FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION
OF HIGHER EDUCATION
«NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS»

Faculty of Business and Management

STUDY OF THE POSSIBILITIES OF USING DEEP LEARNING FOR GOVERNMENTAL CONTRACT RISK PREDICTION

Coursework of **Eliseev Dmitrii**
1st course, Master's program "Business Informatics"

Scientific supervisor
Golov N.E

Moscow, 2019

TABLE OF CONTENTS

INTRODUCTION

Government spends significant part of its budget on procurement of goods and services to meet need of public agencies and organizations annually. In Russia about 25 300 billion rubles was spent on government procurement in 2018 that constituted 24,4% of GDP ("Unified Public Procurement Information System," 2019). Taking into account the amount of money spinning in public procurement, it is important to conduct public procurement effectively.

Transparent and competitive procurement process is one of strategic priorities for Russian authorities. This is one of the reasons why e-procurement system – the Unified Public Procurement Information System (UPPIS) was launched in Russia on 1st January 2011. Since that day information about every public contract has been being collected and stored in UPPIS and is publicly available on its website.

The consolidation of such amount of data with the current level of development of techniques for data processing and analysis represents an opportunity for the government to accelerate its progress in the fight against inefficiency in public procurement. Detection of inefficiency constitute a good example of area where machine learning can be applied, and this area will be in the focus of the current research.

Machine learning is frequently used to find anomalies, abnormal cases or behavior. Some of the examples are fraud detection, credit scoring, cancer prediction etc. Machine learning seems to have promising applications in procurement process and one of them is detection of inefficient contracts. Machine learning can help controlling bodies detect such contracts and timely impact on contractors. As a result, effectiveness of regulatory authorities may increase. Consequently, this might give huge saving in procurement process due to diminishing of operating expenses and might have positive effect on budget usage of Russia, reducing losses on procurement process.

As for practice, machine learning techniques are not used yet to detect inefficient contracts in procurement process in Russia. Instead manual inspections of contracts are conducted.

This work is devoted to development of machine learning model for risk prediction of public contracts. This model should be able to assess riskiness of contract using available in UPPIS data by the moment of signing. Such approach has several advantages over traditional methods such as ability to work every day of the week, ability to build predictions for millions of contracts and to detect difficult patterns which are hidden from human eye. So, the aim of this study is to evaluate possibilities of using machine learning especially deep learning techniques for detection of inefficient contracts.

At this moment, it is important to explain what is considered as inefficiency in this work. Contract is considered as inefficient (bad) if it is terminated by any reason (termination by mutual agreement, one-side termination by customer, termination by court order etc.), contract is defined as efficient (good) in all other cases. The termination of contract is a huge loss for public customer in terms of money and time, that's why termination was chosen as a criterion of efficiency.

Subject of the study is the process of government procurement. **Object of the research** is model for risk prediction of public contracts developed with the help of machine learning techniques and trained on historical data. Risk is defined as probability of contract termination. **Objective** of this work is to create solution based on machine learning and deep learning techniques for risk assessment of contracts governed by 44-FL (federal law about "Contract system in the field of purchase of goods, works, services for ensuring state and municipal needs"). The solution should be a trained model, which takes contract characteristics available in UPPIS as input and returns probability of contract termination as output.

Four machine learning algorithms (logistic regression, random forest, gradient boosting on decision trees, neural networks) were compared to establish the best one for risk prediction of government contracts. Predictive quality was assessed by area under the receiver operating curve (AUC ROC) as it may be important to vary threshold to find optimal values of sensitivity and specificity. Also, F-score and logistic loss were used. Unique sample was collected. The dataset contains 38 variables and more than 300 thousand public contracts throughout Russia for the period between 2016 and 2018.

This work is continuation of the research started in previous year as a part of diploma thesis for bachelor's degree (Eliseev, 2018b; Eliseev & Romanov, 2018). Key results of previous work are listed below.

- Data collection about government contracts for two Russian regions (Yaroslavskaya and Tulskaya oblast). Dataset contains about 37 thousand observations and 36 variables. It is publicly available under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 license (Eliseev, 2018a).
- Comparison of three algorithms for classification – logistic regression, random forest and gradient boosting on decision trees. The best result was achieved by third model with specificity = 0.9 and sensitivity = 0.94.

In this research some changes and improvements were made. Main of them are given below.

- Collection of dataset about public contracts over all Russian regions with 38 variables and more than 300 thousand observations.
- New features (features for territory patterns, level of contract and expertise of customer and supplier in different levels of purchases), more profound preprocessing.
- Change of definition of good contract (in previous study contract was considered as good if it was finished or was terminated by mutual agreement and was executed on more than 60%).
- In addition to classical algorithms from classification used in previous study deep learning model based on neural network was implemented.

The research is composed of six sections including this introduction. In Section 2 background of monitoring process in procurement system of Russia is given and some of the theoretical foundations of deep learning is described. Moreover, challenges that might be faced during implementation of machine learning algorithms for risk prediction of contracts are discussed. In addition, overview of research of machine learning usage for risk prediction is given. Section 3 characterizes methods used – dataset, machine learning algorithms, approach to dealing with imbalanced data, training / testing / validation strategy. In Section 4 the results are presented – main information about model performance is given. Section 5 is devoted to discussion of results, the implication for theory and practice, limitations and direction of further research will be defined. Finally, the conclusion takes place in section 6.

Control and monitoring in public procurement system of Russia

It can be defined four types of procedures that are used to control and monitor government procurement:

- Inspections held by Federal Antimonopoly Service (FAS),
- Consideration of complaints of procurement participants,
- Register of unscrupulous suppliers (RUS) regulated by FAS,
- Control exercised by the Federal Treasury.

All of these methods imply manual work. Basing on analytical report of Ministry of Finance of the Russian Federation, in 2018 about 8500 inspections were made by representatives of FAS and the Federal Treasury ("Consolidated analytical report on the results of the monitoring of purchases, goods and services for the provision of state and municipal," 2019). So, if controlling bodies were working daily, this would mean over 23 inspections per day. Total number of procurement placements in 2018 is more than 3,2 million. Consequently, less than 0,27% from placements were inspected. Controlling agencies have budget as a constrain to possible number of inspections per year, and, as it seen, they are able to check only a few contracts.

Most of inspection made by FAS are unscheduled inspection, they constituted about 97% of inspections in 2018. The object of unscheduled inspection is chosen basing on heuristics or randomly. Therefore, unscheduled inspections detect violations, but they rarely detect contracts with significant violation which are about to be terminated.

In 2018 it was considered over 83,4 thousand complaints and most of them were about regional purchases (46%), then follow complaints about municipal procurement (34%). According to the results of consideration, 28 thousand complaints were found to be substantiated.

In 2018 FAS has added over 7 thousand suppliers in RUS, that is on 27% more than in previous year. Usually, suppliers get included in RUS after one-sided contract termination. This emphasize that there is still a lot of room for improvement in controlling and monitoring processes.

The Federal Treasury has revealed over 1,8 million mismatches in information, added by contractors in UPPIS, and conducted about 1,4 thousand inspections.

It is seen that there are many controlling bodies which are devoted to make procurement process efficient. Meanwhile, such approach seems to be not scalable enough and cannot cover even 1% of contracts. This opens opportunity for suggested in this work approach – automated model for risk prediction of contracts based on machine learning techniques, which will be able to check millions of contracts in minutes for detection the riskiest.

Introduction to deep learning

Deep learning is a subpart of machine learning. The main difference from classical machine learning algorithms is another architecture of a model inspired by structure of biological nervous system. Such models are called artificial neural network (ANN) or just neural networks (NN).

