Forecasting purchase categories by transactional data: a comparative study of classification methods

Egor Shikov and Klavdiya Bochenina

[1] ITMO University, 49 Kronverkskiy prospect, 197101, Russian Federation
shikovegor86@gmail.com, k.bochenina@gmail.com

Abstract. Forecasting purchase behavior of bank clients allows for development of new recommendation and personalization strategies and results in better Quality-of-Service and customer experience. In this study, we consider the problem of predicting purchase categories of a client for the next time period by the historical transactional data. We study the predictability of expenses for different Merchant Category Codes (MCCs) and compare the efficiency of different classes of machine learning models including boosting algorithms, long-short term memory networks and convolutional networks. The experimental study is performed on a massive dataset with debit card transactions for 5 years and about 1.2 M clients provided by our bank-partner. The results show that: (i) there is a set of MCC categories which are highly predictable (an exact number of categories varies with thresholds for minimal precision and recall), (ii) for most of the considered cases, convolutional neural networks perform better, and thus, may be recommended as basic choice for tackling similar problems.

Keywords: purchase forecasting, financial behavior, neural networks, machine learning, transactional data, machine learning.

1 Introduction

Enterprise information systems incorporate different sources of information about actions of employees and clients which further can be used to create 360-degree customer view or to develop a variety of tools for predictive analytics of financial behavior. One of the problems often encountered for bank clients is expanding their debit cards payment experience, that is, increasing the number of different categories of expenses or the intensity of debit card usage. Being informed about expected future payments of a group of clients, a decision maker may provide adaptive and personalized suggestions for increasing the loyalty and improving customer debit card experience. For example, if one expects that customer A will spend amount X in category B in the next month, one may suggest to customer A to spend $1.1 \cdot X$ in category B and get a discount from a partner in this category, or to perform payment in similar category C and to get increased cashback for it.

The transactional data for each customer are formed as a sequence of transactions with timestamp, amount and category of a transaction. Depending on a granularity of hierarchy of payment categories, one may tackle from dozens to hundreds of categories.

That is, to predict payment profile for the next time period, we need to apply a binary classifier (or classifier which predicts a probability of having at least one purchase) for each of the categories. Also, one needs to mention that the frequencies of categories are highly imbalanced, and, according to the nature of the problem, different categories are of different basic predictability (e.g. almost all clients have at least one purchase in category ‘food’ each month, and expenses in ‘petrol stations’ and ‘housing services’ category seem to have less variance than in ‘medical goods’ or ‘building materials’).

The goals of this study are, given a massive dataset on debit card transactions: (i) to perform a comparative study of the efficiency of several modern machine learning approaches for prediction of the categories of expenses for the next time period, (ii) to study the predictability of different purchase categories and to provide the recommendations on categories which are more appropriate for planning the campaign for increasing customer involvement in debit card purchases (in terms of high precision and sufficient recall). Also, we test the consistency of predictions for different months (as patterns of purchases basically vary from month to month) and examine the quality of forecasts for the case when the last month of transactional history is unavailable for the model.

The rest of the paper is organized as follows. Section 2 presents related work on methods which were applied to solve this problem and similar problems earlier. Section 3 gives formal description of the problem, description of data processing workflow and details of implementation of different methods for our problem. Section 4 describes the dataset, the methodic of experimental study and the experimental results. Finally, Section 5 presents conclusion and discussion.

2 Related work

Basically, the problem of forecasting purchase categories may be considered as multivariate time series prediction problem. Depending on the exact problem statement, these time series may be of numerical or categorical variables where each dimension is a single Merchant Category Code (MCC). In such a case, values for a single time unit comprise a vector of amount of purchases in different categories or binary flags marking the existence of at least one purchase for this MCC during given unit. Most of the categories are not presented in the payment sequences of users for a given time step, so basically these vectors are sparse. To make effective predictions, one needs to account not only the interactions between time steps, but also interactions between features. So, the difference between the models (classes of hypotheses) is determined with how these interactions are tackled.

Factorization machines (FMs) use second-order features interactions and are able to infer latent features from a highly sparse dataset using matrix factorization techniques. This method is often used for business cases such as recommender systems. For example, Lee et. al. [1] use FMs for next event prediction task in business processes (namely, loan activities of bank client). There are also variants of FM which incorporate time dependencies between the events such as Factorizing Personalized Markov Chains (FPMC) [2] and Feature-Space Separated Factorization Model (FSS-FM) [3].

