DOI: 10.17323/2587-814X.2023.1.7.17

Predicting customer churn based on changes in their behavior patterns

Yury A. Zelenkov [a]
E-mail: yzelenkov@hse.ru

Angelina S. Suchkova [b]
E-mail: assuchkova_1@edu.hse.ru

[a] HSE University
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

[b] HSE University – Saint Petersburg
Address: 16, Soyuza Pechatnikov Street, Saint Petersburg 190121, Russia

Abstract

Customer retention is one of the most important tasks of a business, and it is extremely important to allocate retention resources according to the potential profitability of the customer. Most often the problem of predicting customer churn is solved based on the RFM (Recency, Frequency, Monetary) model. This paper proposes a way to extend the RFM model with estimates of the probability of changes in customer behavior. Based on an analysis of data relating to 33 918 clients of a large Russian retailer for 2019–2020, it is shown that there are recurring patterns of change in their behavior over a single year. Information about these patterns is used to calculate the necessary probability estimates. Incorporating these data into a predictive model based on logistic regression increases prediction accuracy by more than 10% on the metrics AUC and geometric mean. It is also shown that this approach has limitations related to the disruption of behavioral patterns by external shocks, such as the lockdown due to the COVID-19 pandemic in April 2020. The paper also proposes a way to identify these shocks, making it possible to forecast degradation in the predictive ability of the model.

Keywords: customer churn, customer churn prediction, RFM model, RFM model extension, customer behavior patterns, predictive analytics

Citation: Zelenkov Y.A., Suchkova A.S. (2023) Predicting customer churn based on changes in their behavior patterns. *Business Informatics*, vol. 17, no. 1, pp. 7–17. DOI: 10.17323/2587-814X.2023.1.7.17

Introduction

The concept of customer relationship management (CRM) implies the acquisition and retention of the most profitable customers based on an understanding of their values and the motives that determine behavior [1]. The costs of retention are much lower than those of attracting new customers, and the loss of a customer means the loss of all purchases that he could make during the life cycle [2]. Since not all customers are equally attractive to the company financially, it is extremely important to first determine their profitability, and then appropriately allocate resources to retain them [3].

The problem of optimizing customer retention costs is solved in two stages: the first is customer segmentation, and the second is the prediction of changes in their behavior. For segmentation, clustering methods are usually used, which allow us to divide a set of clients into internally homogeneous groups (classes) which at the same time differ greatly from each other [4]. The goal of models that solve the second problem is to identify customers who can change their group, for example, move from the class of active buyers to the class with low purchasing [5]. This approach is called customer churn prediction. This problem can be reduced to a binary classification problem: using customer data for periods 1, ..., t, train a classifier h that predicts the probability that in period $t + 1$ a customer will remain in the same group, move to a group that generates more income (class label 0), or move to a group with an aggregate lower income (class label 1). Based on these forecasts, companies develop differentiated marketing strategies aimed at retaining customers belonging to class 1 [2].

The most widely used approach to the analysis of customer behavior is the RFM model, which combines data on the time passed since the last activity of the customer (Recency), the number of his purchases for the period t (Frequency) and the total amount of money spent (Monetary) for the same period [6]. According to the traditional approach, the customer database for each of the three RFM dimensions is divided into five equal segments (quantiles). The top 20% of customers get label 5, the next 20% get label 4, and so on. As a result, each customer is associated with a label containing three numbers corresponding to quantiles according to RFM measurements, for example, 534 or 231. Thus, 125 groups of customers with potentially different behavior are allocated. It is obvious that this approach has drawbacks, since it does not guarantee that the selected groups, firstly, are internally homogeneous, or secondly, that they differ greatly from each other. Therefore, recently cluster analysis methods (k-means, self-organizing Kohonen maps, etc.) are more often used; they allow us to divide the customer database using formal metrics [7].

Many authors consider variations of the RFM model, expanding it with additional dimensions, including using the dynamics of customer behavior. For example, a model that considers discounts is proposed in [8], and the duration of a client's stay in a certain cluster is considered in [9].

In the context of the task of predicting the outflow of customers, the discovery of patterns that describe stable trajectories of consumer movement between clusters is one of the most important areas of research. For this, various dynamic models based on pattern identification with such methods as clustering [10, 11], association rules [12], and hidden Markov models [13] are used.

