A Hybrid Machine Learning Framework for E-commerce Fraud Detection

Yury Y. Festa[1], Ivan A. Vorobyev[2]

[1]Sberbank of Russia, Cybersecurity Department, manager; The National Research University Higher School of Economics, junior research fellow;

[2]Sberbank of Russia, Cybersecurity Department, head of direction; The National Research University Higher School of Economics, junior research fellow

E-mail[1]: festa.y.yura@gmail.com

Abstract

We currently see a large increase in e-commerce sector; it is becoming a central trend in the banking industry. Fraudsters keep up with modern technologies, and use weak points in human psychology and security systems to steal money from regular users. To ensure the required level of security, banks began to apply artificial intelligence in their anti-fraud systems. Fraud detection can be formulated as a classification problem with a case-based reasoning or knowledge extraction task with unbalanced classes. In this paper we present a framework of models based on various approaches of artificial intelligence, such as neural networks, decision trees, copula models and others to recognize payment behavior of fraudster. The considered framework is evaluated with different metrics and implemented in an actual anti-fraud system, which leads to an improvement of the system performance. Finally, the interrelation between the anti-fraud system indicators and banks operational risks is discussed in this paper.

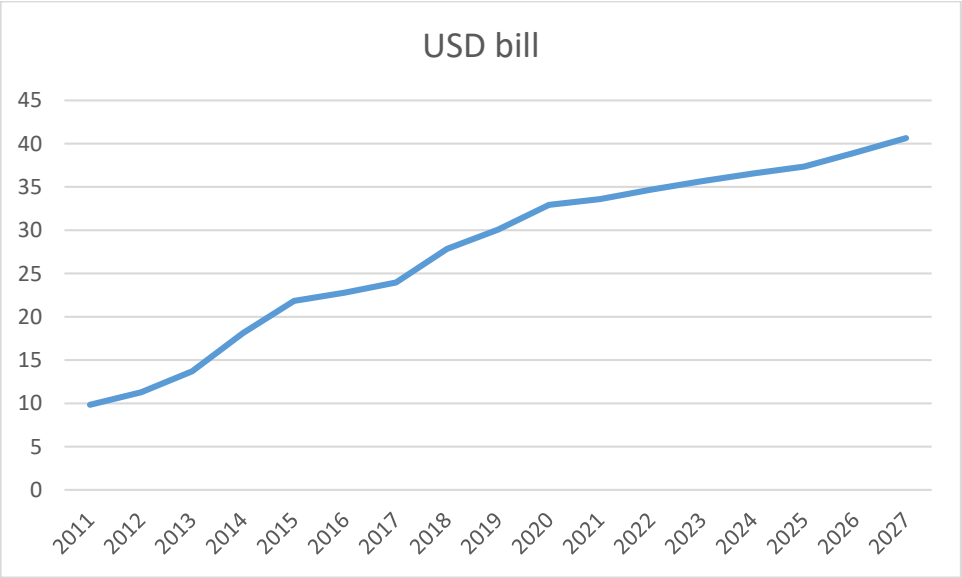
Key words

Machine learning, cyber security, fraud, copula, fraud analysis, classification

1. Introduction

The growth of the transactional business in the banking sector has led to the fact that clients in an online format using smartphones conduct the bulk of transactions (transfers, payments). Banks developing their own mobile applications in which a client can get banking services as quickly and conveniently as possible. However, this convenience is associated with an increase in the risk for cybersecurity, in particular, it becomes easy for fraudsters to use social engineering methods and trick the client into transferring funds to the fraudsters' accounts. The client is always with a smartphone, and during a telephone conversation with fraudsters in a critical situation, the client becomes vulnerable. Separating this social engineering fraud from legitimate transfers and blocking this transaction is a very difficult task: the client makes the operation himself, and the classic patterns for detecting fraud, such as a change of geolocation or a new IMEI number at the client, stop working. To reduce this risk, banks are developing fraud monitoring systems, with elements of artificial intelligence.

The problem of increase in the volume of fraudulent transactions is global. The Nilson Report (2020) predicts $ 40 billion in global fraud by 2027.



Picture.1. Nilson Global Fraud Forecast

In March 2021, the Basel Committee revised its Principles for Sound Operational Risk Management to make the following changes:

- Align principles with the recently revised Basel III operational risk framework;
- Update guidance, where necessary, in the areas of risk management related to information and communication technology;
- Increase the overall clarity of operational risk management principles.

Thus, information security risk management becomes as necessary a component of risk management as other operational risks.

Introducing the elements of artificial intelligence to bank process follows validation and continuous assessment of the quality of the model. This quality is directly related to the bank's financial losses for the formation of a fund for payments to affected customers, the formation of a minimum regulatory capital to cover losses. Low quality model may cause reputational losses, which lead to an outflow of customers and, as a result, an increase in operating costs of marketing activities, growth of the salary fund for attracting additional specialists in the field of sales of banking products.

Evaluation for fraud risk in money transfer use various indicators, obtained from the client's banking profile and mobile application. When a transfer is made between clients of the same bank, additional information for making a decision can also be obtained by evaluating the payee. Most often, when a fraudulent transfer occurs, it is difficult for attackers to determine the region of residence of the victim, his counterparties, detailed account information etc. For this case a random dropper bank account is used as the recipient. In the terminology of bank transactional anti-fraud, a drop is a person who, for a small fee, has a bankcard issued for himself and gives this card to the fraudster. These droppers cannot imitate the behavior of regular clients, so the factors of an atypical payee become a key in identifying such persons.