As analogy to biological neuron the input connectors and output connectors between layers of artificial neurons represent dendrites and axons (Fig. 1). Weights on connectors are updated during back-propagation which is a very simplified analogy to electro-chemical process of biological neuron activation. In other words, the strength of the connections between artificial neurons is strengthened or weaken through repeated training, the same as the power of impulses going between biological neurons transforms during lifecycle of biological creature.

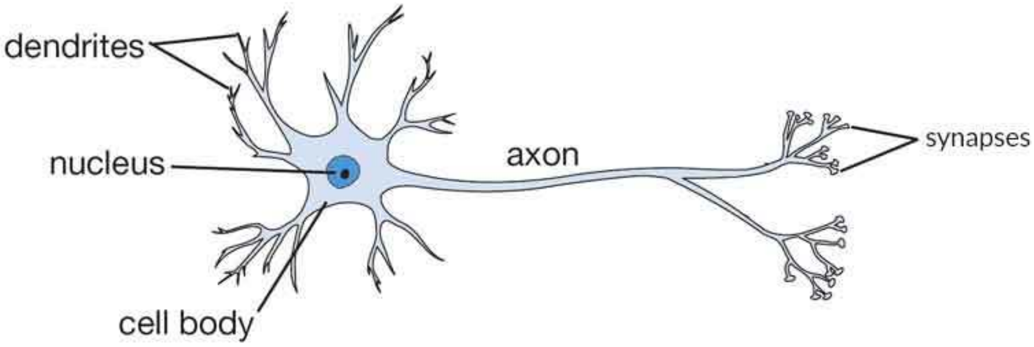


Figure 1. Biological neuron

Artificial neuron takes input (signals) multiplied by weights (strength of impulse), then sum them and add bias (Fig. 2). To add nonlinearity, result is sent to activation function (Sigmoid, Tanh, Linear, ReLU etc.) and output of activation function is sent further to next neurons or output.

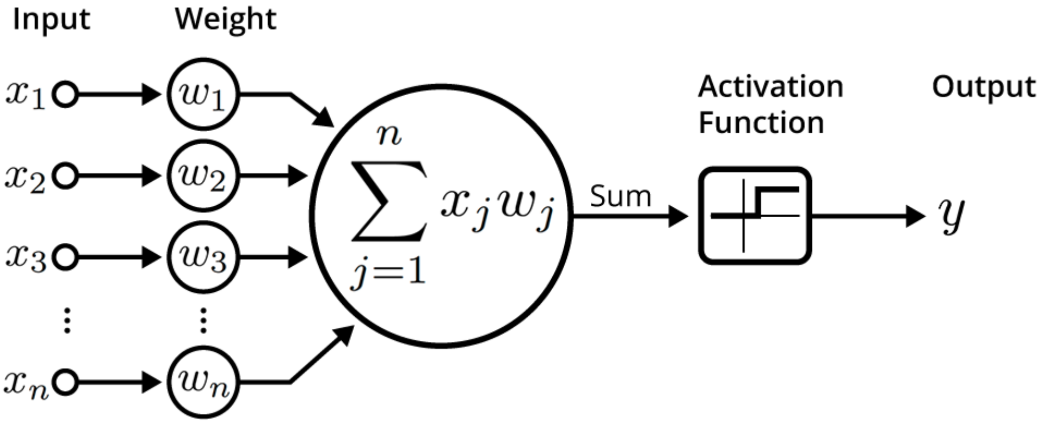


Figure 2. Model of artificial neuron (Saxena, 2017)

NN is nonlinear and non-parametric model and has very wide area of industrial applications because such algorithm is able to detect very difficult patterns in data and is prefect for working with unstructured data (images, text, sound etc.). On the other hand, NN is a “black box” in terms of interpretably, and NN, especially deep NN, demands huge amount of data and computational resources.

NN consists of layers – there are always input and output layers and can be a lot of hidden layers between them (Fig. 3). Usually, term “deep learning” is used for NN with many hidden layers.

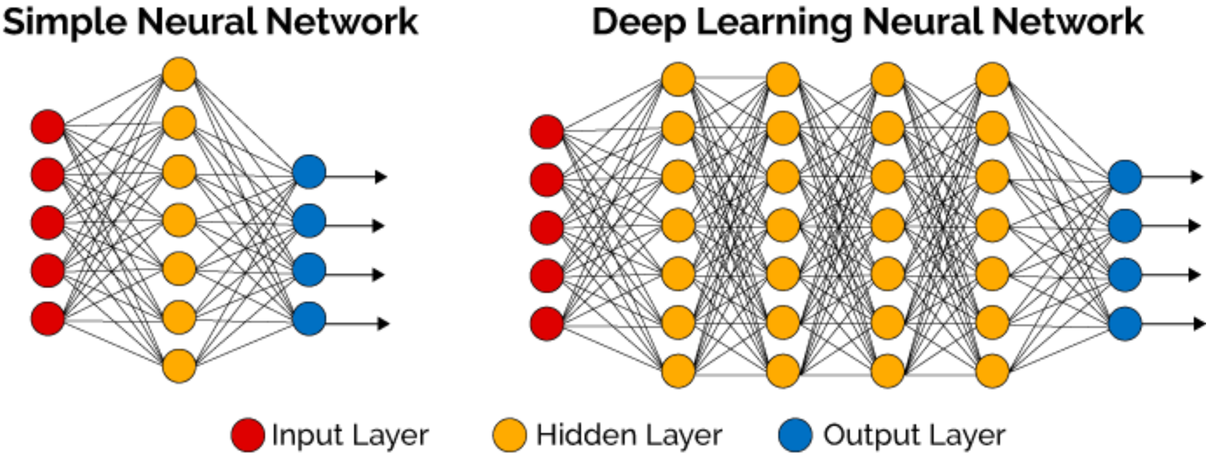


Figure 3. Neural Network Architecture (Vázquez, 2017)

There were invented different type of NN, the main of them are:

- classic feedforward NN (or multilayer perceptron) – are used for typical tasks of classification and regression for structured data;
- recurrent NN (RNN) – are used for working with sequences, e.g. text, video, sound;

- convolutional NN (CNN) – are used for working with images;
- generative NN – are used for creation of new content, e.g. music, art, furniture.

Architecture of the neural network and activation functions are hyperparameters, in other words, they are defined manually. There are best practices for different type of task (classification, regression, clustering etc.) and data (structured data, images, text etc.) but still a lot of time is spent on hyperparameter tuning during developing model based on NN. Usually, the best architecture is achieved by testing several approached.

As classical machine learning algorithms neural network has techniques for regularization to prevent overfitting of model – dropout, L1 and L2 regularization. The main idea of dropout is to randomly drop neurons during training to prevent co-adaption among neurons. In a large network dropout training provides better performance improvement and is more robust than L2 regularization, in opposite, L2 is better for small networks (Phaisangittisagul, 2016).

As for L1 and L2 regularization these approaches suggest adding to cost function one more addend – sum of weights for L1 and sum of squared weights for L2 (Formula 1, 2). Lambda is the regularization hyperparameter.

$$Cost\ function\ (L1) = Loss\ function + \lambda \sum |w| \tag{1}$$

$$Cost\ function\ (L2) = Loss\ function + \lambda \sum |w|^2 \tag{2}$$

As a result, L2 regularization forces the weights to decay toward zero but not exactly zero. L1 regularization reduces weights exactly to zero and is very useful for model compressing and simplification.