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_19

Dependencies between features in FM are linear. To add non-linearity in higher-order feature interactions, Neural Factorization Machine (NFM) was proposed in [4]. NFM may be considered as a generalization of classical FM and shows comparable performance with deep learning models while having simpler, shallower structure. Authors of [5] argue that matrix factorization methods systematically oversmooth distribution of user-item pairs resulting in too high probabilities of unseen items for a given user. To balance between exploration and exploitation, they introduce mixture model with two components, estimated at population and individual levels, correspondingly. The results of the mixture model are compared at seven online and offline user-item datasets and show the advantage of mixture model in terms of log-likelihood of test data and Recall@k.

To add global sequential features to the model (not only between consecutive events but also for non-adjacent cases), different architectures of recurrent neural networks (RNNs) are used. Dynamic Recurrent bAsket Model (DREAM) [6] calculates hidden state of RNN as a function of previous hidden state and latent vector representation of user's basket at time t_i. To get these latent representations, authors use operations of max pooling and average pooling over set of items in the basket (where each item, in turn, is represented as vector). Thus, the model combines representation of current interests of a user with a memory about her interests from previous baskets. Experiments have shown that DREAM outperforms simple baseline models (first-order Markov chains, non-negative matrix factorization) as well as more sophisticated models as FPMC and hierarchical representation model (HRM) [7]. As state-of-the-art model to compare with, different boosting models (as Gradient Boosting Machines in [8]) are also used.

Sequences of transactions may contain repeated patterns similar to graphical primitives and shapes in images. Thus, the next idea for predicting financial behavior is to use convolutional neural networks (CNNs) to find and use these patterns from large arrays of transactional data. This approach was used, for example, for fraud detection in [9] with excellent results on precision and recall of classifier. In [6] an author reports the results on applying CNN for a data about monthly usage of bank products to predict future usage of products.

In our study, the goal is to test existing approaches to predicting user consumption behavior for a problem of forecasting purchase categories in a next time period. For our best knowledge, there were no attempts to perform systematical comparison of predictive ability of these methods for a considered problem. We choose for the evaluation methods from different classes described above, namely recurrent neural networks, boosting and convolutional neural networks, and test them on a real-world dataset of debit card transactions provided by our bank-partner.

3 Problem description

The problem of forecasting purchase categories may be described as follows. There is a set of clients $U = \{u_i\}, i = 1, \dots, N$ where N is a total number of clients. Each client is characterized with a tuple $< F_i, S_i >$ where F_i is a set of static features of client i

DOI: 10.1007/978-3-030-22744-9_19

(such as gender, age), and S_i is a sequence of debit card transactions of client i. This sequence is represented as $S_i = \{< a_{ij}, c_{ij}, y_{ij}, m_{ij} >\}, l = 1, \dots, N_i^S$, where N_i^S – a total number of transactions of i-th client, a_{ij} – an amount of j-th transaction of i-th client, $c_{ij} \in C$ – a category of l-th transaction of i-th client, C – a set of categories (we denote a cardinality of this set, that is, a number of categories, as M).

For each client, his or her transactions may be aggregated by periods of a given length d. If total period of transactions of client u_i is equal to T_i, then the aggregated purchase matrix (APM) is defined as matrix with $K_i = \lceil T_i/d \rceil$ vectors as rows:

$$P_i = \{< n_{i1k}, v_{i1k}, \dots, n_{ijk}, \dots, n_{iMk}, v_{iMk} >, z_{ik}\}_{k=1}^{K}, \qquad (1)$$

where i is an index of client, k is an index of time period, $j = 1, \dots, M$ – an index of category, n_{ijk} – a number of transactions in category l in k-th period of client i, v_{ijk} – a total amount of transactions in category l in k-th period of client i, z_{ik} – a label which marks a 'global' index of k-th period of i-th client. For example, if we aggregate by months, earliest transaction among all of the clients was marked as Jan, 1990, the latest transaction was marked as Jun, 1991, and u_i has transactions from October, 1990 to May 1991, then local index (k) of Jan, 1991 for client u_i will be equal to 4, and global index – to 13. Also, we denote time borders of transactions for set of clients as $B = [< y_b, m_b >, < y_e, m_e >]$ (in our example $y_b = 1990, m_b = Jan, y_e = 1991, m_e = Jun$.

