N. V. Kuznietsova, Cand. Sc. (Eng.), Ass. Prof.

INFORMATION TECHNOLOGIES FOR ANALYZING FINANCIAL ABUSES AT PROZORRO PLATFORM

The paper shows the possibilities of analyzing on-line purchases at ProZorro platform by the data mining methods with the aim to identify the behavior of companies and characteristics by which it is possible to reveal a collusion and unlawful activities during their participation in on-line trading.

Keywords: *logistic regression, neural networks, tender purchases, ProZorro, information technologies.*

Introduction

With the aim to provide effective and transparent purchasing, to create competitive environment in the sphere of public procurement, to prevent corruption in this sphere, to develop fair competition, Verkhovna Rada of Ukraine adopted the Law "On Public Procurement" [1]. This Law adoption and development itself was a significant step in the fight against non-transparent agreements and collusions, which were the reason for squandering the budget finances and overvalued cost of the works. This Law envisages that all services and goods for the sum exceeding 50 thousand UAH are to be carried out using electronic procurement system for selecting a supplier of goods, a provider of services and a performer of works for conclusion of the contract, the customers should adhere to the principles of public procurement implementation stated in the Law. For this purpose "ProZorro" system was created, which is an electronic database for implementation of public state procurement in on-line mode [2]. The system itself provides the portal users with the ability of real-time access to all purchases conducted and thus verifying the procurement transparency, equal access of all the market participants as well as checking directly how the taxes of Ukrainian citizens are spent for purchasing services of the state sector.

The paper aims at analyzing and verifying transparency of tender procurement using data mining methods, implemented in the form of information technologies at SAS platform, in order to reveal possible violations and abuses. Based on the results of the analysis, it is possible to elaborate recommendations on further improvement of the online procurement monitoring system and to reduce financial losses of the taxpayers caused by non-transparency of trades and inaccessibility of real competitive market participants to them.

Monitoring of e-procurement

In order to verify fairness of the conducted trades and purchases, the Law of Ukraine [1] envisages the possibility to declare available abuses through mass media or through public associations and to provide official information from state authorities or from local self-administration bodies in case of violations, identified by the financial control body, or in case of automatically detected risk indicators in the conducted purchases.

The paper will show the possibility to improve automatic risk indicators as special criteria with pre-defined parameters, which enables automatic detection of the signs of legal offences and abuses in procurement procedure.

For the moment, 22 trading platforms are authorized in ProZorro base: Zakupki.Prom.ua, e-tender, Newtend, SmartTender, «Держзакупівлі онлайн», PublicBid and «ПриватМаркет», etc. Their list is constantly updated on the site [2]. The only way to connect to ProZorro system is through one of the platforms.

To detect the signs of legislation violations in the sphere of public procurement, the following data could be used: information disclosed in the electronic procurement system, information contained in the unified state registers, information in databases open to the central executive authority, which implements the state policy in the field of public financial control, data of the state

authority bodies and the bodies of local self-government, enterprises, institutions, organizations, customers and participants of procurement procedures, which can be obtained by the bodies of state financial control in the manner prescribed by the Law.

Analysis of e-procurement in order to establish certain regularities and elaboration of recommendations

To improve the monitoring procedure and the system of automatic detection and rejection of unfair companies, the system implements a certain set of classification algorithms. There is no open access to the settings and parameters, by which classification is carried out, in order to prevent possible manipulations and provision of false statistical data from companies applying for participation in the tender. Therefore, given simulation aims at identifying cause-effect relations and statistical characteristics to ensure qualitative classification as well as at obtaining information for prediction of suspicious purchases required for official application in the order, prescribed by the Law [1], for performing inspection.