Due to the fact that in this work binary classification model is built to differ risky contract from not risky, best practices for solving classification task with neural networks should be mentioned. There is common knowledge about last layer transformation and loss function that produces highest results in classification task – Softmax transformation and cross entropy loss. Let us describe these concepts. By default, neural network produces float values in any range as output, but in classification task probabilities from 0 to 1 describing likelihood of belonging to given classes are needed. This result is achieved with Softmax transformation (Formula 3).

$$y_i = \frac{e^{y_i}}{\sum_{n=1}^{N} e^{y_n}} \tag{3}$$

As a loss function for classification cross entropy loss is recommended. Formula 4 shows equation for cross entropy loss in case of binary classification.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\big(p(y_i)\big) + (1 - y_i) \cdot \log\,(1 - p(y_i)) \qquad (4)$$

where N – number of observations, y_i – true value of the target variable, $p(y_i)$ – predicted probability by model.

Machine learning for risk prediction

Machine learning is widely used for risk prediction. Some of the examples include credit scoring (Baesens, Lessmann, Seow, & Thomas, 2015), cancer detection (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015), prediction of outage of equipment (Settemesdal, 2019). A lot of analysis about usage of machine learning for risk prediction was made especially with application to medicine and healthcare (Kruppa, Ziegler, & Konig, 2012).

But, to the best of our knowledge, there are a few papers, which are covering question of using machine learning algorithms to optimize procurement process. There may be several hinderance to conduct research in this field, such as data absence, restricted access to date or just dirtiness of data.

One of the most relevant paper to the theme of this work was by Gallego et al (Gallego, Rivero, & Martinez, 2018). Authors built three machine learning model – lasso logistic regression, conditional inference tree and a gradient boosting machine using SECOP database of the Colombian government and enriching it with data from several other public authorities Used sample included 2,2 million of observations and more than 300 variables. On average models had achieved AUC ROC from 0.78 till 0.93. SMOTE technique and changing cost function were used for managing with imbalanced data. Model built on data oversampled by SMOTE technique produced about 1% higher AUC ROC. The most important features were size of contract (budget), lag between the day contract was awarded and the first day of execution, geographical and industry-specific variables.

Challenges to consider

Is it important to acknowledge that the use of machine learning models as a tool for detection of risky contracts is not exempt of pitfalls and challenges. The main of them will be described below.

Firstly, it is important to understand limitations of the data and preprocess it very carefully. Bad data as input will produces bad model as output. Sometimes, there is a room for strategic misinformation and data in UPPIS or other sources may be wrong. Not always information about contract, suppliers and customers is consistent, especially when aggregation from different sources of information is needed. In fact, it would be difficult to consolidate high quality data fully covering subject area without cooperation between different government offices and private companies.

Secondly, in classification task there is a common trade-off – distinction between precision and recall, or, in other words, trade-off between false positive and false negative. For instance, aggressive classifier will predict many contracts as inefficient. Consequently, false positive rate will be high and government controlling bodies will have to spend more resources to inspect contracts, but the probability of missing any risky deal will be low. Conversely, a conservative model will minimize false negative. If model predicts that contract is risky, it is very likely that it will be terminated in fact, but this model will have lower recall, missing some bad contracts. Such trade-off may be more effective for public agencies which have scare resources for inspections. Depending on objective function of controlling bodies and available resources, model can be tuned in order to maximize usefulness in terms of desired true positive rate (sensitivity) and true negative rate (specificity).

Thirdly, customers, suppliers and bureaucrats are not static agents and will try to adapt to new conditions after solution implementation. This is a reason why machine learning model should be dynamic and adaptive as well and constantly uses new data to learn. Moreover, data sources are not static as well – UPPIS is developing, in other system API or response format may be changed and so forth. Thus, flexibility is in a demand from both data science and data engineering viewpoint. After deployment in production regular monitoring and support of model will be needed.

METHODS

Code for data collection, data analysis, data preprocessing and model development is available on GitHub (Eliseev, 2019).

Data

Data description

UPPIS serves as a tool where all data about public contracts in Russia is collected and published. Unfortunately, the format how data is presented is oriented for manual analysis. It is difficult to get data in machine readable form without developing web scraper. This prevents scientists from researching government procurement with quantitative methods.

Fortunately, NPO "Krista", developer and integrator of enterprise information systems working in B2G sector, has provided an access to a copy of the database used by UPPIS. This database was deployed in Microsoft SQL Server. Its size was around 100 Gb and it contained data about public contracts since 2011, when UPPIS was launched.

Sample was collected using T-SQL scripts. Contracts signed not earlier than 1st of January 2016 were taken because legislation in area of government procurement is developing quickly and data about previous contracts is too outdated. In total there were over 4,5 million contracts and 3% of them were bad (prematurely terminated).

Contracts, which were purchases from a single supplier, were excluded. These contracts tend to be not risky because only one supplier can provide required services or goods, this is a reason why customer and supplier usually know each other for a long period of time and have strong and trustworthy relationship.

As mentioned above, bad contracts constitute 3% of all contracts – the sample is imbalanced. Undersampling was used to deal with imbalance. Collected dataset contains 572 065 observations about 308 273 public contracts with proportion of bad contracts to good equaled to 1:3. In initial dataset several rows may correspond to one contract. Dataset contains 38 characteristics describing supplier, customer (public organization) and contract (Table 1, Table 2, Table 3). Target variable is described in Table 4.

Table 1. Variables describing supplier

Supplier			
Variable		Type of values[1]	Type of variable[2]
Description	Variable name in code		
1. Number of finished contracts	sup_cntr_num	INT	Q
2. Number of running contracts for current moment	sup_running_cntr_num	INT	Q
3. Number of good finished contracts	sup_good_cntr_num	INT	Q
4. Number of federal contracts	sup_fed_cntr_num	INT	Q
5. Number of regional contracts	sup_sub_cntr_num	INT	Q
6. Number of municipal contracts	sup_mun_cntr_num	INT	Q
7. Average price of contract	sup_cntr_avg_price	FLOAT	Q
8. Average share of penalties from contract price	sup_cntr_avg_penalty_share	FLOAT	Q01
9. Share of contract with one-sided termination by supplier willing	sup_1s_sev	FLOAT	Q01
10. Share of contract with one-sided termination by customer willing	sup_1s_org_sev	FLOAT	Q01
11. Share of contracts without penalties	sup_no_pnl_share	FLOAT	Q01

[1] INT — integer value, FLOAT — floating point number, BOOL — 1 or 0, STRING —

[2] Q — quantitative variable, Q01 — quantitative variable with values in [0, 1], C — categorical variable переменная, CB — categorical binary variable, O — ordinal variable

12. Number of good contracts for OKPD of current contract	sup_okpd_cntr_num	FLOAT	Q
13. Share of contracts with price which differs on less than 20% comparing to price of current contract	sup_sim_price_share	FLOAT	Q01
14. Federal subject where supplier is registered	sup_ter	INT	C

Table 2. Variables describing customer

Customer			
Variable		Type of values	Type of variable
Description	Variable name in code		
15. Number of finished contracts	org_cntr_num	INT	Q
16. Number of running contracts for current moment	org_running_cntr_num	INT	Q
17. Number of good finished contracts	org_good_cntr_num	INT	Q
18. Number of federal contracts	org_fed_cntr_num	INT	Q
19. Number of regional contracts	org_sub_cntr_num	INT	Q
20. Number of municipal contracts	org_mun_cntr_num	INT	Q
21. Average contract price	org_cntr_avg_price	FLOAT	Q
22. Number of finished contracts with current supplier	cntr_num_together	INT	Q
23. Share of contract with one-sided termination by customer willing	org_1s_sev	FLOAT	Q01

