ЕКОНОМЕТРИЧНА МОДЕЛЬ НАЦІОНАЛЬНОГО БАНКУ УКРАЇНИ ДЛЯ ОЦІНКИ КРЕДИТНОГО РИЗИКУ БАНКУ ТА АЛЬТЕРНАТИВНИЙ МЕТОД ОПОРНИХ ВЕКТОРІВ

Дмитро Покідін
Національний банк України

Розвиток економетричних моделей кредитного ризику розпочався із z-моделі Альтмана у 1968 році. З того часу моделі стали більш складними, поширення набули такі методи машинного навчання, як штучні нейронні мережі (ANN) та метод опорних векторів (SVM). Ця стаття фокусується на застосуванні SVM, як моделі для передбачення дефолту. Вона починається із загального теоретичного вступу до SVM та деяких розповсюджених його альтернатив. Далі із використанням цих підходів будуються моделі оцінки кредитного ризику на основі даних НБУ по банківським клієнтам, що дає змогу порівняти точність SVM з точністю інших моделей. Хоча модель SVM загалом більш точна, в силу певних особливостей, описаних у статті, застосування SVM є дещо суперечливим. У статті також презентовано результати моделі логістичної регресії (Logit), яка буде використовуватися НБУ для оцінки кредитних ризиків комерційних банків.

JEL: C45, C51, C52, C53
Ключові слова: машинне навчання, кредитний ризик, скорингова модель.

I. Вступ

Кредитний ризик, по суті, є ймовірністю того, що певний контрагент не зможе виконати свої зобов'язання перед банком щодо повернення позики. Базельський комітет з питань банківського нагляду приділяє велику увагу розвитку належного підґрунтя для кількісної оцінки ризику та заохочує впровадження підходу на основі внутрішнього рейтингу (IRB). Банкам рекомендується розробляти власні внутрішні моделі, аби належним чином оцінювати своїх клієнтів, що сприятиме наявності в банків достатнього обсягу капіталу для покриття очікуваних збитків.

Для розрахунку кредитного ризику НБУ зобов'язує банки використовувати одну конкретну економетричну модель. Причина цього в тому, що багато українських банків не мають добре налагоджених кредитних процесів, заснованих на загальноприйнятих статистичних підходах, тому НБУ ще не може повною мірою довіряти їхній оцінці кредитного ризику. Великі міжнародні групи мають такі процеси, однак вони дуже відрізняються. Крім того, деякі із цих груп користуються уніфікованими моделями, наданими головним офісом, проте вони можуть бути недостатньо адаптовані до української економіки.

▪ *Стаття є перекладом оригінальної статті англійською мовою. У разі будь-яких розбіжностей між оригінальною статтею та її перекладом українською мовою англомовна версія статті має переважний статус.*

У діючій постанові Правління Національного банку України № 23 від 25.01.2012, що визначає регуляторні правила оцінки резервів за активними операціями, роль такої моделі не була настільки важливою. Категорія якості кредиту визначалася як з урахуванням фінансового класу, що був визначений моделлю, так і якістю обслуговування кредиту. Внаслідок цього навіть компанія з поганою фінансовою звітністю з точки зору моделі, але яка ще не має простроченої заборгованості, розглядається як відносно платоспроможна.

Нове положення про оцінку кредитних ризиків передбачатиме, що визначальну роль для встановлення платоспроможності позичальника відіграватиме фінансовий стан позичальника – резервування буде в основному[1] визначатись фінансами підприємства. Тому питання вибору моделі оцінки кредитного ризику набуває більшої ваги. В цій статті із використанням даних українських банків порівнюється точність трьох моделей: лінійна дискримінантна модель (z-модель Альтмана, чи просто LDA), Logit та SVM.

II. Теоретична інформація

2.1. Лінійна дискримінантна модель (LDA)

У 1968 році Едвард Альтман запропонував лінійний дискримінантний аналіз у сфері прогнозування дефолтів (невиконання зобов'язань). З того часу він став дуже популярним, головним чином завдяки своїй простоті і відносно точним результатам. У даний час вона використовується НБУ як основна модель кредитного ризику відповідно до постанови, що має бути скасована (Постанова Правління Національного банку України "Про затвердження Положення про порядок формування та використання банками України резервів для відшкодування можливих втрат за активними банківськими операціями" № 23 від 25.01.2012).