An assumption was made about existence of a certain relationship between duration of a company participation in the deal and its characteristics, namely, whether duration of the company's participation in the deal depends on suspicion in its possible unlawful activity at the platform (i.e. in collusions with other companies). The following data sample was empirically formed:

1. Wins – number of wins of a certain company;
2. Losses – number of the company losses in deals;
3. Sum_of_deals – total sum of deals won;
4. Participations – number of participations in deals;
5. Objections – number of objections submitted by the given company;
6. Date_start – date of starting participation in the system of deals;
7. Date_finish – date of the last participation in deals;
8. IdTenderer – unique number of the tender participant;
9. Suspected – variable, which shows if a company is suspected in illegal collusions with other participants;
10. Churn out – target variable, which is equal to 1 if a company stopped its participation in the deals in a short time period (it is assumed that a company suddenly stopped participation or it was a fictitious company for one deal only).

The company was considered the one, which continues trading, if the period between starting trading at the platform and the time of the last deal was more than 60 days (statistic average duration of business cycles of the given companies at the platform according to official data of ProZorro [2]). It is also worth noting that companies, which participated in the competition at the platform less than three times, were filtered out from the sample.

Input sample contained 3966 cases: 1757 companies, which stopped their participation in a very short time period, and 2209 companies, which continued to participate in the tenders.

The sample was divided into training and verification samples using the stratification method in respect to target variable with the ratio 70 / 30.

Since task statement involved solving the classification problem, it was proposed to use such methods of intellectual data analysis as neural networks, decision trees, logistic regression and Bayesian classifier [3 – 5]. In order to choose the best model, the following criteria could be used [6]:

1. MSE – mean squared error:

$$MSE = E\left((y - \hat{y})^2\right) = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}, \quad (1)$$

where $\hat{y}$ – the dependent variable values, estimated using the constructed mathematical model; y – actual values of the dependent variable;

2) sum of the squared errors (SSE):

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2 \quad (2);$$

3) Akaike's information criterion:

$$AIC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n \quad (3);$$

4) Bayesian – Schwarz's criterion

$$BSC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + n \ln(N) \quad (4),$$

where $n = p + q + 1$ – number of the model parameters, estimated using statistical data (p – number of parameters of the auto-regression part of the model; q – number of the moving-average parameters; „1" appears when the offset is estimated (or intersection, i. e. a_0), N – the sample length;

5) Misclassification Rate is calculated as the ratio between the number of mistakenly predicted parameters and total number of N values:

$$\text{Misclassification Rate} = \frac{number\ of\ the\ mistakenly\ predicted\ parameters}{N} \quad (5).$$

Construction of the described models and computation of the statistical criteria was performed based on the Enterprise Miner information technology [7]. Selection of the best model could be conducted automatically or based on the pre-defined quality criterion. Sequence of the tender procurement analysis is presented in Fig. 1.