24. Share of contract with one-sided termination by supplier willing	org_1s_sup_sev	FLOAT	Q01
25. Share of contracts with price which differs on less than 20% comparing to price of current contract	org_sim_price_share	FLOAT	Q01
26. Type of organization	org_type	INT	C
27. Federal subject where customer is registered	org_ter	INT	C

Table 3. Variables describing contract

Contract			
Variable		Type of values	Type of variables
Description	Variable name in code		
28. Price of contract	price	INT	Q
29. Initial maximum contract price	pmp	INT	Q
30. OKPD (all-Russian classifier of products by type of economic activity)	okpd	INT	C
31. Number of contracts for given OKPD	okpd_cntr_num	INT	Q
32. Number of good contracts for given OKP	okpd_good_cntr_num	INT	Q
33. Level of contract (federal, regional, municipal)	cntr_lvl	INT	C
34. Date when contract was signed	sign_date	STRING	O
35. Date when contract is planned to be finished	exec_date	STRING	O
36. Type of purchase	purch_type	INT	C

37. True if price <= 0,6 * pmp, else False	price_too_low	BOOL	CB
38. True if price > pmp, else False	price_higher_pmp	BOOL	CB

Table 4. Target variable

Target variable			
Description	Variable name in code	Type of values	Type of variables
False if contract is finished, else True (contract was preliminary terminated)	cntr_result	INT	CB

Data analysis and preprocessing

Different types of analysis were conducted to explore data. For **quantitative variable** descriptive statistics were calculated, outliers were checked (Fig. 4), distributions were visualized, correlation analysis and ANOVA test were conducted to test variables impact on target. Most of the charts is left in code in repository (Eliseev, 2019). Please, proceed to GitHub to see more details.

All quantitative variables (Q) have outliers, almost every variable has positively skewed distribution. Before preprocessing all variables had correlation lower than 0,13 with target variable. Top three variable with highest correlation were plan_cntr_len (0,13), okpd_cntr_num (-0,11), sup_running_str_num (0,09). To deal with skewness logarithmic transformation was applied. After that top three variables with highest correlation with target variables were org_cntr_num (0,18), sup_running_cntr_num (0,18), plan_cntr_len (0,15).

To avoid multicollinearity several variables were excluded from sample. Outliers were updated by boundary values (1st and 99th percentile) and the fact of being an outlier was saved to new binary categorical variable. All quantitative variables were scaled and centered, so that mean was equal to zero, and standard deviation – to one.

As for quantitative variables which presented shares (Q01), there was no need for a lot of preprocessing, only some variables were deleted because they were constants (some of them were sup_1s_sev, sup_1s_org_sev, org_1s_sev, org_1s_sup_sev with value equal to zero). This could be expected because one-sided termination is not profitable for both customer and supplier. The last one after one-sided termination will be likely included in listed in the Register of unscrupulous suppliers and will not be able to take part in public procurement for two years.

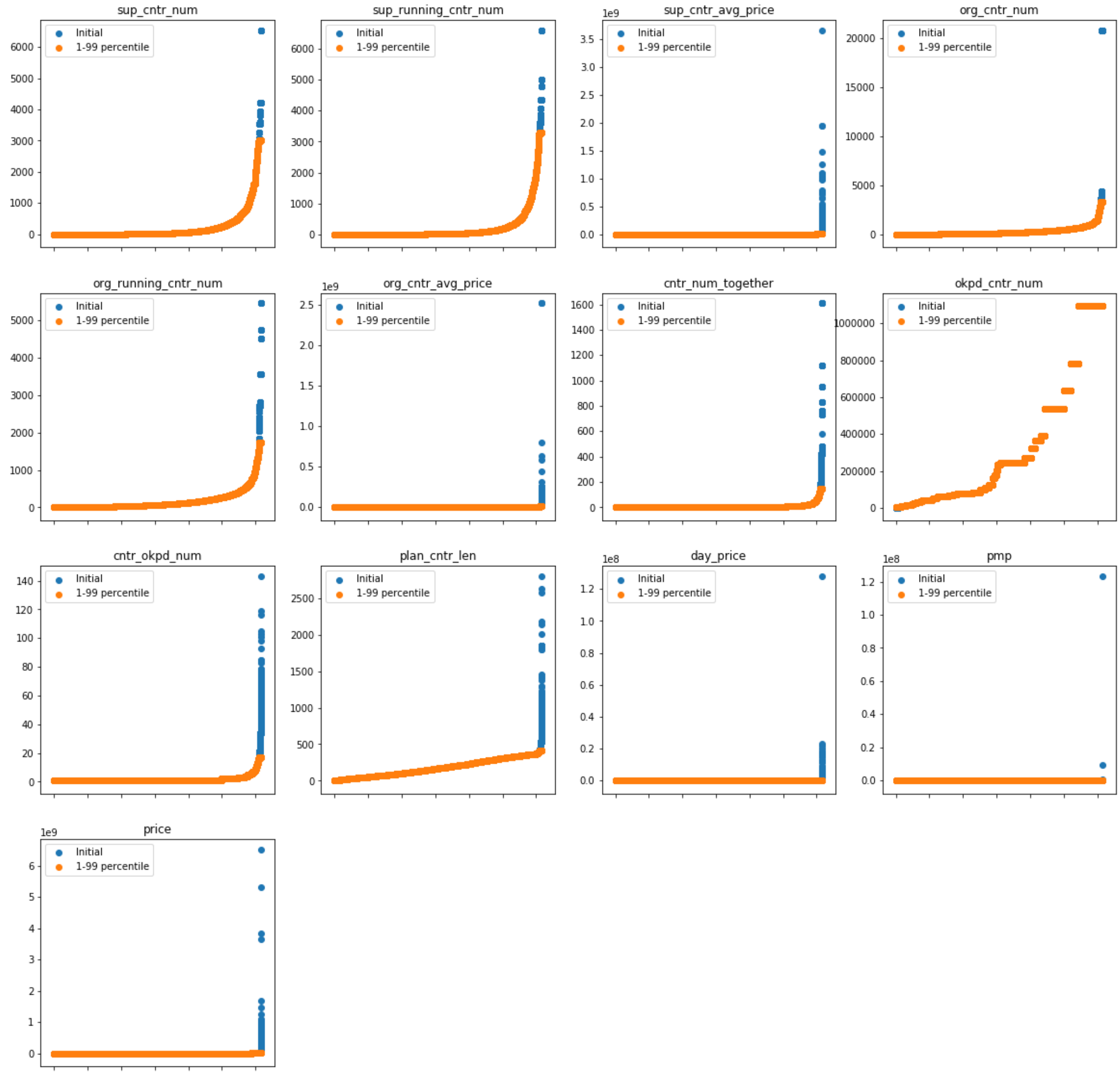


Figure 4. Analysis on outlier for quantitative variables (Q)

For **categorical variables** frequency distributions were visualized and feature importance was tested with χ^2 test and Information Value (Howard, 1966). Before going forward, it is important to describe several extra transformations needed before classical pipeline of preprocessing categorical variables.

As it was mentioned, in initial dataset several rows may describe one contract and reason for this is requirement for different type of products and services in one contract. Type of product and services in Russia is encoded basing on all-Russian classifier of products and 9-digit code (OKPD, in dataset this classification is referred in variable okpd). Okpd variable and its derivatives (`okpd`, okpd_cntr_num, okpd_good_cntr_num, sup_okpd_cntr_num) present industrial patterns and might be important variables.