For a data sample for the study, we used the result of the work of the fraud monitoring system of one of the biggest banks in Russia. The system evaluates the transfer between two clients of this bank, and if it recognizes risk patterns, it blocks transactions. It is proposed to improve the efficiency of fraud detection, and reduce numbers of false positives that negatively affect the customer experience when using the bank's mobile application. In addition, some clients who became fraudster victims may stop trusting a financial institution and stop using remote banking by withdrawing funds from their bank accounts.

The main presupposition for our framework is that fraudsters, having a large number of drop cards, are limited in creating a full client profile for these cards. Legitimate customers who use bankcards for their intended purpose make purchases in various stores, pay bills in restaurants, buy baby products, pet supplies, refuel cars etc. Fraudsters cannot fully imitate such client profile for their dropper cards. It is proposed to consider how the enrichment of the transfer payee's profile will affect two main indicators of the effectiveness of the fraud monitoring system - the numbers of false positives and the recall of fraud detection. The novelty of this research is implementing classification of payments behavior of fraudster and regular payees to managing fraud monitoring indicators.

2. Literature overview

The application of machine learning methods in fraud monitoring discussed in Kewei, X et al. (2021), in which the authors compare traditional machine learning methods such as Naive Bayes and SVM with their approach based on multiple techniques, including feature engineering, memory compression, mixed precision, and ensemble loss. Izotova, A et al. (2021) are considering detecting financial credit card fraud in unbalanced data. The authors consider the homogeneous and heterogeneous Poisson process to determine the probability of fraudulent operations, and compare the metrics of classification algorithms using various ensemble methods like LGBM, XGBoost, CatBoost. The authors also consider the problem of false positives, which are also discussed in our paper. Dubey, S. C et al. (2020) are

considering using the Neural Network algorithm with Backpropagation to score customer credit card transactions. Dornadula, V.N et al. (2019) in their research are using sliding window to aggregate the transactions made by users from different groups to recognize behavioral pattern of the group to choose the fraud pattern. Therefore, they compare metrics of different algorithms such as Isolation Forest, Logistic regression, Decision tree. Amarasinghe, T. et al. (2018) shows a review of machine learning and outlier detection algorithms, like Bayesian Networks, Recurrent Neural Networks, SVM, Fuzzy Logic, Hidden Markov Model, K-Means Clustering, K-Nearest Neighbor that can be integrated into anti-fraud systems.

Han et al. (2013) presented a high dimensional classification method, named the Copula Discriminant Analysis (CODA). The method includes joint distribution of parameters in the naive Bayes classifier with a decision rule that assigns a class label for some observation. He et al. (2016), and Wang et al. (2019) also presented some other researches about CODA. This method is also mentioned in the paper of Eva Scheungrab (2013), who discusses the use of copulas in the discriminant analysis algorithm. In this paper, the author analyzed standard methods for classification like linear discriminant analysis and quadratic discriminant analysis compared with CODA method.

3. Data Structure and Framework

In our research, we use such bank metrics like Fraud Basis Point (FBP) and False Positive Ratio (FPR). To classify drop person in our research we use such methods as Logistic Regression (Logit), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Artificial Neural Network (ANN), Random Forest, Gradient Boosting and CODA. The structure of ANN, Random Forest and Gradient Boosting hyperparameters and description of CODA algorithm are shown below. Coefficients of Logit and marginal effects are shown in Appendix C.

3.1. Notation

To evaluate performance of anti-fraud system we use Fraud Basis Point (FBP) and False Positive Ratio (FPR) as measure of lost or saved money and quality of algorithms. FBP – indicates the level of fraudulent transactions missed by the system. It is measured in basis points (0.01%). The main task of the antifraud monitoring system is to minimize this indicator.

$$FBP = \frac{fraud_l}{turnover} * 10000, \text{ where}$$

$fraud_l$ - volume of transactions the system did not block as fraudulent that later on appeared to be fraudulent when the customer contacted the bank that the transaction was passed without their own consent, $turnover$ - total money turnover in a bank system.

FPR – indicates the ratio of false fraudulent transactions detected by system to sum of all transactions, which passed by the system and it is measured in percepts.

$$FPR = \frac{G}{F_i + G + U} * 100, \text{ where}$$

F_i – transactions blocked by the fraud monitoring system when the client contacted the bank and reported it as a fraudulent transaction, G – transactions blocked by the system when the client informed the bank that they (the client) initiated the transaction, U – transactions blocked by the system but it was impossible to reach the client to confirm authorization.

3.2. Framework

The system for online transfers monitoring between bank clients for possible fraud should evaluate these transactions in a split second. This fact imposes certain restrictions on the use of machine learning methods to detect fraudulent transactions. In particular, in the chosen method, the model should be executed almost instantly for a large number of operations. It is proposed to consider an approach in which the FBP and FPR indicators are managed offline using machine learning methods and these restrictions will not affect the response time of the anti-fraud system.

It is proposed to evaluate all clients who actively use remote banking service in the mobile application for their similarity to the drop profile. This evaluation can be realized on a regular basis and then use the result as an additional parameter in an anti-fraud system online. Note that this approach, provides no limitations in the choice of machine learning methods, and it also becomes possible to use data for analysis from other sources, not only the history of using the mobile application and signs of the current transaction, but also, for example, the history of customer card transactions from other systems.