LDA може використовуватися для мультикласифікації; однак у рамках прогнозування дефолтів ми маємо лише два класи: платоспроможні і неплатоспроможні компанії. Нехай π_i є апріорною вірогідністю класу i, $p\ (x|i)$ – умовним розподілом незалежних змінних *x*. Тоді розподіл апостеріорної ймовірності може бути записаний у такому вигляді:

$$p(i|x) = \pi_i p(x|i)$$

Допускається, що розподіл $p(x|i)$ є багатомірно нормальним:

$$p(x|i) = N(x|\mu_i, \sum),$$

де μ_i - це умовні середні значення, $\sum$ – матриця коваріацій. $\sum$ не має індексу *i*, оскільки допускається, що ця матриця однакова для обох класів[2] .

Оскільки ми маємо лише два класи, позначимо їх як i = 0 та i = 1. Тоді, за умови лінійно роздільних даних, $\pi_0\ N(x|\mu_0,\sum) \neq \pi_1\ N(x|\mu_1,\sum)$ (звертаємо увагу, що π_1 також можливо записати як $(1-\pi_0)$). Ми віднесемо *x* до класу 0, якщо $\pi_0 N(x|\mu_0,\sum) > \pi_1 N(x|\mu_1,\sum)$, та в іншому випадку – до класу 1. Ґрунтуючись на цьому, можемо визначити точку, яка розділяє класи (розділяюча лінія) як

$$\ln\left(\frac{\pi_1 N(x|\mu_1,\sum)}{\pi_0 N(x|\mu_0,\sum)}\right) = 0.$$

Може бути доведено, що ця формула має ідентично представлена у вигляді простого лінійного рівняння

$$w^T x + w_0 = 0,$$

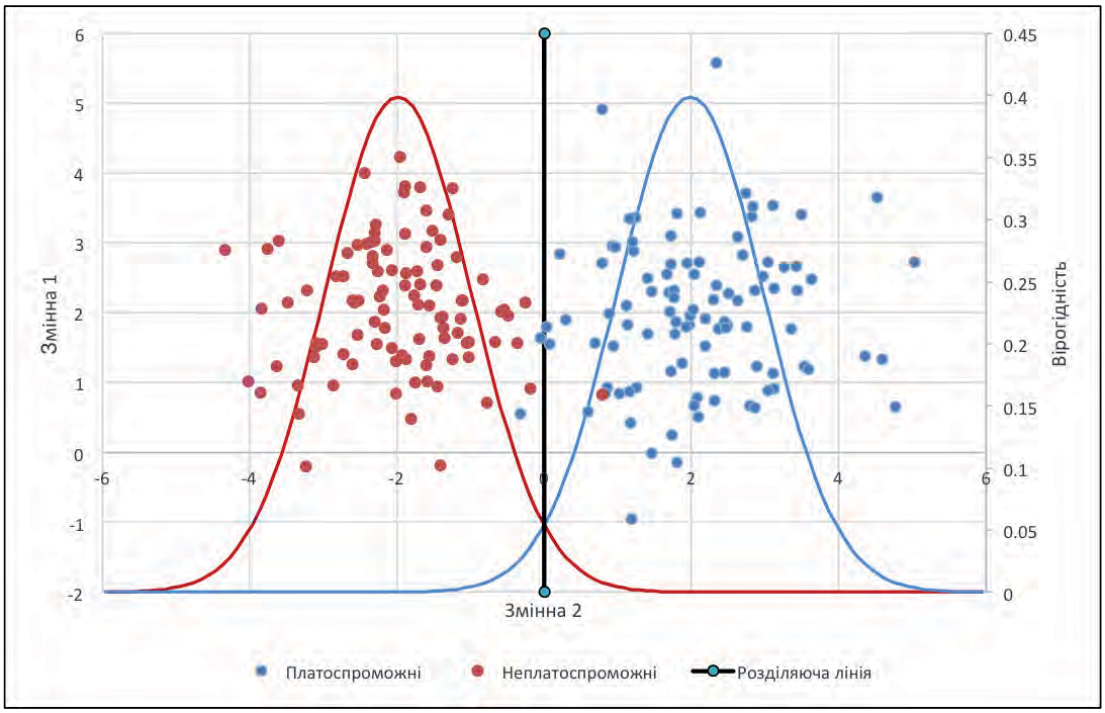
де w – вагові коефіцієнти, які потрібно оцінити.