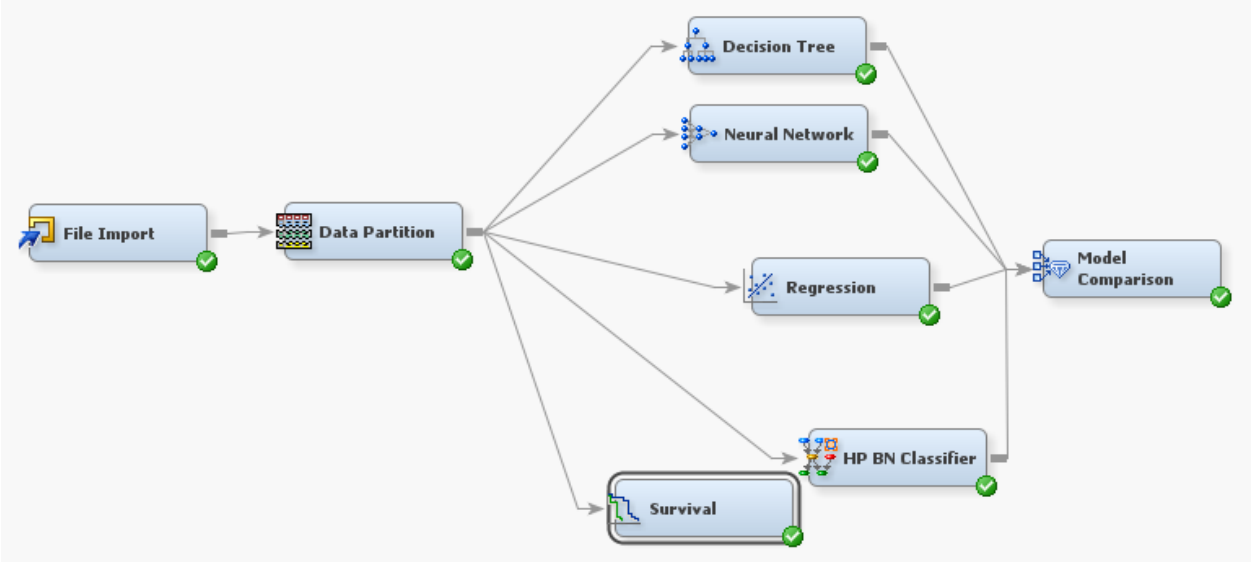

Fig. 1. Sequence of the tender procurement analysis based on the SAS Enterprise Miner information technology

Neural network

We built different types of neural networks [5] with different number of layers, different activation functions, etc. Akaike's criterion was chosen as quality criterion (in order not to make the model too

complex and to achieve balance between the model parameters and its quality). A simple perceptron neural network with 20 hidden layers and standardization of inputs, based on deviation, radial combinative function, logistic activation function and misclassification error training criterion turned out to be the best model for input data. Statistical quality criteria of the best neural network are given in Table 1.

Table 1

Statistical characteristics of the best neural network

Target	Fit Statistics	Statistics Label	Train	Validation
churn_out	_DFT_	Total Degrees of Freedom	2774	-
churn_out	_DFE_	Degrees of Freedom for Error	2743	-
churn_out	_DFM_	Model Degrees of Freedom	31	-
churn_out	_NW_	Number of Estimated Weights	31	-
churn_out	_AIC_	Akaike's Information Criterion	3227.794	-
churn_out	_SBC_	Schwarz's Bayesian Criterion	3411.564	-
churn_out	_ASE_	Average Squared Error	0.197105	0.193835
churn_out	_MAX_	Maximum Absolute Error	0.986715	0.984846
churn_out	_DIV_	Divisor for ASE	5548	2384
churn_out	_NOBS_	Sum of Frequencies	2774	1192
churn_out	_RASE_	Root Average Squared Error	0.443965	0.440267
churn_out	_SSE_	Sum of Squared Errors	1093.537	462.1033
churn_out	_SUMW_	Sum of Case Weights Times Freq	5548	2384
churn_out	_FPE_	Final Prediction Error	0.20156	-
churn_out	_MSE_	Mean Squared Error	0.199332	0.193835
churn_out	_RFPE_	Root Final Prediction Error	0.448954	-
churn_out	_RMSE_	Root Mean Squared Error	0.446467	0.440267
churn_out	_AVERR_	Average Error Function	0.570619	0.567532
churn_out	_ERR_	Error Function	3165.794	1352.996
churn_out	_MISC_	Misclassification Rate	0.317231	0.305369
churn_out	_WRONG_	Number of Wrong Classifications	880	364

Regression model

Since classification problem was solved, as the next model suitable to analyze if a company would cease participation in subsequent tenders (binary output: 0 – “no” or 1 – “ regression cease ”) logistic was chosen, based on the stepwise model construction method with pairwise inputting and outputting of the characteristics from the model. For this model the following characteristics were obtained: $AIC = 3301{,}941$ and $MisclassificationRate = 0{,}328767$.

Decision trees

Then simulation was carried out, based on the decision tree method [3], with different settings of the cut-off rules, the number of descendants and the sub-tree formation algorithms. The tree with minimum misclassification rate turned out to be the best one: $MisclassificationRate = 0{,}313758$. The tree structure is shown in Fig. 2.

