Unsupervised ranking of clients: machine-learning approach to define a “good customer”

Uladzimir Parkhimenka, Mikhail Tatur, Olga Khandogina

***Abstract*—**Ranking of clients is a natural problem for every business. Though usually it can be solved by common sense and intuition of managers, in the case of a big business entity (e.g. global online stores), the problem becomes more complicated with obvious obstacles in derivation of fast and accurate solution. This article deals with the clients ranking problem using machine learning methodology.

***Keywords*—**loyalty ladder; ranking; ecommerce; automatic marketing decision-making; machine learning; data mining & knowledge discovery; latent variable analysis.

I. Introduction

Every company, even the biggest one, faces the problem of limited resources, which is the central economic problem in principle. For example, in a product strategy a company has to decide what product portfolio composition should be optimal, what products or models to produce (and in what quantities), what products or models to discontinue, while there is obvious limitation of production capacity. In an investment strategy, a company deals with defining priorities of assets to invest in, given a limited budget of financial resources.

The same problem a company has in relation to a pool of clients. There is a need to know what clients are worth of marketing activities and efforts and what aren’t. Such knowledge could be used for tailoring individual loyalty program for each client depending on his/her ‘worth’ and ‘quality’; introduction of personal managers for ‘good’ clients; cutting down marketing efforts and costs on ‘bad’ clients; introduction a queue policy in provision of services, etc.

Furthermore, a company could not need only a qualitative solution here (‘bad’ or ‘good’ client), but a quantitative one. A company needs to prioritize its clients, thus making a ranking.

Ranking is sorting of objects in accordance to one or several criteria. The result of ranking is a hierarchy: from the ‘best’ (biggest, most valuable, tallest, etc.) to the ‘worst’ (smallest, least valuable, lowest, etc.). Such an approach is widely used in the society, e.g. a music chart or university rankings.

Ranking is about comparison of several objects with one another; it reveals preferences of a decision-maker and is closely related to choice modelling.

In the case of objects with many characteristics, ranking is a way of simplifying complex relationships between objects. It could result in losing of some information about these objects, but gives a very clear and simple ‘picture’ as a set-off against this lost.

In the business literature, there are many approaches to the problem of clients ranking, usually in the form of assigning clients to several (not too many) categories based on the history of their interactions with a company. For example, companies distinguish a few ‘platinum’, ‘gold’, VIP-, key account-, top-clients contrary to a majority of ‘other’ or ‘normal’ clients.

The key question here is what exactly constitutes the notion of a ‘good’ and ‘bad’ client for a company. There is no definite answer, though financial (e.g. sales volume or profit), indicators seem to be main contributors here, while strategic goals (e.g. competitive and positioning strategy) are also used, but play less significant role.

This article looks at the perspectives of using machine-learning methodology to solve the problem of clients ranking in unsupervised manner: without any expert opinion and learning sample.

II. TRADITIONAL APPROACHES TO CLIENTS RANKING

In the business literature, there is no any universally recognized method for ranking of clients. Though the necessity of distinguishing 'better' client over 'worser' ones is widely taken for granted, e.g. in the framework of the well-known key account management (KAM) methodology or within the concept of loyalty ladder, there is a variety of different approaches to the problem.

Pareto Law states that 20% of clients generate 80% of all sales volume (or profit) [1]. Therefore, the natural consequence of such a statement is to divide clients in two groups (in proportion approximately 20:80) and focus main efforts on the smaller, but more 'worth' part.

ABC analysis, another very popular in the business literature approach to the problem, could be considered as a natural follow-up of the Pareto Law. Its primary focus is inventory [2], but clients can be and are analyzed as well [3, p. 15]. The main idea of this approach is division of all objects into three groups in accordance to their share of sales volume or profit or other outcome (e.g. value added). The *A* group consists of objects that generate approximately 80% of the outcome. To the *B* group belong objects that generate next 15% of the outcome. The *C* group comprises objects that generate the last 5% of the outcome. Theoretically expected that the size of *A* group would be approximately 20%, *B* – 30%, *C* – 50% of the whole clients population.

There are several extensions of classical variant of ABC analysis, for example multi-criteria analysis [4].