Розглянемо ілюстрований приклад із рисунка 1. Червоні точки – це платоспроможні компанії, блакитні – неплатоспроможні. Видно, що їхні середні значення відрізняються. Нормальний розподіл накладено зверху, навколо цих середніх значень. Ми чітко бачимо, що через точку перетину двох розподілів, тобто там, де вірогідність потрапляння до кожного класу однакова, розміщена розділяюча лінія. Якщо точка потрапляє лівіше за лінію, вірогідність бути неплатоспроможним вища, ніж платоспроможним, тому точка класифікується відповідним чином.

1 У постанові залишаються деякі якісні характеристики, які можуть скоригувати результат моделі.

2 Мета припущення – зробити рівняння моделі лінійним. Для ознайомлення з повною теорією можна звернутися до Venables W. N. and Ripley B. D. (2002).

Рисунок 1. Принцип LDA



Вважається[3], що модель дає відносно погані результати, коли вищезгадані припущення не виправдовуються, а це зазвичай так і є – фінансові коефіцієнти рідко розподіляються нормально хоча б тому, що багато з них не приймають значень нижче за 0, та малоймовірно те, що платоспроможні і неплатоспроможні компанії мають однакові матриці коваріацій між фінансовими коефіцієнтами, тому що, інтуїтивно, компанії з діаметрально протилежним статусом платоспроможності можуть мати різну відносність між змінними.

2.2. Логістична регресія

Тоді як LDA є лінійною параметричною моделлю, Logit-модель є нелінійною параметричною моделлю. Модель послаблює припущення багатомірного нормально розподілу та рівності матриць коваріацій і замість цього допускає логістичний розподіл вихідної змінної.

Нехай *x* буде незалежною змінною (у нашому випадку фінансовий коефіцієнт), β – коефіцієнт перед x. Припустимо, рівняння $\beta^T x + \beta_0$ визначає значення змінної z, яка потім іде до логістичної кумулятивної функції розподілу (CDF) $\Phi(.)$ як параметр. Тоді кожна компанія і має власну вірогідність дефолту (PD):

$$PD(\mathrm{x_i}) = \Phi(\beta^{\mathrm{T}}\mathrm{x_i}+\beta_0) = \Phi(z_i). \qquad (1)$$

Очевидна задача – це максимізувати значення (1) для неплатоспроможних компаній (позначимо їх як y = 1) та мінімізувати його для платоспроможних (y = 0). Можемо записати це як

$$\max_\beta \prod_{i=1}^{n}[\Phi(\beta^T x_i + \beta_0)]^{y_i}[1 - \Phi(\beta^T x_i + \beta_0)]^{1-y_i}, \qquad (2)$$

де n – розмір вибірки. Іншими словами, варіюючи β, ми намагаємося максимізувати добуток вірогідності дефолту PD (для неплатоспроможних компаній) чи вірогідності виживання (для платоспроможних компаній), що дорівнює 1 – PD. Ця процедура називається методом максимальної вірогідності (maximum likelihood estimate). Зазвичай логарифм (2) береться для спрощення розрахунків, адже логарифм добутку дорівнює просто сумі логарифмів[4].