`Okpd` variable has more than seven thousand unique values. To decrease number of unique values, okpd was shortened to first two symbols, representing class of product / services such as “education services”, “real estate services”, “electricity, gas, steam and air conditioning” and so on. Then data was reaggregated in accordance with shortened OKPD. After this transformation only 81 unique values were left.

To describe in one row contract with several classes of product, dummy variables were created for `okpd` variable, which were equal to one if such OKPD is presented in contract and zero otherwise. After this transformation one observation described one contract. Similar workaround was implemented for derivative of `okpd` – `sup_okpd_cntr_num`. These transformations added 162 new binary variables: okpd2_0, okpd2_10, …, okpd2_99 and socs_0, socs_10, …, socs_99.

As for `okpd_good_cntr_num` and `okpd_cntr_num`, these variables were transformed in share of good contract among all contracts with given OKPD and stored in three columns as minimum, average and maximum from several shares corresponding to several OKPDs.

Let’s turn back to analysis of categorical variables. In order to calculate IV (formula 6), it is necessary to calculate weight of evidence at first (WoE, formula 5).

$$WoE_{ij} = \log\left(\frac{p_{ij}}{q_{ij}}\right), \qquad (5)$$

- i — categorical variable;
- j — unique value i-variable;
- p_{ij} — how many times value i for variable j is met among good contracts divided by total number of good contracts;
- q_{ij} — how many times value i for variable j is met among bad contracts divided by total number of bad contracts.

$$IV_i = \sum_j \left(p_{ij} - q_{ij}\right) * WoE_{ij}, \qquad (6)$$

- i, j, p_{ij} and q_{ij} are the same as in Formula 5.

IV is impossible to calculate if value is met only among good or only among bad contracts. So, rare values which are noticed in sample in less than 0,5% of cases were grouped in one new value.

In table 5 IV for categorical variables are given. Month of signing the contract seems to have strong discriminative power.

Table 5. IV for categorical variables (C)

Variable	Information value	Discriminative power
Sup_ter	0,07	Useless
Org_type	0,03	Useless
Org_ter	0,17	Medium
Sign_month	0,401	Strong
Purch_type	0,19	Medium
Cntr_lvl	0,15	Medium

If we visualize distribution of good and bad contracts in sample by month, the tendency to have more inefficient contracts in the beginning of year and less in the end is evident (Fig. 5),

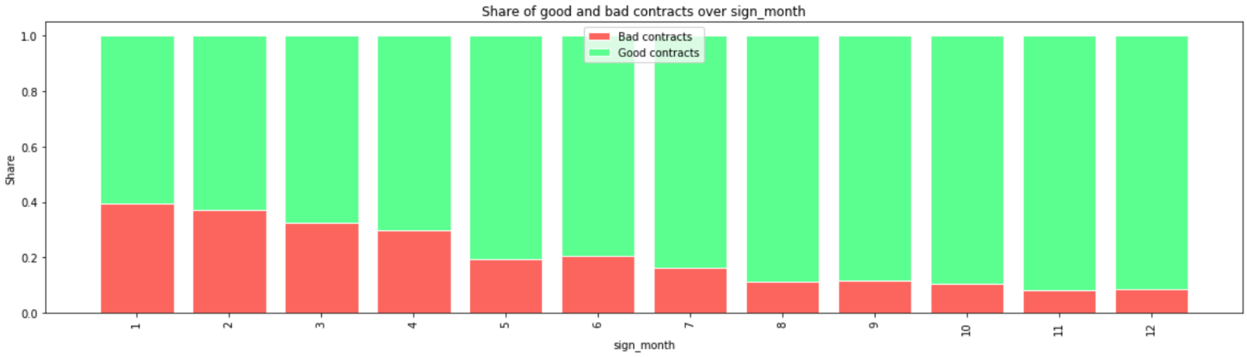


Figure 5. Distribution of good and bad contracts by month of signing

Machine learning algorithms do not work with categorical variables directly, they should be encoded. There are several methods – dummy encoding, encoding by frequencies, encoding by share (frequencies divided by total number of observations) etc. One of the most promising for binary classification is WoE-encoding, when value is replaced by its weight of evidence. This method usually demonstrates better result comparing to other ways of encoding for binary classification task.

To conclude, after analysis and preprocessing some variables were deleted, namely № 9-10, 14, 18-20, 23-24, 37-28, and some were created (Table 6).

Table 6. New variables after data preprocessing

Contract			
Variable		Type of values	Type of variables
Description	Variable name in code		
Planned length of contract (exec_date – sign_date)	Plan_cntr_len	INT	Q
Price of contract per day (price / plan_cntr_len)	Day_price	INT	Q
Month of signing	Sign_month	INT	C
First two symbols (10-99) of all-Russian classifier of products by type of economic activity, okpd2_0 – for undefined class of product / service	Okpd2_0, okpd2_10, …okpd2_99	INT	CB
Supplier			
Share of good contracts among total	Sup_good_cntr_share	FLOAT	Q01
Share of federal contracts among total	Sup_fed_cntr_share	FLOAT	Q01
Share of regional contracts among total	Sup_sub_cntr_share	FLOAT	Q01
Share of municipal contracts among total	Sub_mun_cntr_share	FLOAT	Q01
Share of contract with given OKPD (10-99) among total, socs2_0 – for undefined class of product / service	Socs_0, socs_10, …, socs_99	INT	CB

In addition, it was created 19 binary variables (CB) to reflect outliers which we noticed almost in every quantitative variable.

Model development

Algorithms

In this study it is proposed to use four algorithms for developing a machine learning model for detection of risky contracts: logistic regression, random forest, gradient boosting on decision trees and neural network. Below the ideas standing behind algorithms, their strengths and weaknesses will be briefly described (Muller & Guido, 2017).

The risk prediction is a binary classification task, since the target variable takes two values (0 for ordinary observation, 1 for deviation). For classification task it is recommended to begin solving the problem with **logistic regression**, which is simple parametric linear model and, usually, serves as a baseline for more profound algorithms (formula 7).

$$p = \frac{1}{1 + e^{-(w_0 + w_1 x_2 + \cdots + w_n x_n)}} \tag{7}$$

Key strengths: fast fit and fast prediction, ability to work with sparse data and high-dimensional data, high interpretability. *Key weaknesses*: scaling and normalization is needed, not robust to outliers, linear separating hyperplane (can be made “non-linear” with the help of adding new features based of non-linear transformations of existing characteristics).

The next algorithm proposed to use is a **random forest**. The random forest is a homogeneous ensemble, bagging from a variety of decision trees. The key idea is to build many weak independent learners (shallow trees) on randomly chosen subsamples. Output of the model is average of predictions of trees. Homogeneous ensembles are often used in practice and in machine learning competitions. For random forest it is important to control depth of trees and number of trees to avoid overfitting. *Key strengths*: no need for data scaling and normalization, ability to parallelize training of random forest, ability to work with huge volume of data, robust to outliers. *Key weaknesses:* not interpretable model (not taking into consideration feature importance), performance on sparse data and high-dimensional data may be bad.

The third algorithm tested is the **gradient boosting on the decision trees**. This algorithm, like random forest, is a homogeneous ensemble consisting of decision trees, but differs in the way of building them. For a random forest, trees are built on different subsamples of the total sample (bagging). Boosting in the general case is a method of constructing an ensemble of models, in which each subsequent model is built to correct the errors of the previous model (ensemble of models). So, this algorithm is based on idea of adding new trees to ensemble which will rectify errors of previous trees. In machine learning competitions gradient boosting on decision trees are often found in the best solutions. Also, this algorithm is often used in practice in industrial solutions. Some of popular implementations are XGBoost, LightGBM and CatBoost. *Key strengths*: the same as random forest has, higher performance and prediction building comparing to random forest. *Key weaknesses:* the same as random forest has, slower training, high sensitivity to parameters tuning.

