УДК 004.942

Н. В. Кузнєцова, к. т. н., доц.

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ДЛЯ АНАЛІЗУ ФІНАНСОВИХ ЗЛОВЖИВАНЬ НА ПЛАТФОРМІ PROZORRO

У роботі показано можливості аналізу онлайн-закупівель на платформі ProZorro методами інтелектуального аналізу даних із метою виявлення поведінки компаній та характеристик, за якими можна встановити наявність змови та неправомірної діяльності під час участі в онлайн-торгах.

Ключові слова: *логістична регресія, нейронні мережі, тендерні закупівлі, ProZorro, інформаційні технології.*

Вступ

В Україні з метою забезпечення ефективного та прозорого здійснення закупівель, створення конкурентного середовища у сфері публічних закупівель, запобігання проявам корупції в цій галузі, розвитку добросовісної конкуренції Верховною Радою був прийнятий Закон «Про публічні закупівлі» [1]. Сама розробка та прийняття цього закону було суттєвим кроком з погляду боротьби з непрозорими договорами та змовами, які були основною причиною розтрати бюджетних коштів та завищення вартості робіт. Цей закон передбачає, що всі послуги та товари на суму більш ніж 50 тисяч гривень мають здійснюватися з використанням електронної системи закупівель із метою відбору постачальника товару (товарів), надавача послуги (послуг) та виконавця робіт для укладення договору; замовники повинні дотримуватися принципів здійснення публічних закупівель, установлених у законі. Для цього була створена система «ProZorro» – електронна база даних, за допомогою якої проводять публічні державні торги в режимі онлайн [2]. Сама система надає користувачам порталу можливість переглядати в реальному часі всі закупівлі, які відбуваються, і таким чином перевіряти прозорість закупівель, рівноправність доступу всіх учасників ринку та безпосередньо перевіряти, як витрачають податки громадян України на закупівлю послуг державного сектору.

Метою статті є аналіз та перевірка прозорості тендерних закупівель засобами інтелектуального аналізу даних, реалізованих у вигляді інформаційних технологій на платформі SAS із метою виявлення можливих порушень і зловживань. За результатами аналізу можна напрацювати рекомендації для подальшого вдосконалення системи моніторингу онлайн-закупівель та зменшити фінансові втрати платників податків, пов'язані з непрозорістю та недоступністю до торгів реальних конкурентоспроможних учасників ринку.

Моніторинг електронних закупівель

Для проведення перевірки доброчесності проведених торгів та закупівель Законом України [1] передбачена можливість звернення про наявність правопорушень через засоби масової інформації, через громадські об'єднання, та надання офіційної інформації від державної влади або органів місцевого самоврядування в разі виявлення порушень органом фінансового контролю або в разі автоматичного виявлення індикаторів ризику в проведених закупівлях.

У статті буде показано можливість удосконалення автоматичних індикаторів ризиків як спеціальних критеріїв із заданими наперед параметрами, за якими можна автоматично виявити ознаки правопорушень та зловживань в процедурі закупівель.

На сьогодні в базі ProZorro авторизовано 22 торгівельні майданчики: Zakupki.Prom.ua, e-tender, Newtend, SmartTender, «Держзакупівлі онлайн», PublicBid і «ПриватМаркет» тощо,

перелік яких постійно оновлюється на сайті [2]. Приєднатися до системи ProZorro можна через один із цих майданчиків, інших способів немає.

Для виявлення ознак порушення законодавства у сфері публічних закупівель може використовуватися інформація, оприлюднена в електронній системі закупівель, інформація, що міститься в єдиних державних реєстрах, інформація в базах даних, відкритих для доступу центральному органу виконавчої влади, що реалізує державну політику у сфері державного фінансового контролю, дані органів державної влади, органів місцевого самоврядування, підприємств, установ, організацій, замовників та учасників процедур закупівель, що можуть бути отримані органами державного фінансового контролю в порядку, установленому законом.

Аналіз електронних закупівель із метою виявлення певних закономірностей та напрацювання рекомендацій

Для покращення процедури моніторингу й доопрацювання системи автоматичного виявлення та відсікання недобросовісних компаній у системі реалізовано певну сукупність алгоритмів класифікації. Налаштування та параметри, за якими здійснюють класифікацію, у відкритому доступі відсутні з метою уникнення можливих маніпуляцій та надання хибних статистичних даних від компаній-заявників для участі в тендері. Тому метою моделювання буде виявити причинно-наслідкові зв'язки, статистичні характеристики для якісної класифікації, а також отримати інформацію в якості прогнозу підозрілих закупівель, за якими можна буде офіційно звернутися у встановленому законом порядку [1] для проведення перевірки.