Then, the problem of forecasting purchases in a given category is formulated as follows. Given a set of tuples $U = \{u_i =< F_i, P_i > \}, i = 1, \dots, N$ with static attributes and aggregated purchase matrices for a set of N clients (estimated by period B) predict for a given period $z^* > \max z_{ik}$ for each client u_i and category c_j if there is at least one transaction of this category in the period. That is, our goal is to get a matrix of predictions with clients as rows and categories as columns where non-zero entry with indices i and j means that a client i will spend in category j. Variants of the problem are prediction of n (number of transactions) and v (amount of transactions). In this study, we use fixed $K = K_i$ for all customers in the data set under the assumption that length of B may be larger than K. This means that a dataset may be of any length, but we use for prediction only K last months of client's transactional history (for the training sliding window technique thus should be used).

4 Methods

This section contains short descriptions of methods and some implementation details for a problem from Section 3. The models built for classification and regression task shared the same architecture except the fact that the counts of transactions were used for classification and expenses were used for regression.

4.1 Baseline method (averaging)

As a baseline method to estimate the quality of prediction, we use averaging per each client and each category over a given history of transactions. That is, a probability of purchase in j-th category by i-th client is given by:

$$p_{ij} = \frac{1}{K_i} \cdot \sum_k \mathbb{I}\left(n_{ijk} > 0\right), k = 1, \dots, K_i. \tag{2}$$

Expression (2) provides a frequency of expenses in a given category in terms of periods of aggregation (if time unit is month, then p_{ij} is a fraction of months with expenses in category j in the history of user i).

4.2 Recurrent neural networks

We trained a simple LSTM with sparse vectors of monthly numbers of transactions as inputs and the hidden state of size 128. The network was trained with BPTT (Backpropagation Through Time) with cross-entropy loss for classification and MSE loss for regression.

4.3 Convolutional neural networks

The input layer was constructed as a concatenation of vectors of expenses. We used a simple CNN with 2 Conv2D-layers, and a pooling layer. The final layer as well as losses were the same as in LSTM. MSLE (Mean Squared Logarithmic Error) loss was also tested to reduce the influence of outliers for regression problem.

4.4 Boosting

As an inputs for the algorithm, the following features were estimated:

- minimum, maximum, average values and standard deviation for 6 last months;
- minimum, maximum, average values and standard deviation for 3 last months;
- last month expenses;
- average, minimum, maximum expenses summed over all categories;
- target month, customer age, customer gender.

All these features were concatenated to form a 759-long feature vector.

XGboost with default settings was used, a separate model for every category was trained.

5 Data

We use the dataset provided by our bank-partner (one of the largest regional banks in Russia) with $N = 180\ 000$ and $z_{\max} = 68$ (that is, dataset covers 5 years and 8 months). The customers were chosen to support sufficient level of transactional activity (restrictions are: $N_i^S \geq 200$ and at least 2 distinct categories of transactions in last 48

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22744-9_19

months, $N_i^S \geq 6$ in last 6 months). M (a total number of categories) is equal to 83 (an initial number of categories was equal to 86, but we did not use category ‘Financial services’ (it is the most frequent among the others as it consists of cash withdrawal), and categories ‘Associations and organizations’, ‘Funeral services’ according to their extremely low frequency).

Categories significantly differ in relative expenses and frequencies. Figure 1 illustrates this difference for 10 most frequent purchase categories (abbreviations for the categories are SM for ‘Supermarkets’, MP for ‘Mobile Phones’, BR for ‘Bars and Restaurants’, GS for ‘Gas Stations’, CS for ‘Clothes, Shoes and Accessories’, MG for ‘’Medical Goods’, PS for ‘Personal Services’, C for ‘Cosmetics’, HS for ‘Household Stores’, and SS for ‘Special Stores’). Figure 1a shows average monthly expenses normalized in (0,1] interval, Figure 1b shows the average frequency of purchases for different categories measured at per-month basis, and Figure 1c shows percentage of transactions in different categories. As we will see later, it determines different predictive power of the algorithms for different categories. Clients also differ in the diversity of categories of expenses and frequency of payments (Figure 1d). This suggests the existence of subpopulations with different types of purchase behavior.