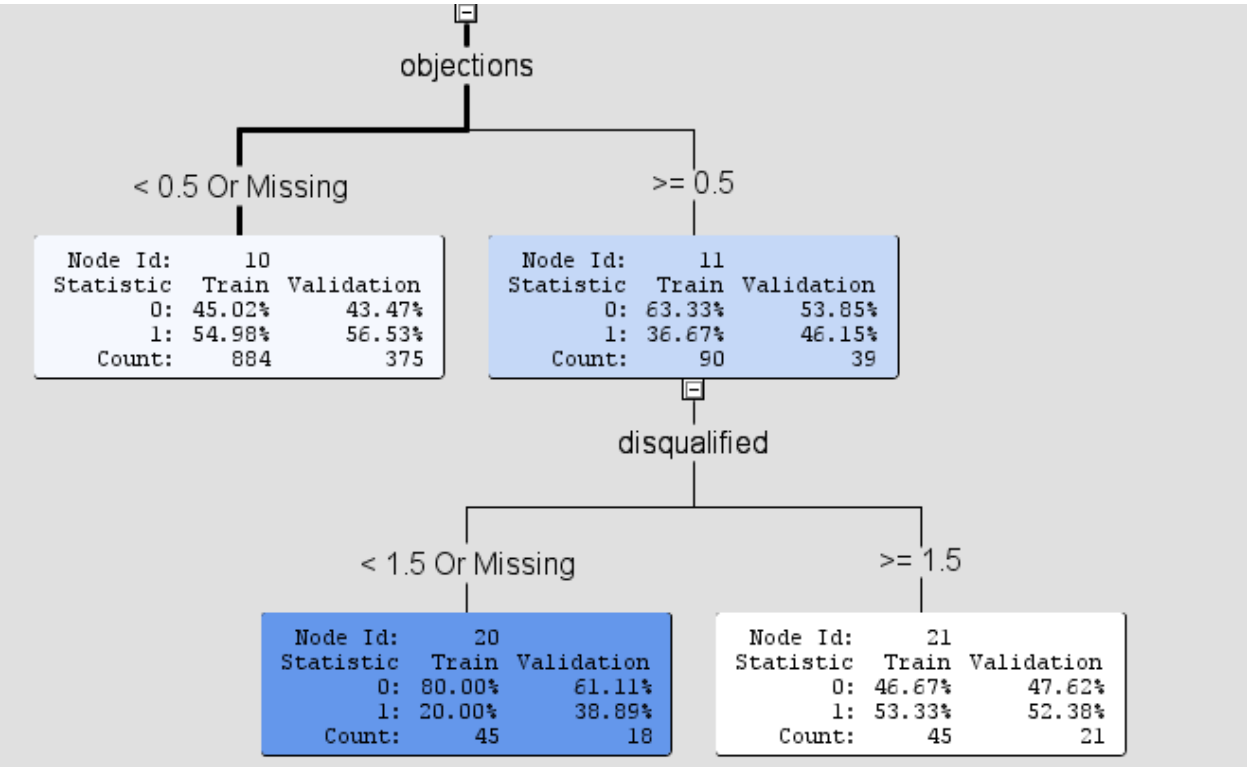

Fig. 2. Structure of the constructed decision tree

The naive Bayesian classifier

In the direct application of the naive Bayesian classifier without additional settings, the percentage of misclassifications was about 50%, which did not make it possible to use such a model for the analysis of companies. After adjustment and increasing the number of partitions, somewhat better results were obtained (Table 2), but still such a model is not recommended for use in practice.