The purpose of this study is to develop a method that allows us to predict a change in customer behavior (i.e., their movement from one class to another) considering information on consumer flows between groups accumulated over previous periods. According to our hypothesis, the intensity of the transition of clients from one class to another varies throughout the year, but there is a stable pattern that repeats from year to year. Thus, we can consider the observed rates of flows of clients from class to class as estimates of the probability that the client will leave the cluster in which he is currently located. The inclusion of this information in the predictive model should significantly increase the accuracy of the prediction.

1. Data and task formalization

To test the formulated hypothesis, a data set was used that includes information on purchases made by customers of a large Russian retail chain in 2019–2020. The data was consolidated monthly, and for each customer the following values were calculated:

- R – the number of days that have passed since the last purchase until the beginning of the current month,
- F – the number of purchases in the current month,
- M – is the total amount of the client's expenses (retail network income) in the current month.

Other data often used in customer churn forecasting tasks (age, gender, marital status, etc.) were not used in this case, since they contain a significant number of gaps and unreliable values. After removing outliers and incomplete data, a set was obtained containing information about 33 918 customers who made at least five purchases in two of the studied years.

At the next stage, stable user clusters were identified. To do this, the data were divided into periods of one month and grouped as follows. Customers who did not make purchases in the month under

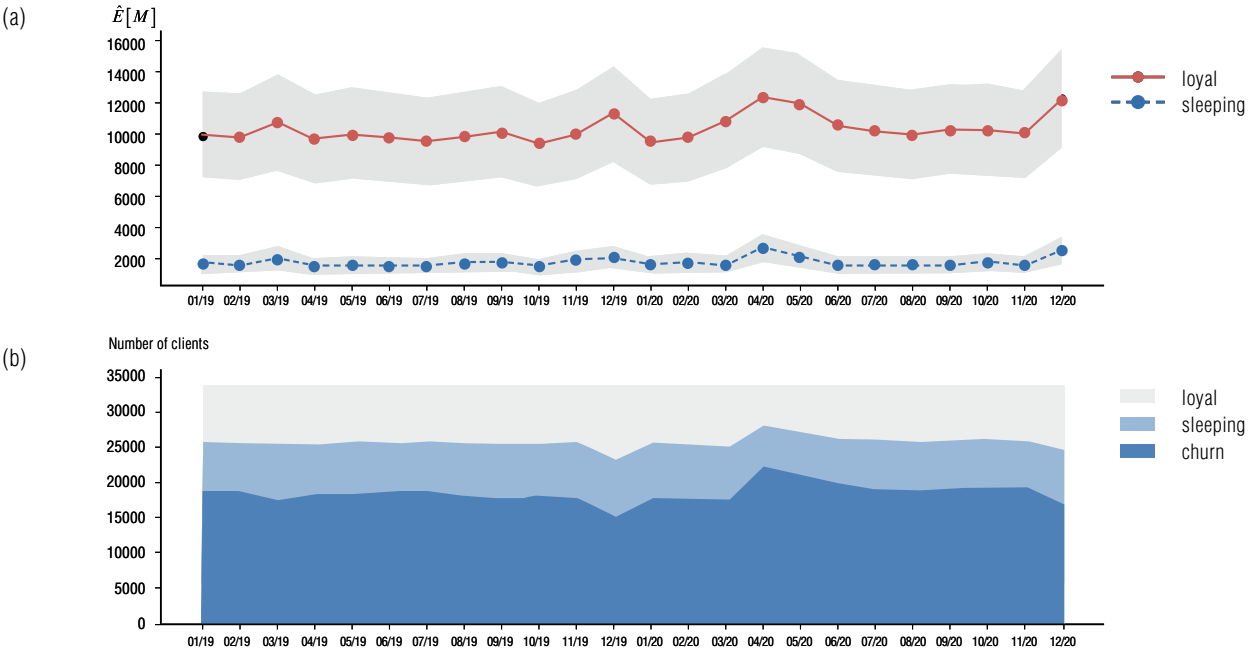

Fig. 1. (a) – average income and 95% confidence interval for clusters of customers who make purchases; (b) – change in the number of clients by clusters.

consideration (obviously, for them $F = 0$ and $M = 0$) were considered to belong to the same cluster and were excluded from the analyzed data set; customers with $F > 0$ and $M > 0$ were segmented using the algorithm k-means. The quality of the resulting separation for a different number of clusters was evaluated using the silhouette score [14]. The maximum values were obtained for splitting into three clusters (one including customers without purchases in the current month, and two identified by the k-means algorithm among customers with purchases). The average value of the silhouette score for 24 months is 0.708, the minimum is 0.649 for December 2019 and the maximum is 0.779 for April 2020.