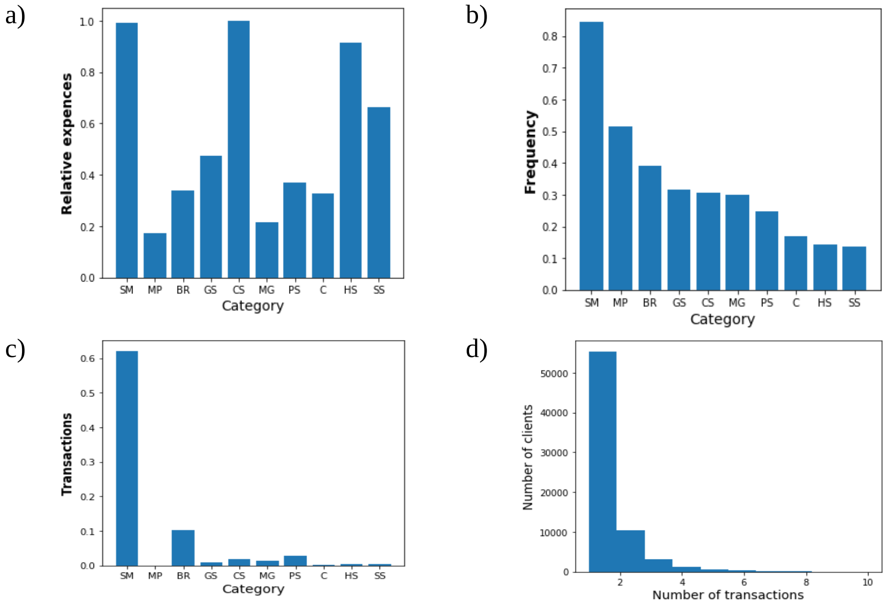
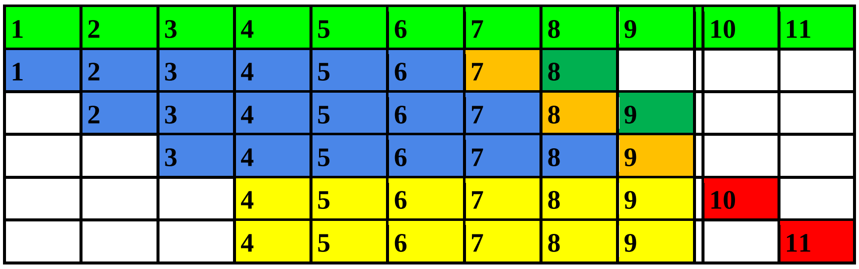


Fig. 1. Properties of expenses in a selected set of 10 most frequent categories: a) normalized average monthly expenses, b) frequency of payments (the average percent of months with purchase in a given category), c) Distribution of parts of transactions in ten most popular categories, d) Distribution of an average number of transactions of a client per month

As our data are time-dependent, for the test set we use all partial sequences S_i of the clients after certain timestamp $z_{\max}$. Figure 2 illustrates the principles of formation of training and test sets for the data of a single client (columns of the table represent data for different k (months), rows of the table represent different ways of using the sequence). In this example, we know $K_i = 9$ months of transactional history of client i, and we want to train the model which predict categories using last 6 months of transactional data. Then, to form the training set, we may use sliding window for each six consecutive months of client's transactions under the condition that month to extract the label from lays within K_i months. One last thing to mention is that there can be introduced a lag between the end of the period used for prediction, and the period for making the forecast. That is, we may consider prediction for the next month or prediction for the month after the next month. It may be useful if there is delay in data collection after the end of the last month.



1	2	3	4	5	6	7	8	9	10	11
1	2	3	4	5	6	7	8			
	2	3	4	5	6	7	8	9		
		3	4	5	6	7	8	9		
			4	5	6	7	8	9	10	
			4	5	6	7	8	9		11

Fig. 2 – Structure of train and test data for a single customer ($k = 1, \dots, 9$ – train months, $k = 10, 11$ – test months). Blue – data for creating $[6 \times (M \cdot 2 + 1)]$ aggregated purchase matrix for training set, orange – data for extracting labels for training set (to predict month $z_{\max} + 1$), dark green – data for extracting labels for training set (to predict month $z_{\max} + 2$), yellow – data to create APM for evaluation of the model, red – data for extracting test labels.

6 Results

The experiments were performed using a desktop PC with the following hardware configuration: Intel i7-7800X, 16 GB RAM, GeForce GTX 1060 6GB. The dataset was divided at train (60 months) and test (11 months). We solve the problem of predicting purchase categories by last 6 months of transactional activity, so for a single client there can be several training examples for different time windows in the training set. We tried lags equal to zero and to one month (so we used both schemes from Figure 2). As we have more than one month in our test period, all the metrics are averaged by a number of test months (to train, each time we use last year of transactional history before test month). Hyperparameters of classifiers were trained as it was explained in Section 5, and the resulting values are shown in Table 2. Further we refer different models as: Average, LSTM, CNN and Boosting.