XYZ analysis is considered to be another one, dynamic extension of the static ABC analysis [5]. This approach is oriented in defining what objects (usually inventory, but in this case – clients) show stable 'behavior' over the period of time and what don't. The 'stability' is measured by statistical variance.

RFM analysis is a popular and traditionally incorporated in CRM systems approach [6, 7, 8]. It aims not directly on ranking, but on clustering of existing clients in accordance to three dimensions: Recency – time from the last purchase, Frequency – number of past purchases during certain period (e.g. year or quarter), and Monetary – volume of purchases during the same period. Nevertheless, it gives some insights what clients are 'better' than others, at least within a specific cluster.

Customer Profitability Analysis (CPA) [9] focuses on measuring profitability of a customer as a main indicator for policymaking. Its logical successor, Customer Lifetime Value approach (CLV) [10, 11], aims at calculating the net present value of a single client (customer) given expected average purchase volume, its frequency, duration of customer "lifetime", cost of capital (discount rate). Ranking is a natural consequence of this analysis (CPA or CLV), though it is never meant explicitly.

III. 'GOODNESS' OF A CLIENT AS A SYNTHETIC LATENT VARIABLE

If we look to the essence of the above-mentioned approaches and switch from the 'business language' to a formal one, we could generalize all approaches as a simple weight-and-sum ranking system:

$$Score_i = \sum_{j=1}^{n} w_j \cdot x_{ij} \quad (1)$$

where $Score_i$ – an estimate of the 'worth' ('quality') of i-th client that is used further for sorting from maximum to minimum to get a final ranking;

w_j – a 'weight' of the j-th characteristic of a client;

x_{ij} – a value of the j-th characteristic of a i-th client;

n – number of characteristic of a client taken into account for ranking.

For simplicity, we skip other two approaches to specify a ranking model (pairwise and listwise).

In traditional methods (see section II of this article) viewed from the perspective of (1), in order to get final ranking one has to specify: (i) set of *n* characteristics of a client (observable, measurable variables) as partial indicators of a client's 'worth' and 'quality'; (ii) values of w_j as 'weights' and 'importance' of a specific characteristic.

In general, characteristics of clients can be of a great variety, e.g. in the case of online stores visitors:

– purchase volume in monetary terms during the specified period;
– purchase volume in-kind during the specified period;
– profit generated by a client during the specified period;
– number of previous visits during the specified period;
– expected Customer Lifetime value;
– RFM-metrics;
– behavioral traits (e.g. number of pages viewed at online store or number of goods put in a 'basket');
– engagement metrics (e.g. number of 'likes' or 'reposts' of relevant content in social media);
– etc.

The chosen set of characteristics and their 'weights' reflects experts' understanding of situation and preferences of owners and top managers of a company through.

Variable $Score_i$ means the 'worth' ('quality') of the *i*-th client. This is unobservable variable, that can not be measured directly.

In statistical terms, it is rational to use latent variable approach [12] or latent synthetic categories approach [13]. We treat both approaches as equivalents. Both of them are oriented toward constructing of a directly unobservable variable as a function of directly observable variables. Formula (1) fits this definition in full.

Usage of latent variable approach allows to make rankings without knowing preferences of experts and without any learning sample.

IV. PROCEDURE OF UNSUPERVISED RANKING OF CLIENTS

Latent variable ranking of clients should include several steps.

Step 1. Defining a set of variables (characteristics of clients). This can reflect specific preferences of decision-makers (e.g. profitability over sales volume) or can be just a set of all possible variables tracked by an information system of a store relevant directly or indirectly to 'worth' or 'quality' of a client. Some examples are given in the Section III of the article.

Step 2. Getting data. It means getting *n*×*m* values x_{ij} for all *m* clients with *n* characteristic.

Step 3. Scaling variables to the interval [0, 1]. This transformation can be done by classical approach:

$$z_{ij} = \frac{x_{ij} - \min\limits_j x_{ij}}{\max\limits_j x_{ij} - \min\limits_j x_{ij}} \quad (2)$$

In the case of reverse relationship between a variable and 'worth' and 'quality' of a client (e.g. number of returns of goods), should be used another formula:

$$z_{ij} = \frac{\max\limits_j x_{ij} - x_{ij}}{\max\limits_j x_{ij} - \min\limits_j x_{ij}} \quad (3)$$

Such an approach does not take into account occurred statistical distribution and possible data outliers, but a data analytic should be aware of that and make corrections if needed.