The idea of evaluating a client's profile for similarity to a drop based on the observations of employees who investigate cases of fraud in banks. Example of a legitimate client's transaction history shown in Table 1, Example of a dropper's transaction history shown in Table 2:

Date and time of transaction	**Card operation type**	**Type of service**	**Shop`s Merchant Category Code**	**Transaction amount**
01.02.2021 13:03	Purchase via pos	Car service	5533	26720,00
01.02.2021 13:10	Purchase via pos	Car service	5533	1500,00
02.02.2021 14:12	Purchase via pos	Gas station	5541	2202,78
08.02.2021 10:00	Purchase via pos	Pet Shop	5995	7399,00
10.02.2021 23:00	Purchase via ecom	Housing payment	4900	4500,00

Table 1. Example of a legitimate transaction history

Date and time of transaction	**Card operation type**	**Type of service**	**Shop`s Merchant Category Code**	**Transaction amount**
02.02.2021 04:03	Card balance request	ATM	6011	0
02.02.2021 04:12	Card balance request	ATM	6011	0
02.02.2021 14:12	Service connection	Gas station	6011	0
03.02.2021 06:33	Cash withdraw	ATM	6011	500
03.02.2021 06:35	Cash replenishment	ATM	6011	500

Table 2. Example of a dropper transaction history

Customers who actively use card services and make purchases naturally make a legitimate transaction history. Scammers do not have such an opportunity and their

transaction history consists of only technical operations, such as requesting a balance, transfers for small amounts, connecting additional banking services, etc.

3.3. Data description and data preparation

For data set in our study, we selected transfer transactions of a large Russian bank in a week, which the fraud monitoring system detected as a suspicious pattern and launched one of the processing scenarios for such transactions, for example, warning a client about possible fraud or completely rejecting an transaction. We also added transaction without triggers of the fraud monitoring system, but during the period under review, the client left a complaint that this operation was fraudulent. The missed and detected fraudulent transaction were combined and considered as a target class for binary classification. False positives are referred to as other class. We mark fraud as 1, false positives as 0 and get the following table (hits_fm):

Payer	**Payee**	**Class label**	**Date of transaction**
cl_1	cl_2	0	20.02.2021
cl_3	cl_4	1	20.02.2021
…	…	…	…
cl_m	cl_k	0	27.02.2021

Table 3. An example of a dataset (hits_fm) formed of a fraud monitoring system triggers.

The data set does not include operations that the fraud monitoring system did not consider suspicious. This follows from the purpose of our paper which is not to build a new classifier to detect fraud, but to study the possibility of managing the FBP and FPR risk indicators using machine learning methods.

We have included all fraudulent transactions and the share of class 1 in our sample is 1.2%. Therefore, we build a baseline from the parameters of the initial classifier of the fraud monitoring system, which we use in evaluation of performance of our framework by calculating the FBP and FPR.

Assuming that in the original classifier precision equals 0.012, and recall equals 1, we calculate f1 measure:

$$F_1 = \frac{2 * precision * recall}{precision + recall} = 0.024$$

We consider this F_1 value as a baseline and, if we can improve it, we can conclude that we have positive influence of the ML algorithm on the FBP and FPR indicators of the fraud monitoring system.

Next, we assume that if class 1 is specified in the hits_fm table, then the payee's profile can be referred to as a drop, in case of 0, the payee's profile is legitimate. For each payee, we can collect data with history of card transactions (Table 4). In order to approximate the experiment to the real transaction history, we use the history of operations that occurred in the two-week period before the first transfer to the client from the hits_fm. For example, the column «Transaction amount in group MCC_1» by client « cl_2 » contains the amount of all payments by the client that he made two weeks before the operation from Table 3. Selected groups of Merchant Categories Code shown in Appendix A.

Bank client	**Transaction amount in group MCC_1**	**Transaction amount in group MCC_2**	**…**	**Transaction amount in group MCC_N**
cl_2	40000	0	…	11112
cl_4	0	30000	…	0
…	…	…	…	…

Table 4. Clients profile data set collected from customer card transactions history

We mentioned that feature selection and feature engineering is not the main aim of this paper. In this case, Table 4 collect to check our proposition about the difference in behavior between a legitimate client and a drop. Finally, this data was filtered by following criteria:

a. Excluded transactions blocked by the system for which bank did not receive feedback from the client about legitimacy of the operation.
b. Excluded false positive transactions blocked by the system.

c. Excluded transactions for which the history of card transactions by payee were not found.

After these filters our data set decreased by 3.9%. We split our data set to test and train sample in a ratio of 80% to 20%. It should be mentioned that algorithms were not validated for out-of-time samples.

In addition, we did not apply oversampling methods to increase the size of the minor class. Descriptive statistics, feature importance and correlation matrix within classes are shown in Appendix B.

3.4. Description of CODA algorithm

The idea of the CODA is to build maximum likelihood functions, taking into account the copula density of explanatory variables within the class. Therefore, CODA discrimination rule is:

$$k = \arg\max(\sum_{i=1} \ln(f_i^k) + \ln(p_k) + \ln\{c^i(F(_1(x_1),...,F_n(x_n)\}\text{, where}$$

f_i^k - denotes the marginal density of variable i in class k, p_k - is the prior probability of class k, $c^i(F(_1(x_1),...,F_n(x_n)$ - probability distribution function (PDF) of the bivariate copula of explanatory variables within class.

To realize CODA method, the following actions were performed:

• Determined the distribution function of the values of each explanatory variable within the class,

• Determined the prior probabilities of the class,

• Determined copula family for explanatory variable within the class,

• Determined the values of the distribution density of the known copulas for each of the observations

• Build a classifier, which takes previously calculated values, and returns a class label for each class.