Table 1. Parameters of different classifiers for categories prediction problem

Method	Parameters
LSTM	LSTM: hidden layer size = 128 Batch size = 64 Dropout rate = 0.2 Optimizer: Adam (learning_rate = 0.01)

CNN	Conv2D_1: nunits = 128, kernel = (2,16) ReLU MaxPooling: (2,2) Conv2D_2: nunits = 64, kernel = (2,2) ReLU Batch size = 64 Dropout rate = 0.2 Optimzer: Adam (learning_rate = 0.01)

Boosting	Nfeatures = 759
	Ntrees = 100
	learning_rate=0.1
	max_depth=3
	subsampling: ON

The output of all considered classifiers for a given client is an M-dimensional vector with probabilities of categories. To transform these probabilities to binary values, one needs to set a threshold to balance between precision and recall of classifier. With our business case as a frame of reference, threshold was set for each category independently to support certain level of precision (80% or 90%). This approach was used because the results of prediction are aimed for launching campaigns for enhancement of customer debit card activities. These campaigns are planned for a restricted audience but require precise identification of target audience. So, thresholds are tuned using training set and are applied to make decisions for test set.

Figure 3 shows Precision-Recall curves for different classifiers for several categories. This curve shows achievable variants of tradeoffs between confidence in predicted values (measured by precision) and the amount of positive cases which will be captured by the model (measured by recall). The better curve is the closest to the right upper corner of the plot. We can see from Figure 3 that for some categories (basically, the most frequent ones) curve achieves the plateau indicating both high precision and high recall (90% precision / 60% recall for Gas stations, 90% / 60% for Bars & Restaurants for the best classifier). For another categories, there is a region with sufficiently high value of precision up to some value of recall (80%/60% for Musical Instruments[1], 80%/40% for Medical Goods, 80%/45% for Municipal Services). For this set of categories, planning marketing campaigns based on predictions is still possible, as there is a highly predictable segment of the audience. If such algorithm operates over several hundreds of thousands of customers, then even 10% of recall may provide sufficient number of clients for setting the campaign. For the third set of categories, we cannot make a prediction with reasonable precision. Usually, these categories are less frequent.

Figure 3 also shows the difference in predictive ability of classifiers. Simple baseline in some cases outperformed more sophisticated algorithms for small values of recall. In such cases, there exists a small segment of customers with repeatable monthly behavior. Complex models like CNN seek for more sophisticated patterns in the data and then provide lower but more stable precision values.

[1] This category mostly contains purchases in monthly paid media services like Apple Music.

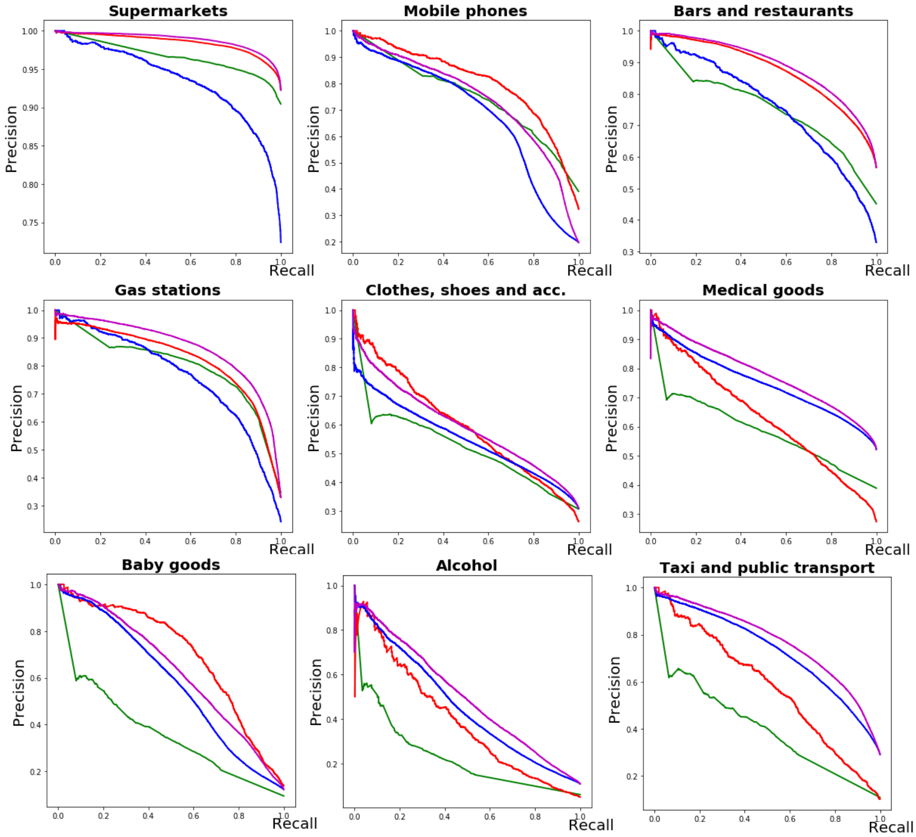


Fig. 3. Precision-Recall curves for selected categories and different classifiers (green – Average, red – CNN, blue – LSTM, magenta - Boosting)

