:

$$X_{standardized}=\ln(X-\min(X)+1), \qquad (2)$$

where X is the variable that will be standardized and *min(X)* is the minimum value of variable X in the dataset.

Additionally, in order to provide us with the possibility to successfully train and test the model, we created separate datasets from our final dataset. For clarity and simplicity reasons, we will introduce a notation that observation where marker FLAG is equal 1 can be called "observation marked 1". Similarly, the observation where marker FLAG is equal 0 can be called "observation marked 0". The rules for creating the separate datasets were the following:

1 In the first step the final dataset was divided into two parts, one contained 70% of observations (dataset A), and the other one contained the remaining 30% (dataset B)

2 We selected all the observations marked as 1 from dataset A and added some randomly selected observations marked as 0 from dataset A. The resulting dataset is defined as training dataset. Since some number of observations that are marked 0 from the dataset A are excluded from training dataset, we can define training dataset as oversampled in comparison to total dataset.

1. The entire dataset B is selected as the testing dataset.

The dataset generation process is graphically described in Figure 1.

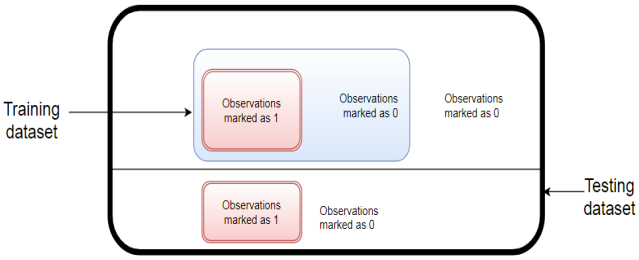

Fig. 1. Graphic explanation of training and testing datasets creation

B. Defining the number of clusters

One of the biggest problems while creating the segmentation model is defining the number of clusters that can be successfully split from the original data. In our work we used (PST2) (introduced by Duda and Hart [3]) to solve this problem.

Pseudo T^2 statistic compares the means between two separate multivariate populations or clusters. The PST2

statistic can be used to determine if two separate clusters should be combined. Larger PST2 values indicate that the two cluster means are different and should not be combined. Conversely, small PST2 values will indicate that the two clusters can be safely combined. The statistic is distributed as an F random variable with v and $v(n_k + n_l - 2)$ degrees of freedom, where v is the number of input variables, n_k is number of observations in cluster k, and n_l is number of observations in cluster l [4].

Any increase in the PST2 value means that the cluster means are increasingly different from each other and should not be joined. The number of clusters just prior to fusions producing an increase is, therefore, a potential solution.

The statistical method for calculating optimal number of clusters was implemented in SAS Enterprise Guide software. The criteria supported an eight cluster solution.

C. Defining the centroids

The next step was to perform the clusterization itself. Using the number of clusters that we defined in previous step, and the training dataset that was defined in step A, we created eight centroids that correspond to the each cluster. This was done using Ward's minimum-variance method of clustering [5]. Ward's method is a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. In our use case, the objective function used was the within-cluster sum of square distances. Ward's method was implemented in SAS Enterprise Guide software.

Using these centroids we performed the segmentation of both testing dataset and final dataset (the entire population).

V. Results

In order to compare the results we generated it was important to select a metric that can be used to check final modal performance. This is important since traditional method of assessing classification accuracy of model cannot be used in datasets where probability of one outcome is very low. For example, let us observe a dataset where 99% of observations are A and 1% is B. If we define a model that marks all observations as A we can be 99% accurate with our predictions. Obviously, this model is not adequate, so there is a need for a different comparison metrics. Our metric of choice was Lift metric [6].

Lift is defined as the ratio of two values: target response divided by average response. We will demonstrate Lift metrics on our previous example; if we define a target group that contains 95% of observations A and 5% percent of observations B. If we calculate the Lift ration we will have:

$$LIFT = \frac{5\%}{1\%} = 5 \tag{3}$$

Now when the comparison metric is selected, we calculated the results for each of the analyzed datasets (training, testing and final (total) dataset).

Table 1 presents the results while analyzing oversampled training dataset. We can see that cluster 4 is distinguished from all other clusters by its large lift value. This means that cluster 4 contains clients that are 4.14 time more likely to purchase housing loan. It is important to notify that, up to this point, the segmentation model has not been allowed to analyze the 1/0 flag which represents that client will/will not purchase housing loan. This means that this is an independent conclusion, based solely on information collected from input variables.