Figure 1(a) shows the change in the average retail chain income $\hat{E}[M]$ from a customer and the corresponding 95% confidence interval for two clusters of buyers making purchases. A cluster that gives a higher average income brings together buyers who can be defined as "loyal," while a cluster of customers that generate less income can be called "sleeping". The third cluster, which unites customers without purchases, we called "churn". *Figure 1(b)* presents the change in the number of clusters.

Analyzing the presented graphs, we can conclude that there are certain patterns that repeat from year to year. For example, in December, the size of the loyal cluster increases and the income generated by it increases simultaneously; this is associated with seasonal holidays (New Year and the corresponding holidays).

An analysis of the significance of exogenous variables was also performed using a method based on measuring the decrease in the accuracy of the model when mixing the values of the attribute of interest (permutation importance [15]). Logistic regression was used as the base classifier because this model is robust to perturbations, and the accuracy of the model was estimated using the area under the receiver operating characteristic curve (AUC) because this metric is not sensitive to class imbalance [16]. The results obtained show that the most significant features (in descending order of significance) are R_t, M_t, F_t, preceding the predicted period $t + 1$, i.e., the process of changing customer behavior under study is Markovian. Data from earlier periods does not affect the quality of the prediction. This is consistent with the results of other researchers [4, 17, 18].

Thus, the formulation of the customer churn prediction problem can be refined as follows. Let X be the set of customer descriptions and $Y = \{0, 1\}$ the set of class labels. We must build an algorithm h: $X \to Y$, capable of classifying an unknown object $x \in X$ according to a known finite training sample $D = \{[R_t, M_t, F_t]_1, y_1, \ldots, [R_t, M_t, F_t]_m, y_m\}$, where $y \in Y$, and the vector $(R_t, M_t, F_t) \in X$ is a feature description of the object.

2. Modeling customer flows

As noted above, many researchers focus on expanding the feature vector including additional features in it, which improves the accuracy of the classification algorithm. In this paper, it is proposed to use information about the dynamics of customer flows between clusters.

This idea was inspired by epidemiological models (EM), which consider the movement of people between different groups: susceptible, infectious, recovered, etc. [19]. The intensity of movement from group A to group B is determined by the transition rate α^{AB}, which determines the proportion of members of group A who moved to B. In most EMs, these rates are considered as exponentially distributed random variables, but in our case, their exact values can be calculated, since the number of all groups is always known. The second difference is that EMs usually assume a limited number of possible trajectories of movement between groups; in our case, the client can move from his group to any other. Considering all the above, the dynamics of customers can be represented by the following difference equations:

$$L_{t+1} = L_t + \alpha_{t+1}^{SL} S_t + \alpha_{t+1}^{CL} C_t - [\alpha_{t+1}^{LS} + \alpha_{t+1}^{LC}] L_t$$
$$S_{t+1} = S_t + \alpha_{t+1}^{LS} L_t + \alpha_{t+1}^{CS} C_t - [\alpha_{t+1}^{SL} + \alpha_{t+1}^{SC}] S_t$$
$$C_{t+1} = C_t + \alpha_{t+1}^{SC} S_t + \alpha_{t+1}^{LC} L_t - [\alpha_{t+1}^{CL} + \alpha_{t+1}^{CS}] C_t.$$

Here L_t, S_t, C_t are the number of clients in the "loyal", "sleeping" and "churn" clusters, respectively, at time t;

α_{t+1}^{AB} is the flow coefficient that determines which part of the clients who are in group A at time t will move to group B in period $t + 1$.

The index $t + 1$ in this case means that the value of this coefficient will become is known only after the moment $t + 1$.

The coefficient α_{t+1}^{AB} can be calculated as

$$\alpha_{t+1}^{AB} = F_{t+1}^{AB} / A_t,$$

where F_{t+1}^{AB} is the number of clients (flow) who moved from group A to B in the interval between t and $t + 1$; is the number of clients in group A at time t.