More systematic analysis of comparative precision is presented below. After setting the thresholds, we obtain values of precision and recall for a fixed precision threshold. Table 2 shows the results for different classifiers for 80% and 90% thresholds for 10 categories with highest average values of recall (over all classifiers) for 80% threshold. For 80% threshold, CNN outperforms LSTM in all categories except three (‘Musical instruments’, ‘Taxi and Public Transport’, ‘Medical Goods’). For 90% threshold, CNN outperforms LSTM in all categories.

Another comparison of predictive ability of classifiers (Table 3) shows different number of categories which are classified with precision larger than precision threshold and have recall larger than recall threshold. CNN support 1.5 times higher number of categories than LSTM for both considered values of thresholds.

Table 2. Comparison of recall values of classifiers for different categories with fixed precision

Category	**80% precision threshold**			
	Average	**CNN**	**LSTM**	**Boosting**
Supermarkets	**1.00**	**1.00**	**1.00**	**1.00**
Mobile phones	0.45	**0.65**	0.44	0.50
Gas stations	0.63	**0.86**	0.70	0.79
Baby goods	0.00	**0.51**	0.30	0.34
Bars and restaurants	0.41	0.76	0.67	**0.81**
Musical instruments	0.00	0.45	0.49	**0.66**
Municipal services	0.00	**0.44**	0.23	0.38
Hosting, TV	0.00	**0.26**	0.04	0.10
Marketing	0.00	**0.24**	0.02	0.05
Taxi and public transport	0.00	0.24	0.45	**0.52**
Medical goods	0.00	0.23	0.34	**0.46**
Clothes, shoes and accessories	0.00	**0.19**	0.01	0.10
Barbershop	0.00	**0.14**	0.01	0.01
Petshops	0.00	0.14	0.09	**0.15**
Cosmetics	0.00	**0.13**	0.01	0.08
Alcohol	0.00	0.11	0.12	**0.15**
Category	**90% precision threshold**			
	Average	**CNN**	**LSTM**	**Boosting**
Supermarkets	**1.00**	**1.00**	**1.00**	**1.00**
Mobile phones	0.00	**0.33**	0.15	0.22
Gas stations	0.00	0.52	0.38	**0.54**
Baby goods	0.00	**0.29**	0.18	0.21
Musical instruments	0.00	0.28	0.25	**0.35**
Bars and restaurants	0.00	0.28	0.52	**0.57**
Municipal services	0.00	**0.26**	0.05	0.04
Marketing	0.00	**0.15**	0.01	0.00
Petshops	0.00	**0.11**	0.01	0.05

Table 3. Number of categories which are classified with precision larger or equal to threshold and fixed recall (0.1)

	LSTM	CNN	Average	Boosting
80% precision	10	16	5	13
90% precision	6	9	1	6

For CNN classifier, we additionally investigated consistency of predictions for consequent test months and quality of predictions for instances of the problem with zero and unit lags. Figure 4 shows P-R curves for categories Baby goods, Gas stations, Restaurants and Supermarkets for several different test months. One can see that although there are some fluctuations of quality for distinct months, the shape of P-R curve remains the same. Figure 5 shows P-R curves for six different categories for zero and unit lags. As expected, the absence of last month influences quality of predictions but the effect cannot be considered as drastic.