3 Pohar M., Blas M., and Turk S. (2004) вивчали поведінку LDA та Logit за умов ненормального розподілу.

4 Hosmer D. W., Lemeshow S. (2000) надають детальніший опис Logit.

Рисунок 2. Принцип Logit

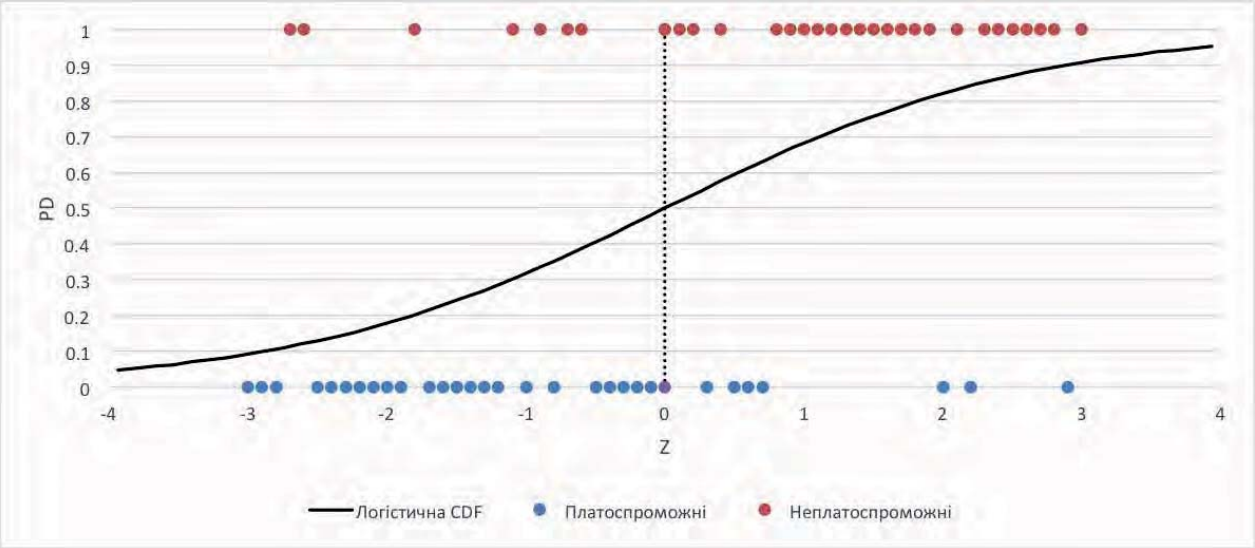


Рисунок 2 дає ілюстрацію Logit, позначення лишаються такими ж. На горизонтальній осі знаходиться z. Після (2) β встановлені на такому рівні, що z у середньому максимально різний між класами. Логістична CDF (чорна лінія) загалом вища для неплатоспроможних компаній. Однак точки, що розташовуються вгорі ліворуч, – це, звісно, помилки моделі, так само, як і точки внизу праворуч.

2.3. Метод опорних векторів

Bernhard E. Boser, Isabelle M. Guyon, Vladimir N. Vapnik (1995), впровадили SVM як нелінійний непараметричний алгоритм для класифікації. Останнім часом він стає дедалі популярнішим у рамках прогнозування дефолтів, оскільки все більше дослідників тестують і далі розвивають цю модель. Західні комерційні банки та рейтингові агентства також зацікавлені в цьому, і багато з них упроваджують SVM та пов'язані навчальні машинні методи в свою діяльність (McKinsey, 2015).

Регулятори також не уникають цих моделей. Дойче Бундесбанк використовував SVM під час кредитного скорингу нефінансових компаній до 2012 (Adrian Costeiu, Florian Negu (2013))[5].

Розглянемо лінійне рівняння $h(x) = w^T x + b$, де x (як звичайно) є вектором незалежних змінних, w є вектором вагових коефіцієнтів і b є перетином. Це рівняння формує розділяючу гіперплощину, коли воно дорівнює нулю.

Так звана "маржа" відіграє ключову роль у SVM. Це, по суті, відстань від точки до розділяючої гіперплощини. Ми можемо розрізняти функціональну маржу і геометричну маржу. Функціональна маржа може бути сформульована як