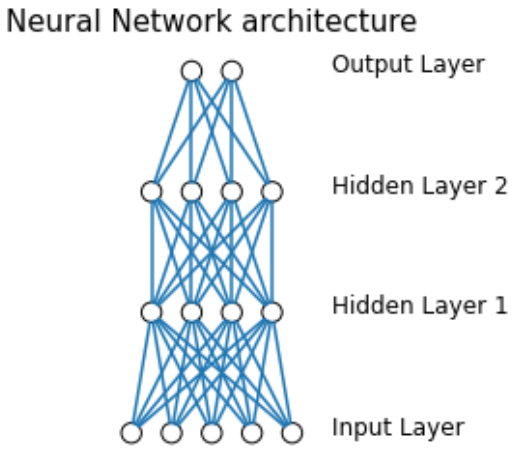
To build CODA classifier source data was reduced to 6 features, selected by ANOVA method, which shown in Appendix B.

3.5. Description of ANN structure, Random Forest and Gradient boosting parameters

To build ANN we used MLPClassifier from sklearn library for Python. We transform our data with «power_transform» approach from module sklearn.preprocessing. We also try «normalize» and «minmax_scale» transform, but it was not affected well. «Lbfgs» solver with «Relu» activation function were used. Structure of ANN and graphic visualization shown in Table 5 and Picture 2. Also source data was reduced to 5 features, selected by ANOVA method.

Layer	Number of neurons
Input layer	5
Hidden layer 1	4
Hidden layer 2	4
Output layer	2

Table 5. Structure of ANN.



Picture 2. ANN structure visualization.

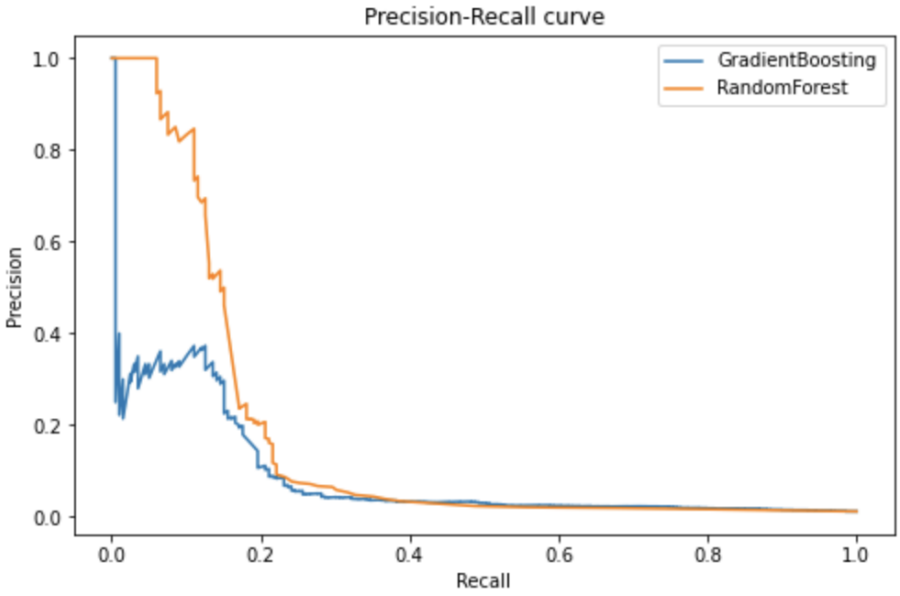
To build Random Forest classifier, we used RandomForestClassifier from Sklearn. We select «log2» as max_features and n_estimators equals 300. For Gradient boosting we used GradientBoostingClassifier from Sklearn with n_estimators equals 300, max_depth equals 7 with «deviance» loss function. Logit, QDA, LDA selected from «statsmodels» library in Python.

3.6. *Classification metrics and framework conclusions.*

Classification metrics are show in Table 6. We select F1 score as a measure of performance for our classification. Therefore, Random Forest and Gradient Boosting shows better performance by this metric, for these methods we plot a precision-recall curve (Picture 3).

	Precision	Recall	F1_score
Random Forest	0,69	0,12	0,21
XGBoost	0,28	0,15	0,2
QDA	0,01	0,68	0,03
Logit	0,21	0,02	0,03
Baseline	**1,0**	**0,012**	**0,024**
LDA	0,4	0,01	0,02
ANN	1	0,01	0,01
CODA	0,01	0,01	0,01

Table 6. Classification metrics.



Picture 3. Precision-recall curve for Random Forest and Gradient Boosting.

The results of experiments showed the efficiency of the proposed framework - integrating the history of customer card transactions as additional features in evaluation of transfers initiated in the bank's mobile application. This approach can be developed to improve the performance indicators of FBP and FPR by generating new features from history of card transactions and search for the best parameters for machine learning classifier. At this stage, we can conclude that classifiers based on decision tree analysis show the best result in terms of managing FBP and FPR indicators.