A comparative analysis of the flow coefficients is shown in *Figs. 2–4*. The top graphs in each figure represent the values of the coefficients for the months of 2019 and 2020 (in this case, the subscript of the variable means the year), as well as the difference between them. Areas shaded in different shades represent different seasons (winter, spring, summer, and autumn). It can be seen that there is no seasonality associated with the time of year. The vertical dotted line corresponds to December, when, as noted above, there is an increase in the number of purchases. It follows from these graphs that the values of the coefficients for different years are quite close, and the difference between them tends to zero (marked with a horizontal dotted line). The only exception is observed in April. This is because a lockdown was introduced in April 2020 due to the COVID-19 pandemic, which led to a decrease in the activity of buyers.

The bottom pair of graphs in each figure shows the distribution of the values of the respective coefficients for both years under study, as well as the kernel den-

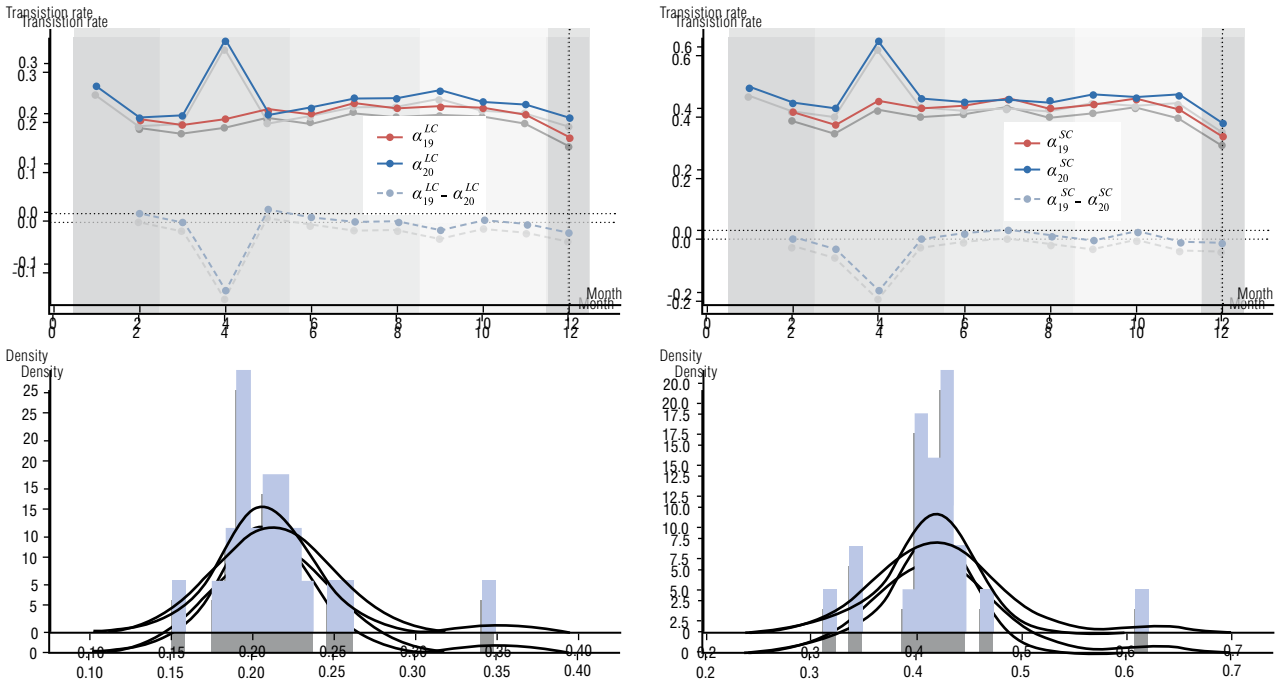

Fig. 2. Flow coefficients to cluster C (churn).