If we observe remaining clusters, we can conclude that cluster 5 represents the near "average" behavior regarding purchasing housing loans. Finally, all remaining clusters demonstrate that the clients they contain are less prone to purchase housing loan in comparison to the average client.

TABLE 1: SEGMENTATION RESULTS – TRAINING DATASET

Cluster	***Percentage of clients***	***Lift***
1	22.1	0.33
2	08.8	0.82
3	12.4	0.83
4	11.7	4.14
5	16.1	1.02
6	12.5	0.24
7	10.2	0.40
8	06.2	0.50

Table 2 presents the results while analyzing testing dataset. Because this dataset is not oversampled, there are small differences in percentages of clients that are member of respective clusters. If we analyze the lift metrics and compare them to the ones presented in Table 1, the overall results are the same – again cluster 4 is the dominant one. Since this dataset is completely new to the model, these results validate our proposed model and verify that our results are significant.

TABLE 2: SEGMENTATION RESULTS – TESTING DATASET

Cluster	***Percentage of clients***	***Lift***
1	24.5	0.62
2	09.1	0.48
3	12.7	0.68
4	07.7	4.50
5	14.9	1.31
6	15.7	0.56
7	07.3	0.90
8	08.1	0.27

Table 3 presents the results while analyzing total available data. The percentages of clients that are member of respective clusters are very similar to the one presented in Table 2. Also, cluster 4 is the dominant regarding the lift metrics and this result confirms that our model does not

create any mistakes when flooded with non-successful (0) observations.

On the other hand, remaining clusters show slight differences in comparison to the data presented in Table 1 and Table 2. The overall conclusion would be the following:

1. Clients that belong in cluster 4 are most prone to purchasing housing loan and should be provided with the corresponding offer via preferred communication channel,
2. Clients that belong in cluster 5 represent average client behavior regarding purchase of housing loan,
3. Clients that belong to all other segments represent below average behavior

TABLE 3: SEGMENTATION RESULTS – TOTAL DATASET

Cluster	***Percentage of clients***	***Lift***
1	24.3	0.52
2	09.3	0.68
3	12.5	0.78
4	07.9	5.16
5	15.3	0.96
6	15.6	0.40
7	07.1	0.69
8	08.0	0.61

VI. CONCLUSION

There are several papers that focus on the client behavior modeling using segmentation, such as [7], [8]. In [7] bank telemarketing data was analyzed and classification method was used as pre-processing activity in order to create under sampled training dataset that was later used as input for various modeling techniques. On the other hand, paper [8] presents customer response model supported by random forests and under sampling algorithm that analyze client demographic information, contact details and socio-economic data.

In this paper we presented a solution that can combine the theoretical knowledge of data analytics and modeling, and a real life problem from banking industry. The results that we generated can be helpful by decreasing the costs in marketing department. Additionally, the presented results can be used in order to increase the efficiency when contacting client. In other words, the number of positive response to number of client contacts generated ratio can be significantly increased.

It is also important to note that the results generated via segmentation can be used in combination with any other client targeting or evaluation method (e.g. [9], [1]) and can hence provide even better results. This paper can also be used as a starting point for further work that can be based on some alternative modeling methods as well as augmented input variable data set.

REFERENCES

[1] A. Machauer, S. Morgner, "Segmentation of bank customers by expected benefits and attitudes," *International Journal of Bank Marketing*, 2001, vol. 19, no. 1, pp. 6-18.

[2] W.R. Smith, "Product differentiation and market segmentation as alternative marketing strategies," *Journal of Bank Marketing*, July 1956, vol. 21, no. 1, pp. 3-8.

[3] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1973

[4] R. Matignon, *Data mining using SAS enterprise miner*, John Wiley & Sons, vol. 638, 2007

[5] J.H. Ward Jr., "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American statistical association*, vol. 58, no. 301, 1963, pp. 236-244

[6] D. S. Coppock, "Why lift?" *Data modeling and mining, Information Management Online*, 2002.

[7] M. Amini, j. Rezaeenour, E. Hadavandi, "A cluster-based data balancing ensemble classifier for response modeling in Bank Direct Marketing," *International Journal of Computational Intelligence and Applications*, 2015, vol. 14, no. 4

[8] V.L. Miguéis, A.S. Camanho, J. Borges, "Predicting direct marketing response in banking: comparison of class imbalance methods," *Service Business*, 2017, vol. 11, no. 4, pp. 831-849

[9] G. Elliott, W. Glynn, "Segmenting financial services markets for customer relationships: a portfolio-based approach," *Service industries journal*, vol. 18, no. 3, 1998, pp. 38-54.