5. Fraud and operational risks

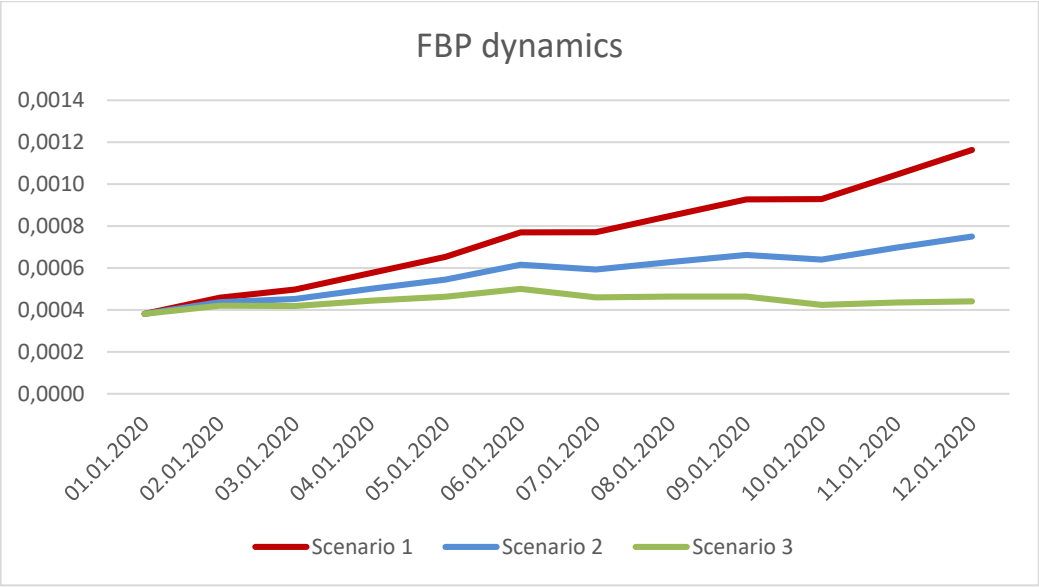
Previously, we did not consider the parameters FBP and FPR in terms of business value. These indicators are useful in assessing the quality of the anti-fraud system and in managing the risks associated with remote banking processes. FBP shows the overall efficiency of fraud monitoring systems in terms of missed funds and allows you to assess how close the bank is to the critical values of international payment systems or local banks regulator. FPR also shows the effectiveness of the system, however, the bank always has to balance the FBP and FPR, choosing the best ratio of these indicators. Bank also has to follow the constant level of approval rate, determinable by international payment systems and other members of e-

commerce. Theoretically, a fraud monitoring system can block all operations, and then FBP will tend to a minimum, but high FPR value leads to increases of reputational risk, which may lead to customer churn and negative reviews in mass media, and increased costs for offline analysis of suspended operations. The load on the fraud monitoring system and related remote banking systems is also increasing, which can lead to disruption of servers and provoke IT incidents. Including machine learning methods in bank process follows continuous validation and assessment of the quality of the model. This quality is directly related to the bank's financial losses for the formation of a fund for payments to affected customers and the formation of a minimum regulatory capital to cover losses. Suppose a case in which the volume of fraudulent attempts tripled in a calendar year, and the bank's turnover increased by 52%. Each line reflects the dynamics of the FBP indicator, depending on three different scenarios:

1) The model is not validated or updated.

2) The model is validated, but its performance remains constant. This case is better than the previous one in relation to the FBP dynamics; however, with the growing volume of fraud, the dynamics remains negative.

3) The model is validated, the performance is growing, and FBP has a positive trend.



Picture 4. Three different scenarios of FBP dynamics.

Due to the growth of the market for remote banking systems, it becomes necessary to introduce regulatory requirements from the regulators and FPB and FPR can become such indicators. These indicators can be decomposed within a financial institution, taking into account its organizational structure, for example, antifraud system downtime costs, call center costs, IT equipment costs, hiring specialists, etc. Therefore, a framework for the dependence of the macro indicators FBP and FPR on the work of the company's structural divisions can build.

6. Conclusions

In this article, we have proposed a possible scenario for managing cyber risk using machine learning techniques. A framework for improving the antifraud system using data enrichment from the history of the client's operation is presented. The proposed framework showed an increase in the quality metrics of the algorithm compared to the previous solution, which led to an increase in the macro indicators FBP and FPR measured at F1-score (Table 6.). The importance of FBP and FPR management from the point of possible losses in realization of cyber risks of external fraud is shown.

Further, the authors plan to extract more features from a client’s transaction history, try to optimize hyperparameters of machine learning algorithms. In addition, we will decompose FBP and FPR indicators to show approaches to managing of cyber risks.

Appendix A. Description of Merchant Category Code group

Column label	Merchant Category Code group
A	Car rent
C	Cash withdrawals, money transfers
F	Restaurants, fast food
H	Hotels
J	Utilities
O	Professional services (medicine, healthcare, etc.)
R	Retail Stores
T	Wholesale Stores
U	Gambling
X	Flights
Z	Money transfers
Q	Banking services in the office

Table 7. Description of MCC groups.