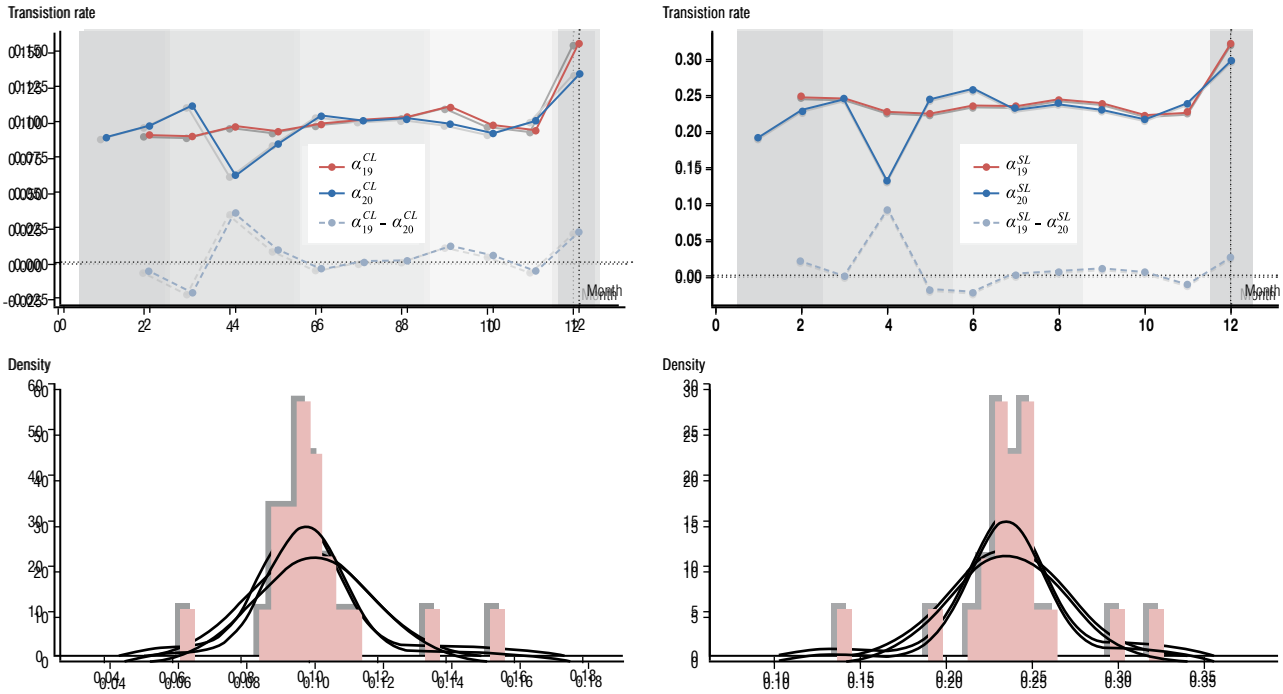

Fig. 3. Flow coefficients to the cluster L (loyal).