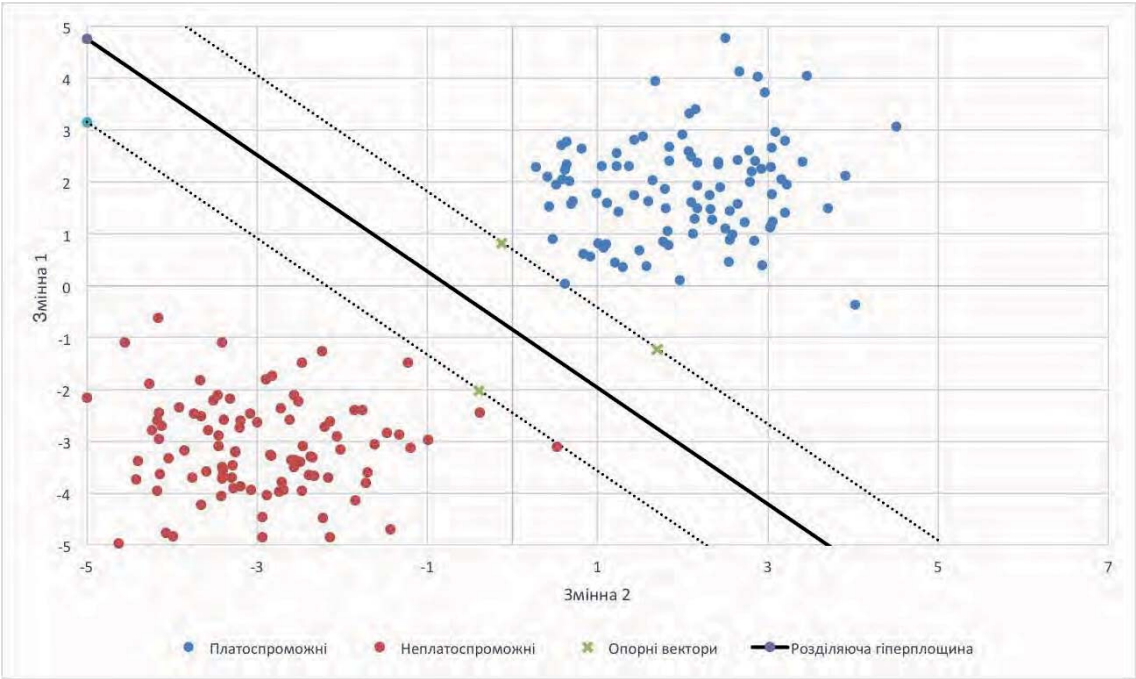
$$y = y(w^T x + b), \qquad (3)$$

де y вказує на змінну, яка набуває значення +1, якщо компанія не виконує зобов'язань (є дефолтною) і –1 – в іншому випадку. Таким чином, ми класифікуємо компанію як дефолтну, якщо значення (3) більше від нуля. Чим більше таке значення, тим більше ми впевнені в нашому прогнозуванні. Рисунок 3 ілюструє SVM на прикладі лінійно абсолютно роздільних класів. Розділяюча гіперплощина розташована таким чином, що "маржа" між найближчими точками (опорними векторами) та площиною максимальна. Зверніть увагу, що для лінійного випадку повинно бути щонайменше три таких точки. В іншому випадку лінія може бути проведена безкінечно багатьма способами. Тобто точки ніби слугують опорою для лінії. Звідси і назва методу.

Тим не менше (3) не може бути надійною мірою, тому що після зміни w і b (множення або ділення на довільне число) прогнозування залишається тим самим, але значення (3) змінюється (тобто ми можемо зробити його яким завгодно великим, що може оманливо свідчити про надійне прогнозування).

5 У 2012 модель було замінено на складнішу інтегральну модель, яка складається з кількох допоміжних.

Рисунок 3. Принцип SVM



Для подолання цієї проблеми надалі запровадимо поняття геометричної маржі. Замість того, щоб просто використовувати w і b, нормалізуємо їх таким чином, що вони тепер стали $\frac{w}{||w||}$ та $\frac{b}{||w||}$. Це означає, що тепер параметри стандартні і їхня довжина дорівнює одиниці. Формула (3) має тепер такий вигляд:

$$\gamma = \frac{y(w^T x+b)}{||w||}.$$

Принципом SVM є знайти набір параметрів, який максимально збільшує мінімальну маржу точок кожного класу (як на рисунку 3). Це зроблено шляхом формулювання оптимізаційної проблеми[6]:

$$\min_{w,b} \frac{||w||}{2} + C\sum_{i=1}^{n} \varepsilon_i \qquad (4)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1, \qquad i = 1, \dots, n$$

$$\varepsilon_i, w \geq 0, \qquad i = 1, \dots, n, \; n \text{ - розмір вибірки.}$$