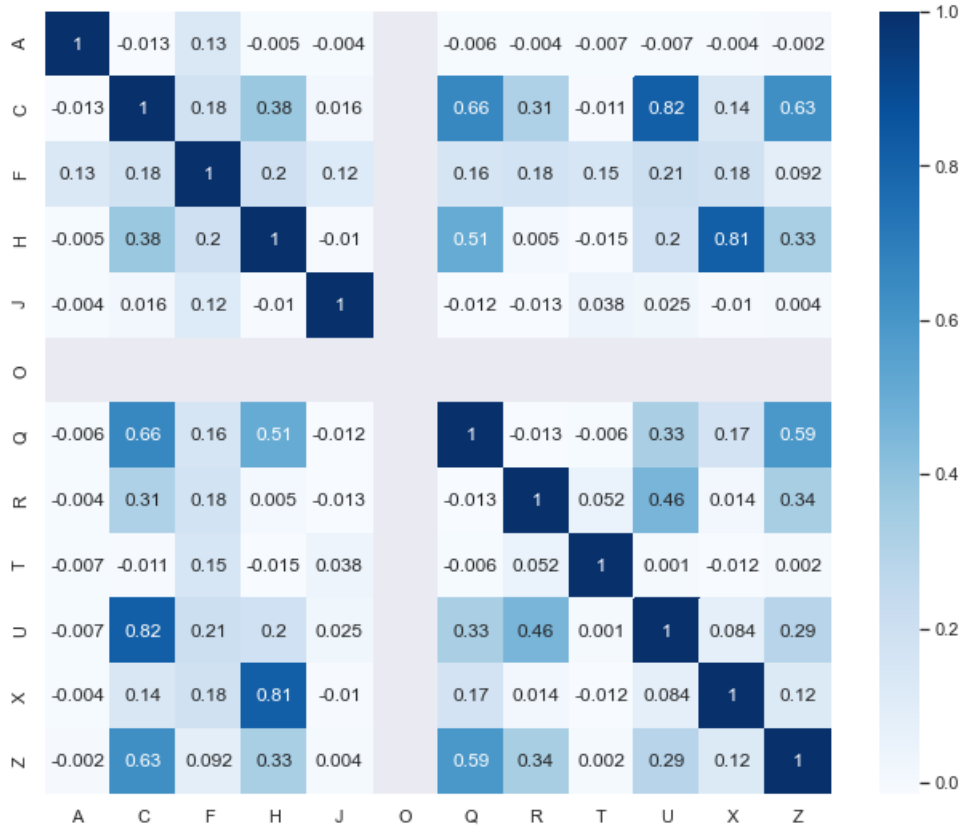
Appendix B. Data description.

	A	C	F	H	J	O	Q	R	T	U	X	Z
mean	110	417560	2345	1729	222	24	749	73161	420	296545	3737	250011
std	2931	1028781	122660	81139	2768	1150	15540	260628	17320	924418	78068	1150206
min	0	0	0	0	0	0	0	0	0	0	0	0
25%	0	4230	0	0	0	0	0	2475	0	1500	0	300
50%	0	37975	0	0	0	0	0	11317	0	20600	0	16904
75%	0	310000	950	0	0	0	0	42673	0	150732	0	114501
max	4,23E+05	2,37E+07	3,49E+07	7,75E+06	2,50E+05	1,99E+05	1,48E+06	2,20E+07	3,33E+06	2,52E+07	1,99E+07	2,37E+08

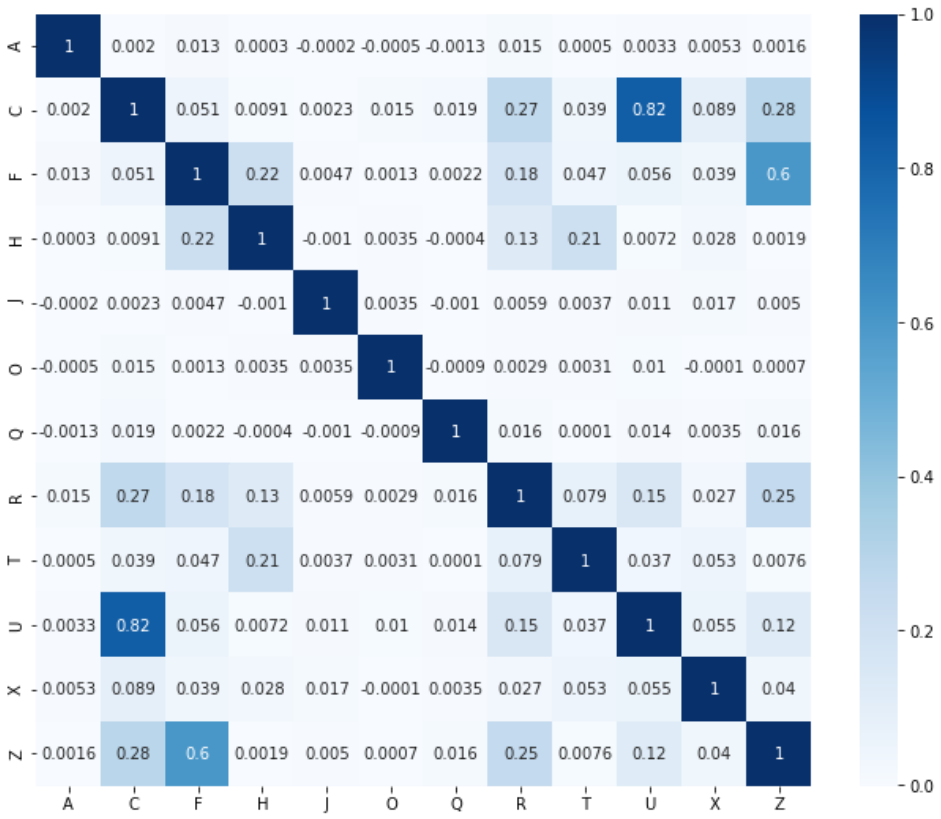
Table 8. Descriptive statistics for data in class 0

	A	C	F	H	J	O	Q	R	T	U	X	Z
mean	16	192822	902	117	111	0	5211	70896	149	131215	323	153351
std	415	531192	3444	1229	1461	0	47155	221827	1024	436348	2858	451968
min	0	0	0	0	0	0	0	0	0	0	0	0
25%	0	3000	0	0	0	0	0	1297	0	0	0	953
50%	0	19800	0	0	0	0	0	7494	0	5071	0	11900
75%	0	99800	431	0	0	0	0	36972	0	33510	0	68801
max	1,26E+04	4,58E+06	7,26E+04	2,76E+04	2,98E+04	0,00E+00	5,40E+05	2,57E+06	1,57E+04	4,78E+06	5,69E+04	5,65E+06