Step 4. Making a principal component analysis. This would give a set of principal component scores (client's 'coordinates' in principal components space) – s_{il}, where *i* denotes the *i*-th client, and *l* denotes *l*-th principal component.

Step 5. Defining the number of principal components for ranking function. Following the approach of Aivazian [13], we use 55% as a threshold for the variance explained by the first principal component.

If this threshold is not achieved, the second, third and other principal components should be taken into construction of the ranking function until the explained variance goes beyond 55%. In formal way, this implies the inequality:

$$\frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_k+\cdots+\lambda_p} \geq 0.55 \quad (4)$$

where λ_l – eigenvalue of the *l*-th principal component;

k – number of first principal components which should be taken into construction of the ranking function;

p – number of principal components obtained.

Step 6. Specifying the ranking function. Latent variable which reflects the 'worth' and 'quality' of a client is now can be defined as a weighted (by principal components eigenvalues) sum of first *k* principal component scores

$$\widehat{Score}_i = \sum_{l=1}^{k} \lambda_l \cdot s_{il} \quad (5)$$

Thus, score weights are being determined by principal component analysis algorithm without a need of expert opinions on the weights.

Step 7. Calculating the values of the ranking function. The function (5) is used to get $Score_i$ for each client ($i \in [1, n]$).

Step 8. Sorting. All clients are sorted by $Score_i$ from maximum to minimum.

V. SYSTEM OF AUTOMATIC ONLINE CLIENTS RANKING

Proposed approach could be implemented within automatic marketing decision-making in an online store. The author's point of view on this issue is presented in Figure 1.

Clients' actions are stored in a database and used for constructing (calculating) a set of variables (characteristic) that indicate 'worth' and 'quality' of a specific client (see Section III). These variables are scaled to the interval [0, 1] and processed by the principal component analysis algorithm. In accordance to the proposed in Section IV, procedure of ranking number of principal components is defined (based on the needed level of explained variance).

Finally, calculation of ranking values is made with subsequent sorting of clients.

In line with the ranking results, a marketing strategy for each client is designed and implemented (e.g. automatically defined amount of discount).

The efficiency of ranking and related marketing strategies should be assessed (e.g. clients loyalty measures over time, stability of ranking results, etc.) and corrective measures, if needed, should be introduced. Strictly saying, such correction is only possible in (i) redefining a set of variables, (ii) changing the needed level of explained variance and (iii) the logic of designing a marketing strategy based on the ranking results. It is an open question whether all of this can be done automatically in principle without intervention of experts.

VI. CONCLUSIONS AND FURTHER WORK

In the article, an approach to unsupervised (without experts and learning sample) clients ranking has been proposed. The approach is based on the known latent variable analysis. Under such a latent variable the ‘worth’ and ‘quality’ of a client is considered.