Висувалось припущення, що існує певна взаємозалежність між тривалістю участі компанії в торгах та її характеристиками, а саме: чи залежить тривалість участі компанії в торгах від того, чи була вона запідозрена в неправомірній діяльності на платформі (у змовах з іншими компаніями). Емпіричним шляхом сформована така вибірка даних:

1. Wins – кількість виграшів певної компанії;
2. Losses – кількість програшів компанії в торгах;
3. Sum_of_deals – загальна сума виграних торгів;
4. Participations – загальна кількість участі в торгах;
5. Objections – кількість скарг, які подавала ця компанія;
6. Date_start – дата початку участі в системі торгів;
7. Date_finish – дата останньої участі в торгах;
8. IdTenderer – унікальний номер учасника тендеру;
9. Suspected – змінна, що показує, чи була фірма запідозрена в неправомірних змовах з іншими учасниками.
10. Churn out – цільова змінна, що дорівнює 1, якщо компанія переставала брати участь у торгах через короткий термін (уважають, що компанія раптово перестала брати участь або це була фіктивна компанія для одного торгу).

Компанію вважають такою, що продовжує вести торги, якщо проміжок часу між початком торгів на платформі і часом останнього торгу складає більше 60 днів (середня статистична тривалість бізнес-циклів компаній на платформі – згідно з офіційними даними ProZorro [2]). Варто також зазначити, що з вибірки було відфільтровано компанії, що провели та брали участь у конкурсі на платформі менше, ніж тричі.

Вхідна вибірка містила 3966 випадків, із них 1757 компаній, що перестали брати участь через досить короткий термін, а 2209 компаній продовжували брати участь у тендерах.

Вибірку було розбито на навчальну та перевірну вибірки за методом стратифікації щодо цільової змінної у відношенні 70/30.

Оскільки постановка завдання передбачає розв'язання задачі класифікації, то було запропоновано використати такі методи інтелектуального аналізу даних, як нейронні мережі, дерева рішень, логістичну регресію та Баєсівський класифікатор [3 – 5]. У якості критеріїв

обрання кращої моделі можуть бути використані [6]:

1) середньоквадратична похибка (MSE – mean squared error):

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

де $\hat{y}$ – оцінені значення залежної змінної за допомогою побудованої математичної моделі; y – фактичні значення залежної змінної;

2) сума квадратів похибок (SSE):

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

3) інформаційний критерій Акайке:

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

4) критерій Баєса – Шварца:

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

де $n = p + q + l$ – кількість параметрів моделі, які оцінюють за допомогою статистичних даних (p – кількість параметрів авторегресійної частини моделі; q – кількість параметрів ковзного середнього; „1" з'являється тоді, коли оцінюють зміщення (або перетин, тобто a_0), N – довжина вибірки.

5) частку неправильної класифікації (Misclassification Rate) обчислюють як відношення помилково спрогнозованих значень щодо загальної кількості значень N:

$$Misclassification\ Rate = \frac{кількість\ хибно\ спрогнозованих\ значень}{N} \quad (5).$$

Побудова описаних моделей та розрахунок статистичних критеріїв здійснювали на основі інформаційної технології SAS Enterprise Miner [7], обрання кращої моделі може бути здійснено автоматично або на основі заданого критерію якості. Послідовність аналізу тендерних закупівель представлена на рис. 1.

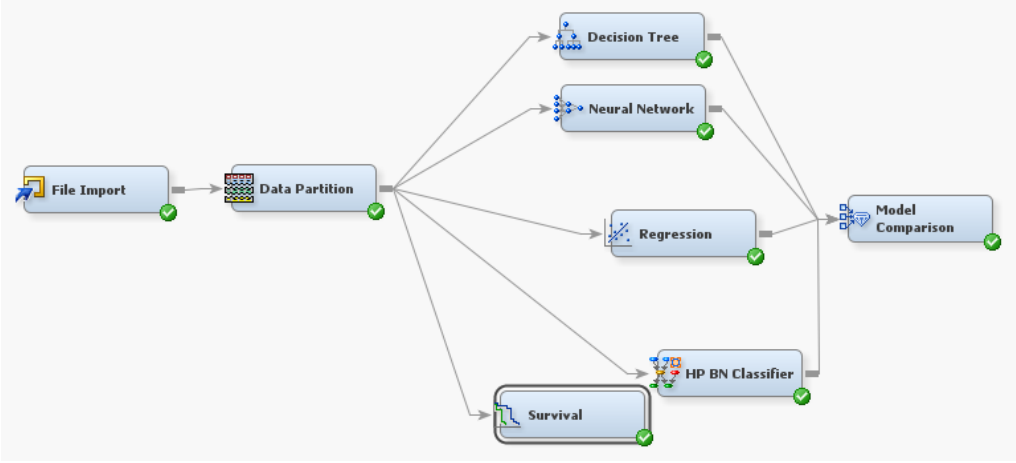

Рис. 1. Послідовність аналізу тендерних закупівель на основі інформаційної технології SAS Enterprise Miner