Finally, neural network will be trained to compare its performance with three before mentioned algorithms. *Key strengths*: ability to detect very difficult patterns in data, ability to work with instructed data (images, text, sound etc.), support of incremental and transfer learning. *Key weaknesses*: need to scale and normalize data, not interpretable model, demand a lot of computational resources, sensitive to parameters tuning. Some of points mentioned above are summarized in Table 7.

Table 7. Main feature of used algorithms

Algorithms	**Type of hyperplane**	**Interpretability**	**Robust to outliers**	**Demand of computational resources**	**Need to scale and normalize data**	**Efficiency with high-dimensional data**
Logistic regression	Linear	Yes	No	Low	Yes	Yes
Random forest / Gradient boosting	Nonlinear	No	Yes	Medium	No	No
Neural network	Nonlinear	No	No	High	Yes	Yes

As for practical realization, for logistic regression and random forest *sklearn* package for Python is used, for gradient boosting on decision trees – *XGBoost* library. *Pytorch* is chosen as a framework for neural networks.

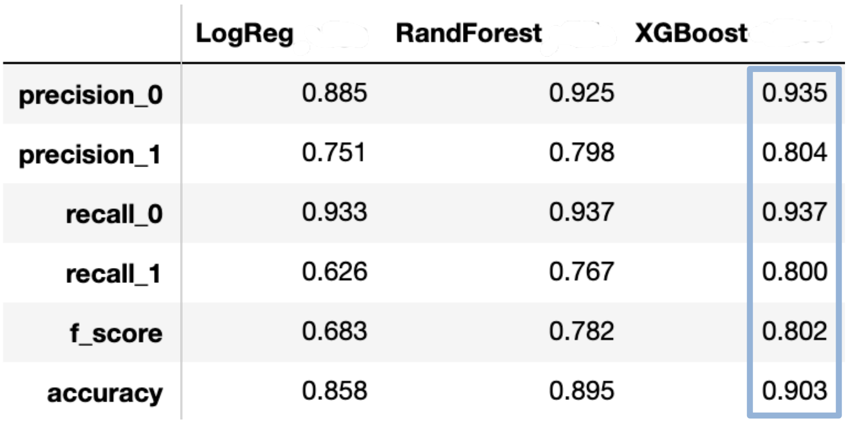
Training, testing, validation

Sample was divided in two groups 80% for training and 20% for validation with saving proportions of bad and good contracts inside subsamples. The first part of sample was used for training model and tuning parameters on cross-validation. Taking available computational resources into account, it was decided to use 2-fold cross-validation for hyperparameters tuning. The second subsample was used for defining final quality of models.

Hyperparameters tuning was conducted with the help of Bayesian Optimization methods realized in Hyperopt (Distributed Asynchronous Hyper-parameter Optimization) library ("Hyperopt: Distributed Asynchronous Hyper-parameter Optimization," 2019).

RESULTS

After tuning of parameters of logistic regression, random forest and gradient boosting the best models were chosen and main metrics were calculated on validation sample (Fig. 6).

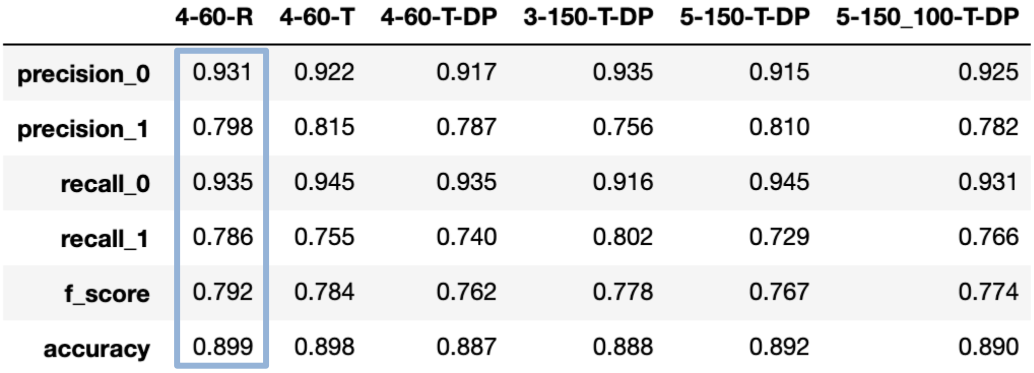


	LogReg	RandForest	XGBoost
precision_0	0.885	0.925	0.935
precision_1	0.751	0.798	0.804
recall_0	0.933	0.937	0.937
recall_1	0.626	0.767	0.800
f_score	0.683	0.782	0.802
accuracy	0.858	0.895	0.903

Figure 6. Main metrics on validation sample for logistic regression, random forest and gradient boosting

The best quality demonstrate model based on gradient with specificity = 0,94 and sensitivity = 0,8, ROC AUC = 0,96. The difference between precision and recall on good class (0) differ on about 13 percentage points from the same metrics on bad class (1).

There were six types of different neural networks trained starting from four hidden fully connected layers with 60 neurons in each layer and finishing with network with five hidden layer and 150 neurons in each of them. The activations functions varied between models from ReLU to Tanh, sigmoid activation function showed much worse results. For regularization there was used dropout after each hidden layer for last four neural networks. Column names represent number of hidden layers, number of neurons in each hidden layer, activation function and usage of dropout as regularization technique.



	4-60-R	4-60-T	4-60-T-DP	3-150-T-DP	5-150-T-DP	5-150_100-T-DP
precision_0	0.931	0.922	0.917	0.935	0.915	0.925
precision_1	0.798	0.815	0.787	0.756	0.810	0.782
recall_0	0.935	0.945	0.935	0.916	0.945	0.931
recall_1	0.786	0.755	0.740	0.802	0.729	0.766
f_score	0.792	0.784	0.762	0.778	0.767	0.774
accuracy	0.899	0.898	0.887	0.888	0.892	0.890

The highest quality achieved the simplest model with specificity = 0.94, sensitivity = 0.9, ROC AUC = 0.96. The values of metrics of neural network are almost identical to result of gradient boosting and are slightly better than metrics of models based on logistic regression and random forest.

Let us compare best model in more detail with paying attention to probabilistic power of classifiers, because, in the end, model should give probabilities which provide more information to decision maker than 0 or 1 (Fig. 7, 8).