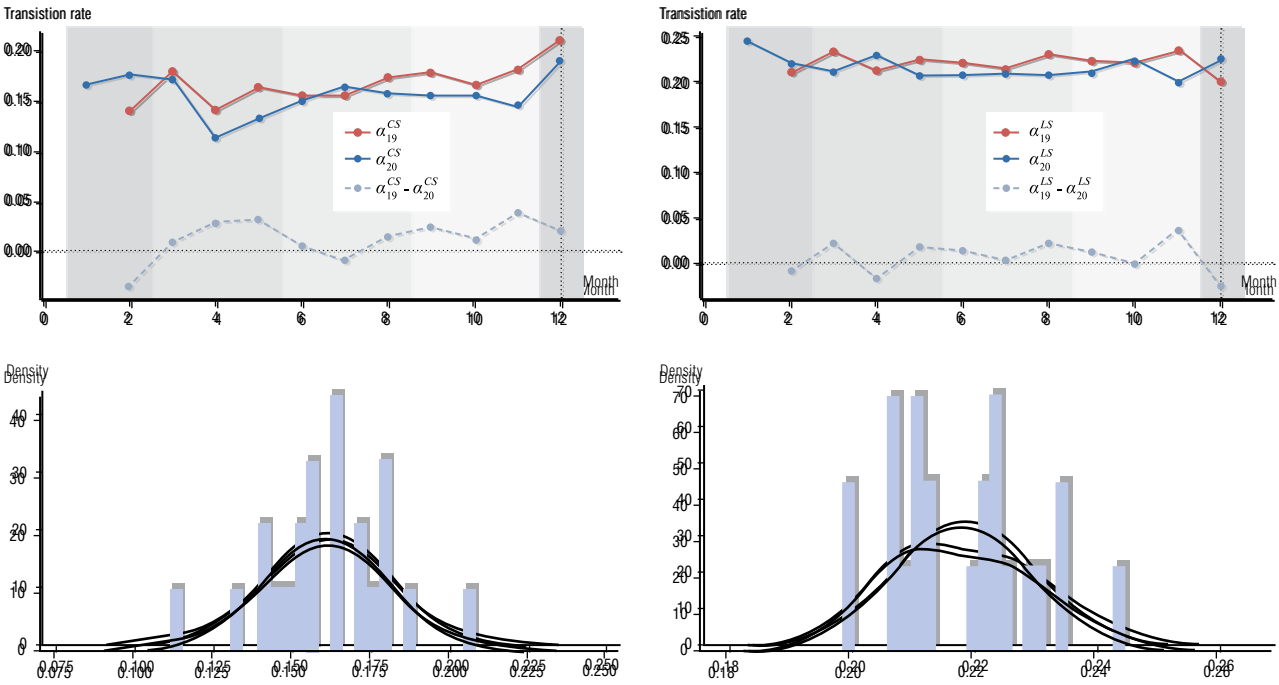

Fig. 4. Flow coefficients to the cluster S (sleeping).

sity estimation (KDE) of this distribution and the normal distribution with the mean and variance calculated from the observed values. It can be seen from these graphs that the distribution of the values of the flow coefficients is close to normal. Also, based on the information presented, it can be assumed that the time series representing the values of the coefficients for both years are stationary.

To test the assumption of stationarity of time series, an Augmented Dickey-Fuller test (ADF) was performed. The values of the corresponding statistics are presented in *Table 1*, column ADF. The results obtained prove that the null hypothesis about the presence of unit roots and, consequently, the non-stationarity of the series is rejected for all coefficients except α^{CL} and α^{LS}. However, if we exclude the observation corresponding to April 2020, which introduces the greatest disturbances, then all the series become stationary (column ADF_{-4}).

In addition, to assess the similarity of the coefficients for the two years studied, two measures were calculated (*Table 1*): cosine similarity (CS) and mean absolute percentage error (MAPE)

$$CS(P,N) = \frac{\sum_{i=1}^{n} P_i N_i}{\sqrt{\sum_{i=1}^{n} P_i^2}\sqrt{\sum_{i=1}^{n} N_i^2}},$$

$$MAPE(P,N) = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{P_i - N_i}{P_i}\right|.$$

In this case, P and N are vectors of flow coefficients for 2019 and 2020, respectively.

Cosine similarity is the cosine of the angle between two vectors in n-dimensional space, where n is the number of values in the sequence. As follows from *Table 1*, the angle between vectors of the flow coefficients is almost zero, i.e., the directions of the vectors coincide. Relatively high *MAPE* values are also explained by the results of the lockdown due to the COVID-19 pandemic. This is proved by comparing the *MAPE* (%) and $MAPE_{-4}$ (%) columns calculated respectively on all data and the data from which April was excluded.

Thus, it can be assumed that the inclusion of information about flows between groups in the number of features when solving the problem of predicting the outflow of customers will improve the accuracy of classification. Using the coefficients α, we can calculate

Table 1.

The results of testing for stationarity and similarity measures of coefficients α for 2019 and 2020

Coefficients		ADF	ADF_{-4}	*CS*	*MAPE* (%)	$MAPE_{-4}$ (%)
Flows to cluster C	α^{LC}	–5.088*	–4.121*	0.982	16.0	9.2
	α^{SC}	–4.594*	–4.855*	0.993	10.4	6.8
Flows to cluster L	α^{CL}	–2.777**	–7.968*	0.990	10.9	8.3
	α^{SL}	–5.367*	–6.612*	0.992	8.4	5.0
Flows to cluster S	α^{CS}	–1.431	–10.757*	0.992	12.4	11.7
	α^{LS}	–0.995	–5.649*	0.997	7.5	7.5

$^{*}p < 0.01$; $^{**}p < 0.1$.

Table 2.

The value of the *TNR*, *TPR*, *AUC* и G_{mean} metrics for logistic regression

Period	$D_1 = [R, F, M]$		$D_2 = [R, F, M, \hat{p}, \hat{v}]$		$\frac{G_{mean}(D_2)}{G_{mean}(D_1)}$	$\frac{AUC(D_2)}{AUC(D_1)}$
	G_{mean}	*AUC*	G_{mean}	*AUC*		
04/20	0.811	0.812	0.856	0.863	1.055	1.063
05/20	0.739	0.745	0.741	0.747	1.004	1.003
06/20	0.781	0.781	0.000	0.500	0.000	0.640
07/20	0.767	0.767	0.846	0.854	1.102	1.113
08/20	0.750	0.750	0.750	0.750	1.000	1.000
09/20	0.733	0.734	0.839	0.850	1.145	1.158
10/20	0.758	0.758	0.841	0.853	1.109	1.125
11/20	0.743	0.743	0.842	0.854	1.134	1.148
12/20	0.742	0.742	0.831	0.839	1.120	1.131