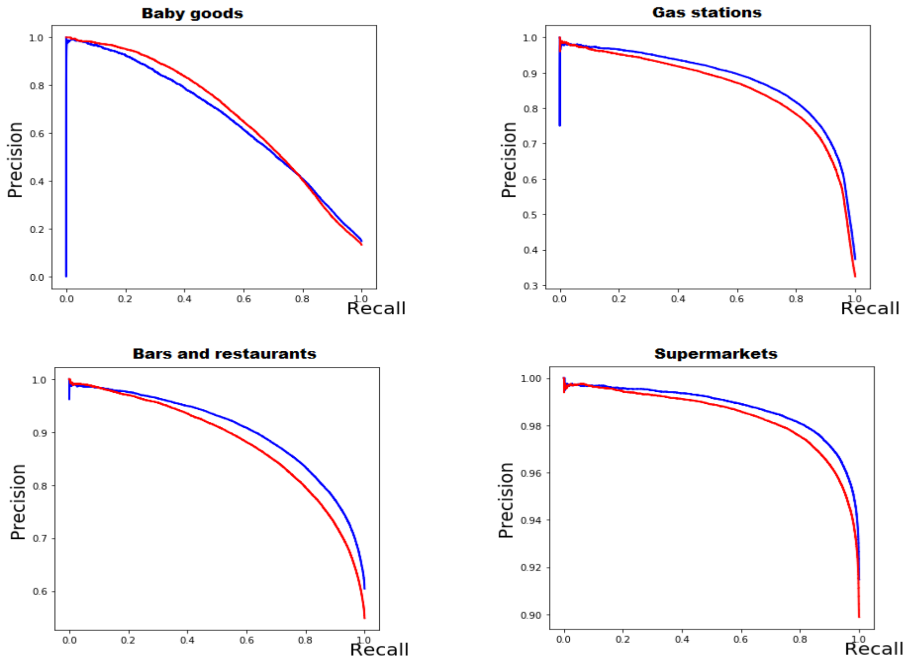


Fig. 4. Precision-Recall curves for different test months and different categories (blue – October 2018, red – November 2017)

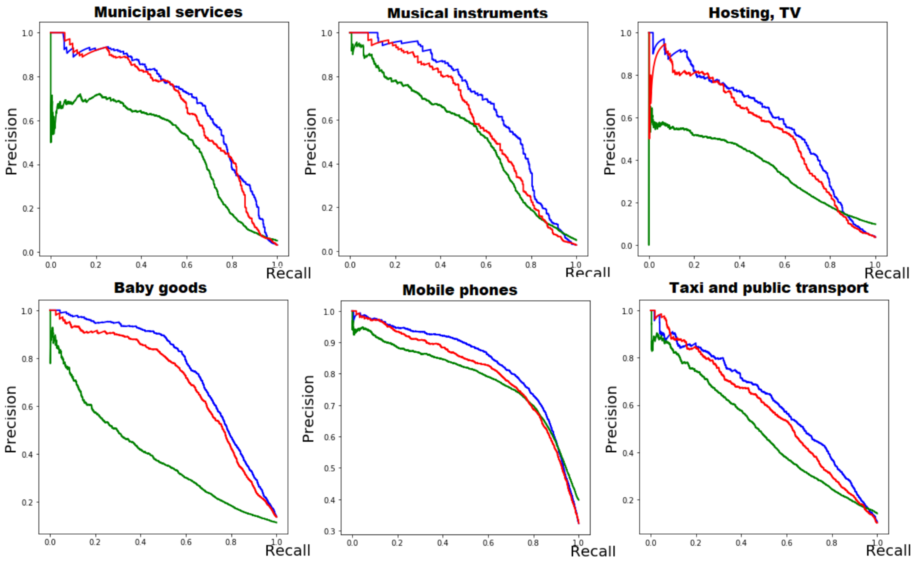


Fig. 5. Precision-Recall curves for different categories with zero and unit lags (prediction for the next month – blue, prediction for the month after the next – red; green – Average prediction)

The algorithms may be used to solve the problem in a regression statement (to predict total sum of purchases in a given category for the next month). Table 4 provides results on MAPE values for different methods. One can see that CNN also provide the best percentages for all of the considered categories. The nearest competitor is twice as worse on average. Boosting algorithm does not provide any advantages compared to simply taking median value of purchases in that category. Average value gives worser percentage than median due to outliers in the sums of transactions. Finally, LSTM shows highest MAPE (almost 4 times higher than CNN). The values of MAPE for CNN vary from 77% for the pet shops to 235% for household stores with average value of MAPE equal to 116%. This indicates that the absolute values of MAPE are still too large to be used for planning marketing campaigns.

Table 4. Mean absolute percentage error for different methods (‘Abs’ field) and ratio to the best achieved value (‘Rel’ field).

	Average		**Median**		**CNN**		**Boosting**		**RNN**	
	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
Supermarkets	160	1.62	143	1.44	**99**	1.00	398	4.02	333	3.36
Mobile phones	108	1.35	99	1.24	**80**	1.00	107	1.34	149	1.86
Bars and restaurants	186	1.92	157	1.62	**97**	1.00	189	1.95	280	2.89
Gas stations	135	1.62	120	1.48	**81**	1.00	131	1.62	165	2.04
Clothes, shoes and accessories	333	1.35	287	2.47	**116**	1.00	251	2.16	213	1.84
Medical goods	223	1.92	193	1.99	**97**	1.00	209	2.15	261	2.69
Personal service	300	1.62	244	2.26	**108**	1.00	219	2.03	375	3.47
Cosmetics	214	1.35	192	2.04	**94**	1.00	199	2.12	213	2.27
Household stores	701	1.92	559	2.38	**235**	1.00	394	1.68	472	2.01
Special stores	594	1.62	471	2.79	**169**	1.00	369	2.18	1085	6.42
Baby goods	255	1.35	228	2.15	**106**	1.00	186	1.75	283	2.67
Building Materials & Supplies	708	1.92	549	3.64	**151**	1.00	257	1.70	345	2.28
Hosting, TV	109	1.62	102	1.01	**101**	1.00	260	2.57	80	0.79