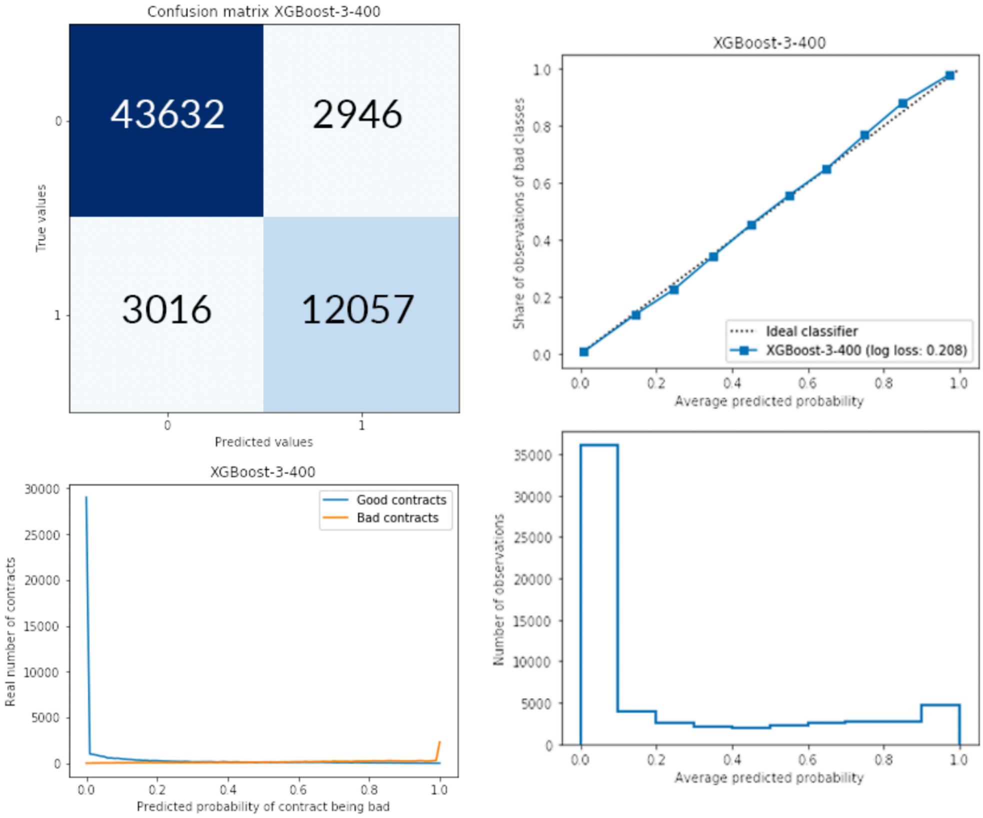


Figure 7. Confusion matrix and probabilistic power of classifier (XGBoost)

Confusion matrix are similar for both models but model based on XGBoost demonstrates slightly better results. Both classifiers have f-score approx. 0.8 and almost identical strong probabilistic prediction power (LL = 0.21 vs LL = 0.23). Thus, the result achieved by developed neural network are almost identical to model based on gradient boosting on decision trees.

The models are perfectly calibrated. This means, that their output of the `predict_proba` method (or just output) can be directly interpreted as a confidence level ("Comparison of Calibration of Classifiers," n.d.). In other words, a well calibrated (binary) classifier will classify the observation such that among them to which it gave a probabilistic prediction close to 0.8, about 80% actually will belong to the positive class.

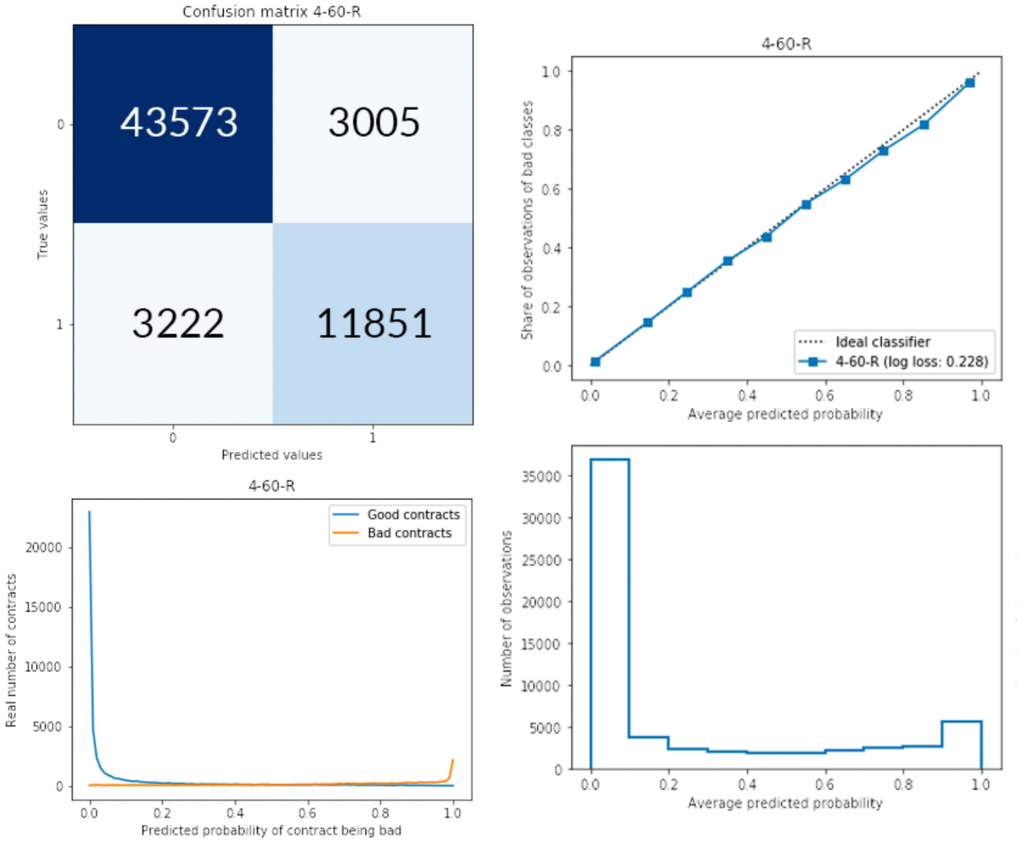
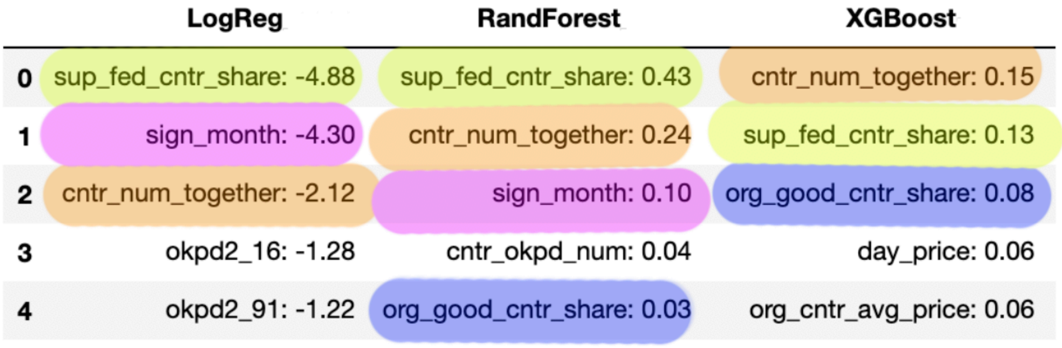


Figure 8. Confusion matrix and probabilistic power of classifier (neural network)

One more thing is left to stay. Classical algorithms natively support assessment of importance of features for prediction. On Fig. 9 top five important features for models are listed.



	LogReg	RandForest	XGBoost
0	sup_fed_cntr_share: -4.88	sup_fed_cntr_share: 0.43	cntr_num_together: 0.15
1	sign_month: -4.30	cntr_num_together: 0.24	sup_fed_cntr_share: 0.13
2	cntr_num_together: -2.12	sign_month: 0.10	org_good_cntr_share: 0.08
3	okpd2_16: -1.28	cntr_okpd_num: 0.04	day_price: 0.06
4	okpd2_91: -1.22	org_good_cntr_share: 0.03	org_cntr_avg_price: 0.06

Figure 9. Top 5 important features

Number of finished together contracts (cntr_num_together) is in all lists. This is reasonable, the longer contractors know each other, the less probability that something will go wrong. Share of finished contracts on federal level by supplier (sup_fed_cntr_share) is also important variable for every model. The federal contracts are controlled especially thoroughly and to win tender on federal contract is not simple. Likely, suppliers who are able to execute federal contracts are not new on the market and value their reputation. There are two variables that are met in 2 out of 3 lists – month of signing contract (sign_moth) and proportion of good contract of customer in total (org_good_cntr_share). As it was shown in “Data analysis and preprocessing“ section, share of bad contracts gradually decline during the year, the most of risky contracts are signed in January and the least number of bad contracts is witnessed in December. And `org_good_cntr_share` reflects experience of public customer in procurement process because, apparently, not only suppliers violate rules.