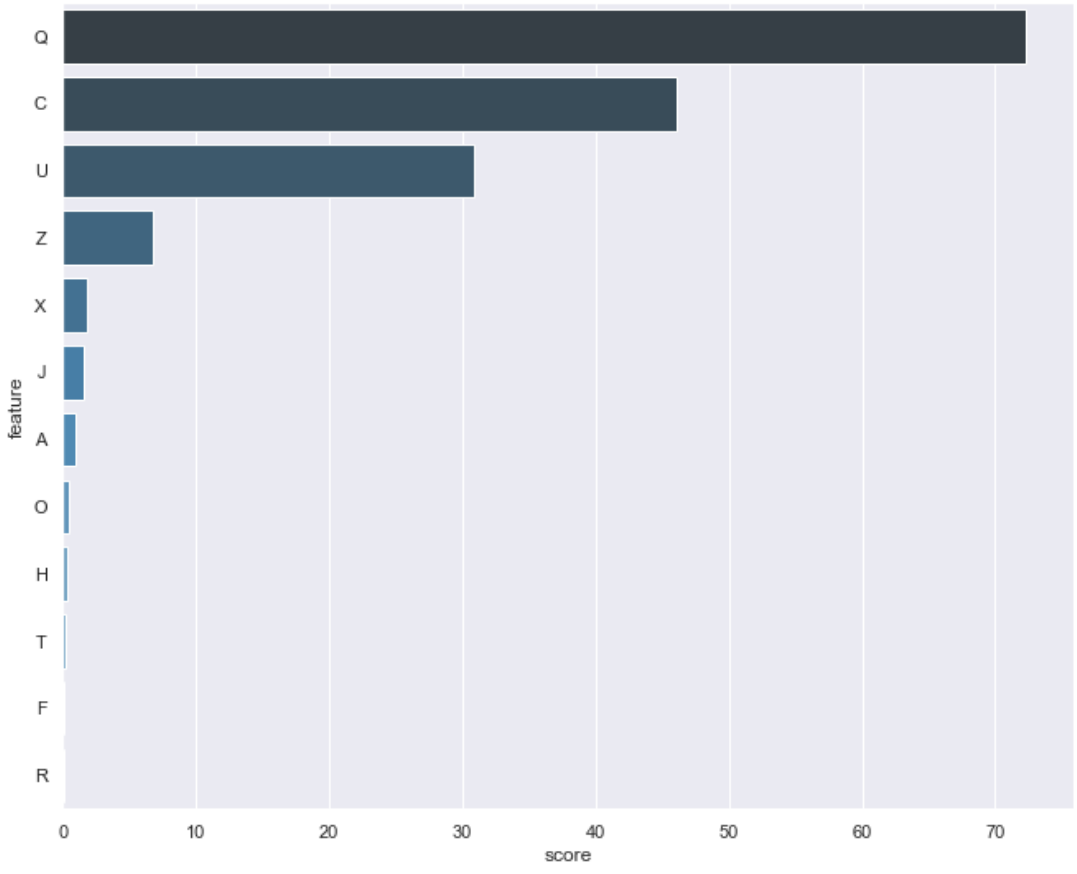
Table 9. Descriptive statistics for data in class 1



Picture 3. Correlation matrix within class 1



Picture 4. Correlation matrix within class 0



Picture 5. Feature importance calculated by ANOVA

Appendix C. Logit model summary.

In our study we have complete quasi-separation because of a fraction 0.31 of observations can be perfectly predicted. So, marginal effects are not significant. Models summary and marginal effects of modules in full data set and reduced data set are shown in Pictures 6, 7, 8, 9.

Logit Regression Results

Dep. Variable:	rec_class	No. Observations:	
Model:	Logit	Df Residuals:	
Method:	MLE	Df Model:	11
Date:	Thu, 15 Jul 2021	Pseudo R-squ.:	-3.573
Time:	12:34:23	Log-Likelihood:	-19541.
converged:	False	LL-Null:	-4272.8
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
A	-0.0005	0.000	-2.791	0.005	-0.001	-0.000
C	-3.52e-06	3.87e-07	-9.098	0.000	-4.28e-06	-2.76e-06
F	-0.0014	4.58e-05	-29.658	0.000	-0.001	-0.001
H	-7.772e-05	3.64e-05	-2.136	0.033	-0.000	-6.42e-06
J	-0.0005	6.34e-05	-7.685	0.000	-0.001	-0.000
O	-1.7831	715.828	-0.002	0.998	-1404.781	1401.215
R	-3.031e-05	1.2e-06	-25.225	0.000	-3.27e-05	-2.8e-05
T	-0.0018	0.000	-12.989	0.000	-0.002	-0.002
U	-9.545e-06	5.63e-07	-16.952	0.000	-1.06e-05	-8.44e-06
X	-0.0005	4.69e-05	-11.077	0.000	-0.001	-0.000
Z	-1.252e-05	4.88e-07	-25.669	0.000	-1.35e-05	-1.16e-05
Q	1.669e-06	1.8e-06	0.925	0.355	-1.87e-06	5.2e-06

Picture 6. Logit model summary for full data set.