an estimate $\hat{p}$ of the conditional probability $p(y = 1|X)$ of moving a customer from a group with high purchase costs to a cluster with lower costs, i.e., from the loyal cluster to the sleeping or churn clusters and from the sleeping cluster into the churn cluster. In accordance with the conditions formulated above, the assessment of the probability that the client belongs to class 1 is defined as

$$\hat{p}_t = \alpha_t^{LS} + \alpha_t^{LC}, \ if\ cluster = loyal$$
$$\hat{p}_t = \alpha_t^{SS}, \quad if\ cluster = sleeping.$$

Since the estimate of $\hat{p}_{t+1}$ for the forecast period is unknown, we introduce an additional variable $\hat{v}_t$, that considers potential changes in $\hat{p}_{t+1}$ based on the data of the previous year

$$\hat{v}_t = \hat{p}_{t+1-q} - \hat{p}_{t-q},$$

where q is the time lag corresponding to the duration of the pattern of repetitive customer behavior. In this case $q = 12$.

3. Results and discussion

To test the effectiveness of the proposed approach, a numerical experiment was conducted to predict the outflow of customers based on a logistic regression model. The model was trained on two data sets: the first one included the metrics $D_1 = [R, F, M]$; the second one was extended by the variables рен за счет предложенных здесь переменных $D_2 = [R, F, M, \hat{p}, \hat{v}]$ proposed here. The task of the model was to determine the class of the object in the period $t + 1$ from the data in period t. Training, respectively, was carried out on the data set $[(R, F, M, \hat{p}, \hat{v})_{t-1}, y_t]$.

The results of serial testing of the model on validation samples over different periods are presented in *Table 2*. The metrics used are the area under the receiver operating characteristic curve (AUC) and the geometric mean

$$G_{mean} = \sqrt{TPR \cdot TNR},$$

where *TPR* and *TNR* are the proportion of correctly classified objects of true positive rate ($y = 1$) and true negative rate ($y = 0$), respectively. The choice of this metric in addition to AUC is justified by the fact that, other things being equal, the geometric mean has a higher value for balanced predictions for both classes [16].

As follows from *Table 2*, the inclusion of variables $\hat{p}$, $\hat{v}$ provides an increase in the accuracy of the predictive model (in most cases, more than 10%) during periods when there are no external shocks. Thus, the proposed approach can be used to develop individual strategies for retaining users in a relatively stable time.

At the same time, the model turns out to be useless when the influence of external disturbances is catastrophic, and this influence manifests itself with a delay (see results for the period 06/20). In this case, the predictive ability of the model corresponds to random guessing ($AUC = 0.5$), and all predicted objects are classified as belonging to class 0. This is quite understandable, since to predict the period $t = 6$ based on data from period $t = 5$, a model is used that is trained to classify objects at time $t = 5$ on data $t = 4$, when the behavior of customers has changed dramatically due to the lockdown. This means that in April 2020 there was a violation of the customer behavior pattern based on which estimates of the probability of behavior change are built. However, this situation is not critical since the decrease in the predictive ability of the model due to external shocks is quite foreseeable and can be taken into account when using it.

A signal warning about a potential decrease in the accuracy of the model is a significant deviation at the moment t of the current values of the coefficients α^{AB} from the values recorded for this moment in previous years. This deviation can be detected by analyzing the graphs presented in *Figs. 2–4* or by statistical methods. If there is such a deviation for the forecast at $t + 2$, it is advisable to use a model trained on the data $D_1 = [R, F, M]$.

Conclusion

The paper shows that in the analyzed retail network there is a repeating pattern of customers moving between groups with the same behavior lasting one year. Using information about customer flows between these groups allows you to assess the likelihood of a change in their behavior. Compared with the traditional RFM model, the forecasting accuracy is increased by more than 10%.

It also demonstrates the limitations of the proposed approach, which are associated with a violation of the behavior pattern due to external shocks. A method for identifying such a violation is proposed, making it possible to predict the degradation of the predictive ability of the model.