Table 2

Statistical criteria for the naive Bayesian classifier

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.240782	2.41E-01
DIV	Divisor for ASE	5548	2384
MAX	Maximum Absolute Error	0.849634	8.18E-01
NOBS	Sum of Frequencies	2774	1192
RASE	Root Average Squared Error	0.490695	4.91E-01
SSE	Sum of Squared Errors	1335.859	575.7147
DISF	Frequency of Classified Cases	2774	1192
MISC	Misclassification Rate	0.460707	0.463926
WRONG	Number of Wrong Classifications	1278	553

Comparative analysis of the results and selection of the best model

Since classification problem was solved, the choice of the best model was conducted, based on the criterion of the number of misclassified examples, on the validation sample. This is due to the specificity of certain methods and the tendency to adapt to the training sample and, therefore, comparison with the use of validation sample was more justified. Simulation results are presented in

Table 3.

Table 3

Results of classification by different methods

Model	Misclassification Rate
Neural network	0.305369
Logistic regression	0.307047
Decision tree	0.313758
Naive Bayesian classifier	0.463926

Thus, a neural network, which makes it possible to predict with the accuracy of 70 % whether a company will continue to participate in public trading, turned out to be the best model for analyzing the data of ProZorro system. The absence of effective mechanisms for removing unfair participants from the trading system was confirmed and two potential groups of companies-ghosts – single-day and permanent – were found.

Conclusions

Analysis of tender purchasing and ensuring the possibility to participate in trades for all the participants irrespective of their relation to definite state bodies or collusions will enable achieving significant breakthrough in the fight against corruption, will be a powerful stimulus for small companies – Ukrainian producers – to participate in tenders and enable receiving additional revenues to the budget of Ukraine. The constructed models made it possible to classify such tender applications in the system, which contained collusion characteristics, to detect suspicious companies – participants of trades, which were only marionettes so that trades could take place. Complications in the construction and classification and, respectively, in obtaining accurate results were due to the inability to determine whether suspicious companies were real because they were not excluded from participation in trades, i.e. the existing classification model at on-line platform did not detect them and classified as real participants of trades. However, the results of the analysis and applications from corresponding public associations, mass media could be the reason for additional monitoring and inspection in order to confirm or to reject such suspicion. Analysis of the system itself is useful in terms of obtaining statistical information as to average number of participants in trades, effectiveness of the normative restrictions on the access to trading sites, convenience and transparency of the public procurement. Further research will aim at refinement of the proposed analysis by constructing behavioral models for predicting behavior of the real participants of trades and detecting illegal and untypical trade participants.

REFERENCES

1. The Law of Ukraine “About Public Procurement” [Electronic resource] / Verkhovna Rada of Ukraine. – Mode of access: http://zakon2.rada.gov.ua/laws/show/922-19. (Ukr).

2. ProZorro: public procurement [Electronic resource] / Mode of access: https://prozorro.gov.ua/. (Ukr).

3. Chubukova I. A. Data Mining / Chubukova I. A. – M .: Binom LBZ, 2008. – 384 p. (Rus).

4. Bruno G. R. Lean Compendium. Introduction to Modern Manufacturing Theory / Bruno G. R. – Springer International Publishing, 2018.

5. Zaichenko Y. P. Fundamentals of Intellectual Systems Design / Zaichenko Y. P. – K. : Slovo, 2006. – 352 p. (Ukr).

6. Bidiuk P. I. Analysis of time series / P. I. Bidiuk, V. D. Romanenko, O. L. Timoshchuk. – Kyiv : Polytechnica, 2013. – 600 p. (Rus).

7. Kuznietsova N. V. Information Technologies of Data Processing and Analysis in Financial Risk Management / N. V. Kuznietsova // Information Technologies and Special Security. – IPRI, 2015. – №1. – P. 86 – 98. (Ukr).

Editorial office received the paper 06.03.2018.
The paper was reviewed 12.03.2018.

Kuznietsova Nataliya – Cand. Sc. (Eng.), Ass. Prof. of the Department of Mathematical Methods of System Analysis, e-mail: natalia-kpi@ukr.net.

The Institute of Applied System Analysis of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".