Sport goods	215	1.35	205	2.11	**97**	1.00	190	1.96	3103	31.99
Taxi and public transport	276	1.92	217	1.75	**124**	1.00	235	1.90	501	4.04
Alcohol	157	1.62	146	1.59	**92**	1.00	163	1.77	167	1.82
Pet shops	141	1.35	131	1.70	**77**	1.00	124	1.61	143	1.86
Municipal services	286	1.92	254	1.21	**210**	1.00	384	1.83	307	1.46
Bookstore	174	1.62	165	1.83	**90**	1.00	159	1.77	181	2.01
Medical centers	231	1.35	215	1.90	**113**	1.00	190	1.68	98	0.87
Mean	275	2.23	234	1.93	**116**	1.00	230	1.98	437	3.93

7 Conclusion

Information about consuming goods and services by a set of customers may further be used to develop different kind of personalization strategies. In this study, we consider extraction of the meaningful information to plan personalized marketing campaigns based on forecasting purchase categories for the next time period from large arrays of transactional data. For the case study, we use massive dataset provided by our industrial partner, one of the largest regional Russian banks.

To compare different machine learning algorithms, we state the problem of forecasting purchase categories as a set of binary classification problems. Data analysis shows high level of heterogeneity both in payment behavior of different clients and in different categories. In general, we observe that frequent categories are of significantly higher predictability.

We compare the results of classification and regression on a set of 83 MCC categories for recurrent neural networks, convolutional neural networks and boosting algorithm together with simple baselines as mean and median. The results of classification show that CNN outperform other competitors in terms of higher recall for a fixed precision in a majority of categories. Also, it allows to forecast larger number of categories with a minimum threshold on precision. Our study shows that there exists a set of categories which may be predicted with high accuracy and thus can be used for planning marketing campaigns. The results show their consistency on different months and for the case when data about last month before the test period is not available. As for regression problem, CNN outperforms the nearest competitor in two times. However, the resulting MAPE values are still high and may be used only as a benchmark. Further step here may be distinguishing the customer segments with more predictable expenses, stating the problem as multi-class classification and testing different classes of models.

Acknowledgements

This research is financially supported by The Russian Science Foundation, Agreement №17-71-30029 with co-financing of Bank Saint Petersburg.

References

1. Lee W.L.J. et al. Predicting process behavior meets factorization machines // Expert Syst. Appl. 2018. Vol. 112. P. 87–98.
2. Rendle S., Freudenthaler C., Schmidt-Thieme L. Factorizing personalized Markov chains for next-basket recommendation // Proceedings of the 19th international conference on World wide web - WWW '10. New York, New York, USA: ACM Press, 2010. P. 811.
3. Cai L. et al. Integrating spatial and temporal contexts into a factorization model for POI recommendation // Int. J. Geogr. Inf. Sci. 2018. Vol. 32, № 3. P. 524–546.
4. He X., Chua T.-S. Neural Factorization Machines for Sparse Predictive Analytics // Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17. New York, New York, USA: ACM Press, 2017. P. 355–364.
5. Kotzias D., Lichman M., Smyth P. Predicting Consumption Patterns with Repeated and Novel Events // IEEE Trans. Knowl. Data Eng. 2019. Vol. 31, № 2. P. 371–384.
6. Yu F. et al. A Dynamic Recurrent Model for Next Basket Recommendation // Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16. New York, New York, USA: ACM Press, 2016. P. 729–732.
7. Wang P. et al. Learning Hierarchical Representation Model for NextBasket Recommendation // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15. New York, New York, USA: ACM Press, 2015. P. 403–412.
8. Sheil H., Rana O., Reilly R. Predicting purchasing intent: Automatic Feature Learning using Recurrent Neural Networks // SIGIR 2018 eCom. 2018. P. 9.
9. Zhang Z. et al. A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection // Secur. Commun. Networks. 2018. Vol. 2018. P. 1–9.