In conclusion, we also list possible ways to improve the efficiency of the proposed method:

- Using a sample spanning more than two years. This will make it possible to more accurately determine the average values of the α^{AB} coefficients, as well as consider their trends.
- Selection of periods of shorter duration (for example, a week). Potentially, this can make it possible to detect patterns of behavior change of higher frequency, which will increase the accuracy of forecasting in the short term.
- Use of more complex models than logistic regression. ■

References

1. Yeh I.C., Yang K.J., Ting T.M. (2009) Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, vol. 36(3), pp. 5866–5871. https://doi.org/10.1016/j.eswa.2008.07.018

2. Kotler P., Armstrong G. (2006) *Principles of Marketing*, 11th ed. NY: Pearson Prentice Hall.

3. Huang S.C., Chang E.C., Wu H.H. (2009) A case study of applying data mining techniques in an outfitter's customer value analysis. *Expert Systems with Applications*, vol. 36(3), pp. 5909–5915. https://doi.org/10.1016/j.eswa.2008.07.027

4. Wei J.-T., Lee S.-Y., Wu H.-H. (2010) A review of the application of RFM model. *African Journal of Business Management*, vol. 4(19), pp. 4199–4206. https://doi.org/10.5897/AJBM.9000026

5. Vafeiadis T., Diamantaras K.I., Sarigiannidis G., Chatzisavvas K.C. (2015) A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

6. Hughes A.M. (1994) *Strategic Database Marketing*. NY: Probus Publishing.

7. Ernawati E., Baharin S.S.K., Kasmin F. (2021) A review of data mining methods in RFM-based customer segmentation. *Journal of Physics: Conference Series*, vol. 1869(1), 012085. https://doi.org/10.1088/1742-6596/1869/1/012085

8. Heldt R., Silveira C.S., Luce F.B. (2021) Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research*, vol. 127, pp. 444–453. https://doi.org/10.1016/j.jbusres.2019.05.001

9. Peker S., Kocyigit A., Eren P.E. (2017) LRFMP model for customer segmentation in the grocery retail industry: A case study. *Marketing Intelligence and Planning*, vol. 35(4), pp. 544–559. https://doi.org/10.1108/MIP-11-2016-0210

10. Chen Y.L., Kuo M.H., Wu S.Y., Tang K. (2009) Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications*, vol. 8(5), pp. 241–251. https://doi.org/10.1016/j.elerap.2009.03.002

11. Hosseini M., Shabani M. (2015) New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics*, vol. 3(3), pp. 110–121. https://doi.org/10.1057/jma.2015.10

12. Akhondzadeh-Noughabi E., Albadvi A. (2015) Mining the dominant patterns of customer shifts between segments by using top-k and distinguishing sequential rules. *Management Decision*, vol. 53(9), pp. 1976–2003. https://doi.org/10.1108/MD-09-2014-0551

13. Lemmens A., Croux C., Stremersch S. (2012) Dynamics in the international market segmentation of new product growth. *International Journal of Research in Marketing*, vol. 29(1), pp. 81–92. https://doi.org/10.1016/j.ijresmar.2011.06.003

14. Rousseeuw P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

15. Breiman L. (2001) Random Forests. *Machine Learning*, vol. 45(1), pp. 5–32.

16. Zelenkov Y., Volodarskiy N. (2021) Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers. *Expert Systems with Applications*, vol. 185, 115559. https://doi.org/10.1016/j.eswa.2021.115559

17. Fader P.S., Hardie B.G., Lee K.L. (2005) RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, vol. 42(4), pp. 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

18. Lumsden S.A., Beldona S., Morrison A.M. (2008) Customer value in an all-inclusive travel vacation club: An application of the RFM framework. J*ournal of Hospitality & Leisure Marketing*, vol. 16(3), pp. 270–285. https://doi.org/10.1080/10507050801946858

19. Bjørnstad O.N., Shea K., Krzywinski M., Altman N. (2020) The SEIRS model for infectious disease dynamics. *Nature Methods*, vol. 17, pp. 557–558. https://doi.org/10.1038/s41592-020-0856-2

About the authors

Yury A. Zelenkov

Dr. Sci. (Tech.);

Professor, Department of Business Informatics, Graduate School of Business, HSE University, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: yzelenkov@hse.ru

ORCID: 0000-0002-2248-1023

Angelina S. Suchkova

Student, BSc Program «International business and management», Saint Petersburg School of Economics and Management, HSE University – Saint Petersburg, 16, Soyuza Pechatnikov Street, Saint Petersburg 190121, Russia;

E-mail: assuchkova_1@edu.hse.ru