Logit Marginal Effects

Dep. Variable:	rec_class
Method:	dydx
At:	mean

	dy/dx	std err	z	P>|z|	[0.025	0.975]
A	-8.267e-29	1.37e-24	-6.06e-05	1.000	-2.68e-24	2.68e-24
C	-5.607e-31	9.26e-27	-6.06e-05	1.000	-1.81e-26	1.81e-26
F	-2.165e-28	3.58e-24	-6.06e-05	1.000	-7.01e-24	7.01e-24
H	-1.238e-29	2.04e-25	-6.06e-05	1.000	-4.01e-25	4.01e-25
J	-7.763e-29	1.28e-24	-6.06e-05	1.000	-2.51e-24	2.51e-24
O	-2.841e-25	4.58e-21	-6.21e-05	1.000	-8.97e-21	8.97e-21
R	-4.828e-30	7.97e-26	-6.06e-05	1.000	-1.56e-25	1.56e-25
T	-2.83e-28	4.67e-24	-6.06e-05	1.000	-9.16e-24	9.16e-24
U	-1.521e-30	2.51e-26	-6.06e-05	1.000	-4.92e-26	4.92e-26
X	-8.28e-29	1.37e-24	-6.06e-05	1.000	-2.68e-24	2.68e-24
Z	-1.994e-30	3.29e-26	-6.06e-05	1.000	-6.46e-26	6.45e-26
Q	2.658e-31	4.39e-27	6.06e-05	1.000	-8.6e-27	8.6e-27

Picture 7. Logit model marginal effects for full data set.

Logit Regression Results

Dep. Variable:	rec_class	No. Observations:	
Model:	Logit	Df Residuals:	
Method:	MLE	Df Model:	4
Date:	Thu, 15 Jul 2021	Pseudo R-squ.:	-4.224
Time:	12:34:23	Log-Likelihood:	-22320.
converged:	True	LL-Null:	-4272.8
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
Q	-1.729e-06	1.45e-06	-1.194	0.232	-4.57e-06	1.11e-06
C	-1.013e-05	4.89e-07	-20.723	0.000	-1.11e-05	-9.17e-06
U	-2.006e-05	7.17e-07	-27.989	0.000	-2.15e-05	-1.87e-05
Z	-1.876e-05	5.7e-07	-32.907	0.000	-1.99e-05	-1.76e-05
X	-0.0011	6.57e-05	-16.862	0.000	-0.001	-0.001

Picture 8. Logit model summary for reduced data set.

Logit Marginal Effects

Dep. Variable:	rec_class
Method:	dydx
At:	mean

	dy/dx	std err	z	P>|z|	[0.025	0.975]
Q	-1.043e-14	9.31e-15	-1.121	0.262	-2.87e-14	7.81e-15
C	-6.112e-14	1.75e-14	-3.486	0.000	-9.55e-14	-2.68e-14
U	-1.21e-13	3.46e-14	-3.493	0.000	-1.89e-13	-5.31e-14
Z	-1.132e-13	3.28e-14	-3.447	0.001	-1.77e-13	-4.88e-14
X	-6.681e-12	1.68e-12	-3.989	0.000	-9.96e-12	-3.4e-12

Picture 9. Logit model marginal effects for reduced data set.

References

Payment Fraud Statistics, Trends & Forecasts (2020), n.d. URL https://www.merchantsavvy.co.uk/payment-fraud-statistics/ (accessed 7.2.21).

Kewei, X., Peng, B., Jiang, Y., Lu, T., 2021. A Hybrid Deep Learning Model For Online Fraud Detection, in: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). Presented

at the 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 431–434. https://doi.org/10.1109/ICCECE51280.2021.9342110

Izotova, A., Valiullin, A., 2021. Comparison of Poisson process and machine learning algorithms approach for credit card fraud detection. Procedia Computer Science 186, 721–726. https://doi.org/10.1016/j.procs.2021.04.214

Dubey, S.C., Mundhe, K.S., Kadam, A.A., 2020. Credit Card Fraud Detection using Artificial Neural Network and BackPropagation, in: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). Presented at the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 268–273. https://doi.org/10.1109/ICICCS48265.2020.9120957

Dornadula, V.N., Geetha, S., 2019. Credit Card Fraud Detection using Machine Learning Algorithms. Procedia Computer Science 165, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

Amarasinghe, T., Aponso, A., Krishnarajah, N., 2018. Critical Analysis of Machine Learning Based Approaches for Fraud Detection in Financial Transactions, in: Proceedings of the 2018 International Conference on Machine Learning Technologies - ICMLT ’18. Presented at the the 2018 International Conference, ACM Press, Jinan, China, pp. 12–17. https://doi.org/10.1145/3231884.3231894

Han, F., Zhao, T., Liu, H., 2013. CODA: High Dimensional Copula Discriminant Analysis. Journal of Machine Learning Research 16, 43.

He, Y., Zhang, X., Wang, P., 2016. Discriminant analysis on high dimensional Gaussian copula model. Statistics & Probability Letters 117, 100–112. https://doi.org/10.1016/j.spl.2016.05.018

Wang, B., Sun, Y., Zhang, T., Sugi, T., Wang, X., 2020. Bayesian classifier with multivariate distribution based on D-vine copula model for awake/drowsiness interpretation during power nap. Biomedical Signal Processing and Control 56, 101686. https://doi.org/10.1016/j.bspc.2019.101686

Scheungrab, E. (2013). Copula based discriminant analysis with application (Master thesis, Technical University of Munich, Munich, German). Retrieved from https://mediatum.ub.tum.de/doc/1166376/1166376.pdf