`Cntr_okpd_num` shows how many contracts were conducted with given OKPD, how this class of products or services is popular on market. If it is not popular, then there are less suppliers and market may be less competitive, and, consequently, more corruptive. `One_day_price` may help detect contracts with unusual prices, when contract implies long run duties, but price is low and vice versa. `Org_cntr_avg_price` represents the average contract price customer has experience to work with. If public agent conducts a contract with a price significantly higher than average price of contracts in past, this may be a signal that something may go wrong.

DISCUSSION

Implication for theory and practice

This research shows the opportunity of using machine learning in procurement process, while a few studies are covering this theme now. The ideas and built models can be used to facilitate starting pilot project in practice and research in this field.

Limitations

As it is usual for machine learning techniques, limitations of application of built models are rooted in sample used for training. In this study it was used data for contracts governed by 44-FL. Meanwhile, there are contracts which are conducted with accordance with 223-FL (federal law about “Procurement of goods, works, services by certain types of legal entities”). 44-FL regulates all purchases of all government customers and fully regulates the conduct of the trade procedure. 223-FL is significantly less strict and regulates general procurement principles only. Public organization, depending on its type, can conduct purchases either following 44-FL or 223-FL. It was decided to start study from more regulated public contracts because it is easier to get trustworthy results on small datasets. It is very likely that contracts characteristics regulated by 44-FL and 223-FL differ a lot and model developed in this study is only applicable for first type of contracts.

Further research

Despite the achievement of a good result in developing a model for the risk prediction of public contracts, a large room for further research remains. One of the directions of study might be testing alternative methods of dealing with imbalanced data – such as oversampling (SMOTE) or changing cost function. Another area is enrichment of dataset with variables from other sources of information except UPPIS. This may be data from the Federal Tax Service, Unified Federal Register of Bankruptcy Information, Federal Bailiff Service, Supreme Court of the Russian Federation and other sources of information. Finally, analysis of text data attached to contract (commercial offer) may be included.

CONCLUSION

Public authorities in Russia spend trillions of rubles annually in procurement process. Taking this into account, everyone understands that transparency is crucial to facilitate effective budget usage and a lot of is done towards this direction. Web-based platform for collection data about contracts – the Unified Public Procurement Information System (UPPIS) was launched. The legislation is constantly evolving to meet needs of contractors better. And by this moment, level of technology, price of computational resources and storage, volume of collected data presents an opportunity to leverage data analytics and machine learning techniques in order to boost transparency and predictability in procurement process.

One of example of machine learning application in public purchases is risk prediction of contracts. In this work possibilities of application machine learning and, in particular, deep leaning techniques for risk assessment of public contracts were studied. Objective of work was to develop machine learning model that are able to predict riskiness of contracts regulated by 44-FL basing on data stored in UPPIS.

To achieve this goal papers in areas of application machine learning for risk prediction were analyzed and challenges of usage machine learning techniques in procurement process were described. Moreover, sample with 38 characteristics and more than 300 observations of real contracts was collected. To develop the model four machine algorithms (logistic regression, random forest, gradient boosting on decision trees, neural network) were used.

As a result, two similar model were developed basing on gradient boosting n decision trees and neural network. They demonstrate strong probabilistic power and similar values of metrics: specificity = ~0.94, sensitivity = ~0.8 and AUC ROC exceeding 0.95. These results were achieved without aggregation data from any external sources except UPPIS. Hence, in general sense, there is a lot of room for achievement higher quality of models with the help of information from other sources.

Arguably suggested application of machine learning in procurement process might optimize controlling and monitoring, which is conducted manually at this moment. Developed model is very scalable and can access millions of contracts per day, giving to them probability of being preliminary terminated or, in other word, “risk score”. Depending on needs and resources of controlling bodies, models can be tuned to be more aggressive or more conservative to maximize value from solution.

LITERATURE

Baesens, B., Lessmann, S., Seow, H., & Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: a ten-year update. *European Journal of Operational Research*, *54*(6), 627–635.

Comparison of Calibration of Classifiers [Scikit Learn]. (n.d.). Retrieved May 25, 2019, from https://scikit-learn.org/stable/auto_examples/calibration/plot_compare_calibration.html

Consolidated analytical report on the results of the monitoring of purchases, goods and services for the provision of state and municipal. (2019). Retrieved May 25, 2019, from Ministry of Finance of the Russian Federation website: https://www.minfin.ru/common/upload/library/2019/04/main/Svodnyy_analiticheskiy_otchet_2018_itog.docx

Eliseev, D. (2018a). Government procurement data for Yaroslavskaya and Tulskaya oblast. Retrieved May 25, 2019, from Zenodo website: https://zenodo.org/record/1244260

Eliseev, D. (2018b). Risk Prediction of Contracts in Government Procurement Process Using Machine Learning Algorithms. Retrieved May 25, 2019, from GitHub website: https://github.com/DmitryEliseev/diploma-project-2018

Eliseev, D. (2019). Repository for research in field of Russian government procurement. Retrieved May 25, 2019, from GitHub website: https://github.com/DmitryEliseev/government-procurement-analysis

Eliseev, D., & Romanov, D. (2018). Machine learning: predicting the risk of public procurement. *Open Systems. DBMS*, (2), 42–44.

Gallego, J., Rivero, G., & Martinez, J. D. (2018). Preventing rather than Punishing: An Early Warning Model of Malfeasance in Public Procurement. *Serie Documentos de Trabajo*, (222), 1–33.

Howard, R. (1966). *Information value theory*. 22–26.

Hyperopt: Distributed Asynchronous Hyper-parameter Optimization. (2019). Retrieved May 25, 2019, from https://github.com/hyperopt/hyperopt

Kourou, K., Exarchos, T., Exarchos, K., Karamouzis, M., & Fotiadis, D. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17.

Kruppa, J., Ziegler, A., & Konig, I. (2012). Risk estimation and risk prediction using machine-learning methods. *Human Genetics*, *131*(10), 1639–1654.

Muller, A., & Guido, S. (2017). *Introduction to machine learning with Python*. O'Reilly Media.

Phaisangittisagul, E. (2016). *An Analysis of the Regularization between L2 and Dropout in Single Hidden Layer Neural Network*. 174–179. Retrieved from http://uksim.info/isms2016/CD/data/0665a174.pdf

Saxena, S. (2017, October 26). Artificial Neuron Networks (Basics) | Introduction to Neural Networks. Retrieved May 25, 2019, from Becoming Human website: https://becominghuman.ai/artificial-neuron-networks-basics-introduction-to-neural-networks-3082f1dcca8c

Settemesdal, S. (2019). Machine Learning and Artificial Intelligence as a Complement to Condition Monitoring in a Predictive Maintenance Setting. *SPE Oil and Gas India Conference and Exhibition*, 1–20. Dubai: Society of Petroleum Engineers.

Unified Public Procurement Information System. (2019). Retrieved May 20, 2019, from http://zakupki.gov.ru/

Vázquez, F. (2017, December 21). Deep Learning made easy with Deep Cognition. Retrieved May 25, 2019, from Becoming Human website: https://becominghuman.ai/deep-learning-made-easy-with-deep-cognition-403fbe445351