Нейронна мережа

Було побудовано різні види нейронних мереж [5], із різною кількістю шарів, різними активаційними функціями тощо. Критеріями якості було обрано інформаційний критерій Акайке, щоб не занадто ускладнювати саму модель та досягти балансу між параметрами та якістю моделі. Найкращою моделлю для вхідних даних виявилась проста перцептронна нейронна мережа із 20 прихованими шарами та стандартизацією входів на основі відхилення, радіальною комбінаційною функцією, логістичною активаційною функцією та критерієм навчання помилки класифікації. Статистичні критерії якості для найкращої нейронної мережі наведено в таблиці 1.

Таблиця 1

Статистичні характеристики найкращої нейронної мережі

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

Регресійна модель

Оскільки розв'язували задачу класифікації, то наступною моделлю, доцільною для виконання аналізу, чи перестане компанія брати участь у наступних тендерах (бінарний вихід: 0 – «ні» або 1 – «перестане»), було обрано логістичну регресію за методом побудови моделі stepwise з попарним уведенням та виведенням характеристики з моделі. Для цієї моделі були отримані такі характеристики: $AIC = 3301{,}941$ та $MisclassificationRate = 0{,}328767$.

Дерева рішень

Далі здійснювали моделювання на основі методу дерев рішень [3] із різними налаштуваннями правил відсікання, кількості нащадків та алгоритмів формування піддерев. Найкращим деревом виявилося дерево з мінімальним рівнем неправильної класифікації: на

навчальній вибірці $MisclassificationRate = 0{,}327325$ та перевірній вибірці: $MisclassificationRate = 0{,}313758$. Структура дерева наведена на рис. 2.

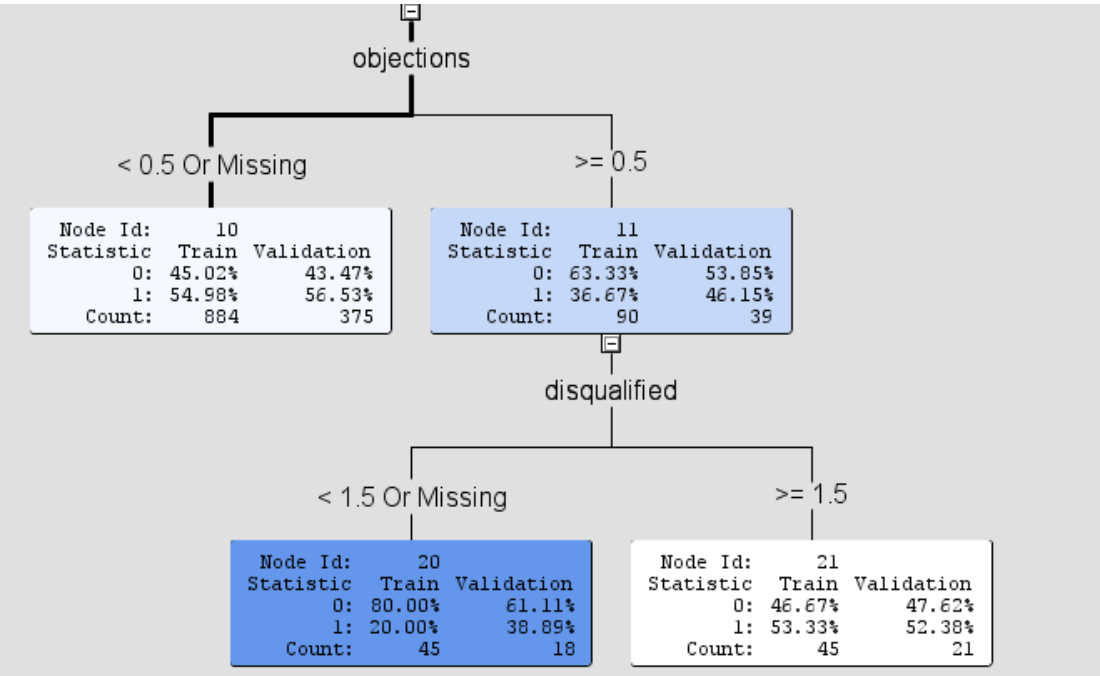

Рис. 2. Структура побудованого дерева рішень

Наївний Баєсівський класифікатор

За прямого використання наївного Баєсівського класифікатора без додаткових налаштувань відсоток неправильної класифікації становив близько 50%, що не дозволяло використовувати таку модель для аналізу компаній. За налаштування й збільшення кількості розбиттів було отримано дещо кращі результати (табл. 2), проте все одно таку модель не рекомендуємо для застосування на практиці:

Таблиця 2

Статистичні критерії для наївного Баєсівського класифікатора

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

Порівняльний аналіз результатів та вибір кращої моделі

Оскільки розв'язували задачу класифікації, то вибір кращої моделі здійснювали на основі критерію кількості неправильно класифікованих прикладів на перевірній вибірці. Це пов'язано зі специфічністю певних методів і схильністю прилаштовуватись до навчальної

вибірки, тому порівняння на перевірній вибірці є обґрунтованішим. Результати моделювання наведено в таблиці 3.

Таблиця 3

Результати класифікації різними методами

Модель	Misclassification Rate
Нейронна мережа	0.305369
Логістична регресія	0.307047
Дерево рішень	0.313758
Наївний Баєсівський класифікатор	0.463926

Отже, найкращою моделлю для аналізу даних ProZorro виявилась нейронна мережа, яка з точністю 70 % дозволяє спрогнозувати, чи буде продовжувати компанія брати участь в публічних торгах. Було підтверджено відсутність дієвих механізмів вилучення недоброчесних учасників із системи торгів, а також знайдено дві потенційні групи компаній-привидів – одноденні та ті, що постійно діють.

Висновки

Аналіз тендерних закупівель і можливість участі в торгах усіх учасників без прив'язки до конкретних державних органів чи змов дозволить нашій країні здійснити суттєвий прорив у боротьбі з корупцією, надасть потужний поштовх для участі невеликих компаній українських виробників у тендерах і дозволить отримати додаткові надходження до бюджету України. Побудовані моделі дозволили класифікувати в системі тендерні заявки, що містять характеристики змови, виявити підозрілих учасників торгів, які виступали лише як маріонетки для того, щоб торги відбулися. Складнощі в побудові і класифікації, а відповідно й точності отриманих результатів були пов'язані з неможливістю точно визначити, чи підозрілі компанії є реальними чи ні, оскільки вони не були відхилені від участі в торгах, тобто чинна класифікаційна модель на онлайн-платформі пропустила і класифікувала їх як реальних учасників торгів. Проте результати аналізу і звернення відповідних громадських об'єднань, засобів масової інформації можуть стати причиною для проведення повторного моніторингу й розслідування, щоб підтвердити або спростувати таку підозру. Проведений аналіз самої системи є корисним із погляду виявлення статистичної інформації щодо середньої кількості учасників торгів, ефективності встановлених нормативних обмежень для доступу до торговельних майданчиків, зручності і прозорості публічних закупівель. У подальших дослідженнях передбачаємо виконати вдосконалення запропонованого аналізу шляхом побудови поведінкових моделей для прогнозування поведінки реальних учасників закупівель і виявлення шахрайських та нетипових учасників торгів.

СПИСОК ЛІТЕРАТУРИ

1. Закон України Про публічні закупівлі [Електронний ресурс] / Верховна Рада України. – Режим доступу: http://zakon2.rada.gov.ua/laws/show/922-19.

2. ProZorro: публічні закупівлі [Електронний ресурс] / Режим доступу: https://prozorro.gov.ua/.

3. Чубукова И. А. Data Mining / Чубукова И. А. – М. : Бином ЛБЗ, 2008. – 384 с.

4. Bruno G. R. Lean Compendium. Introduction to Modern Manufacturing Theory / Bruno G. R. – Springer International Publishing, 2018. – 149 p.

5. Зайченко Ю. П. Основи проектування інтелектуальних систем / Зайченко Ю. П. – К. Слово, 2006. – 352 с.

6. Бидюк П. И. Анализ временных рядов / П. И. Бидюк, В. Д. Романенко, О. Л. Тимощук. – Киев : Политехника, 2013. – 600 с.

7. Кузнєцова Н. В. Інформаційні технології обробки та аналізу даних у фінансовому ризик-менеджменті / Н. В. Кузнєцова // Інформаційні технології та спеціальна безпека. – ІПРІ, 2015. – №1. – С. 86 – 98.

Стаття надійшла до редакції 06.03.2018 р.

Стаття пройшла рецензування 12.03.2018 р.

Кузнєцова Наталія Володимирівна – к. т. н., доцент кафедри математичних методів системного аналізу, e-mail: natalia-kpi@ukr.net.

Інститут прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».