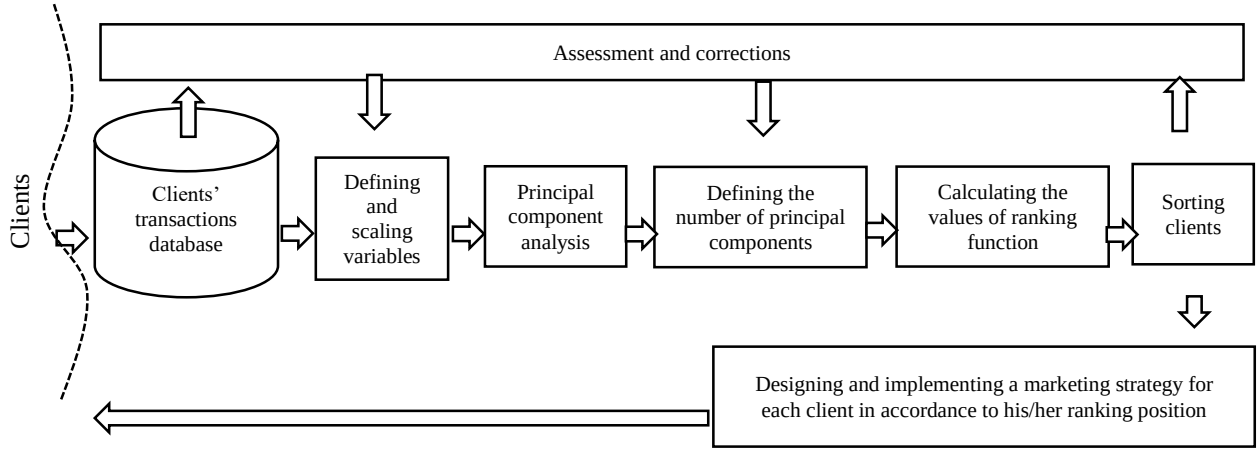

Fig. 1 Conceptual model of the system of automatic clients ranking

Besides this, the conceptual model of the system of automatic clients ranking in the framework of automatic decision-making in online stores has been introduced.

The main question of further research whether it is correct in principle to synthesize an estimate of ‘worth’ and ‘quality’ of client from a set of partial indicators using not expert judgments on real-life importance of indicators, but mathematical properties of the data.

The second question whether ranking based on several principal components (not a single one, the first component) has a business meaning at all, because having two or more principal components means a set of latent variables.

Till now the proposed conceptual model is no more than a concept and it’s still open whether it can be implemented in real context of an online store. As well as it’s not theoretically clear (i) how a marketing strategy can be tailored in accordance to ranking results, (ii) what efficiency measures could be used in the system logic.

Finally, there is an urgent need to test the proposed approach on real datasets.

VII. ACKNOWLEDGEMENTS

The authors thank two anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

REFERENCES

[1] Klepacki, B., and I. Dziedzic-Jagocka, "Stock Management with the example of the production/logistic company Vive Textile Recykling", Аграрна економіка, Volume 3, № 1-2, 2010, pp. 120-130.

[2] M. Karthick, S. Karthikeyan, and M.C. Pravin, “A Model for Managing and Controlling the Inventory of Stores Items based on ABC Analysis”, Global Journal of Researches in Engineering, Volume 14, Issue 2, 2014.

[3] B. Noche, “ABC-/XYZ Analysis Introduction”, https://www.uni-due.de/imperia/md/content/tul/download/en_ss2015_lm01_le_abc_analysis.pdf

[4] Flores, Benito E., and D. Clay Whybark, "Multiple criteria ABC analysis", International Journal of Operations & Production Management, 1986, 6(3), pp.38-46.

[5] Scholz-Reiter, B., Heger, J., Meinecke, C. and Bergmann, J., “Integration of demand forecasts in ABC-XYZ analysis: practical investigation at an industrial company”, International Journal of Productivity and Performance Management, 61(4), pp.445-451.

[6] Khajvand, M., Zolfaghar, K., Ashoori, S. and Alizadeh, S., "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study", Procedia Computer Science, 2011, 3, pp.57-63.
[7] Chang, H.C. and Tsai, H.P., "Group RFM analysis as a novel framework to discover better customer consumption behavior", Expert Systems with Applications, 2011, 38(12), pp.14499-14513.
[8] Hosseini, S.M.S., Maleki, A. and Gholamian, M.R., "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty", Expert Systems with Applications, 2010, 37(7), pp. 5259-5264.
[9] Foster, G., Gupta, M. and Sjoblom, L., "Customer profitability analysis: challenges and new directions", Journal of Cost Management, 1997, 10, pp.5-17.
[10] Berger, P.D. and Nasr, N.I., "Customer lifetime value: Marketing models and applications", Journal of interactive marketing, 1998, 12(1), pp.17-30.
[11] Jain, D. and Singh, S.S., "Customer lifetime value research in marketing: A review and future directions", Journal of interactive marketing, 2002, 16(2), pp.34-46.
[12] Guarino, C., Ridgeway, G., Chun, M. and Buddin, R., "Latent variable analysis: A new approach to university ranking", Higher Education in Europe, 2005, 30(2), pp.147-165.
[13] Aivazian, S., "Synthetic indicators of quality of life: Construction and utilization for social-economic management and comparative analysis", Austrian Journal of Statistics, 2005, 34(2), pp.69-77.