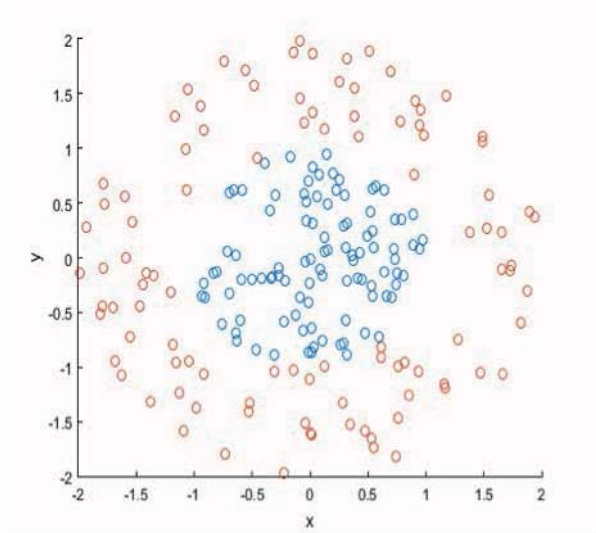
ε_i у цій формулі є параметром, який дозволяє певну частку похибки (для нероздільного випадку), С контролює кількість таких неправильних класифікацій. Якщо *С* занадто велике, то буде менше випадків неправильної класифікації і водночас зростає ризик перепідгінки.

Також кращою SVM робить використання функцій кернела. Функції кернела перетворюють функціональну форму вхідних змінних і переносять їх у багатомірну площину (випрямляюча площина). У випрямляючій площині точки, які були лінійно нероздільними в оригінальному просторі, як правило, можуть бути легко розділені. Принцип кернела найкраще проілюструвати на такому прикладі. Припустимо, в нас є тільки два фінансових коефіцієнти, які можна використати для прогнозування (x та y). Якщо це так, то ми працюємо всього у двох вимірах. Розглянемо рисунок 4 а). Нехай червоні кола будуть платоспроможними компаніями, а сині – неплатоспроможними. Очевидно, що жодна лінія не може відокремити точки одна від одної. Але що, коли ми не хочемо обмежувати себе лише двома вимірами? Давайте трансформуємо точки таким чином, щоб у них тепер був ще один, третій вимір із координатою ($x^2 + y^2$). Отримане зображення проілюстровано на рисунку 4 b). Результатом цього є те, що точки можуть бути розділені лінійно.

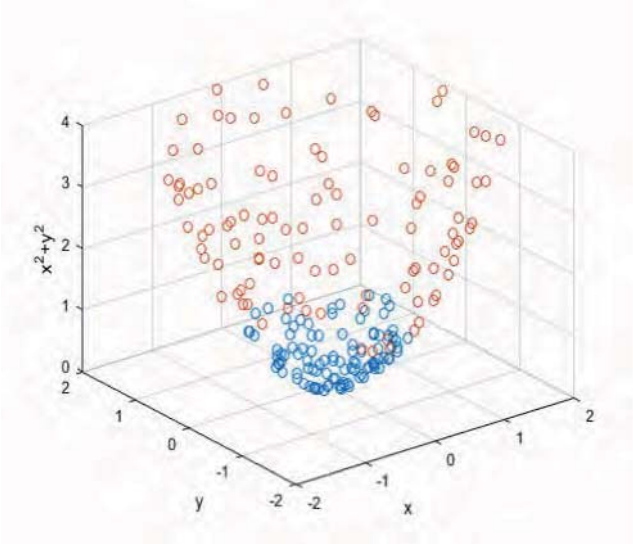
6 Детальне виведення не є метою цієї статті. Для поглибленої теорії див. Andrew Ng, Stanford University, CS229 Lecture notes.

Рисунок 4. Ілюстрація перетворення кернела

a) Оригінальна площина



b) Випрямляюча площина



Наступним кроком є представлення дуальності Лагранжа, формула в (4) називається первинною формою. Її дуальна форма (вже із функцією кернела) має такий вигляд:

$$\max_a \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j a_i a_j K(x_i, x_j); \qquad (5)$$

$$\text{s.t. } 0 \le a_i \le C, \quad i = 1, \dots, n;$$

$\sum_{i=1}^{n} a_i y_i = 0$, $i = 1, \dots, n$; n - розмір вибірки.

a' в (5) є множниками Лагранжа з первинної форми. Зверніть увагу, що параметр C обмежує $a's$ зверху. Під час оптимізації більшість a перетворюються на нуль, ненульові а відповідають опорним векторам.

Тепер формула для прогнозування має такий вигляд:

$$h(x) = \sum_{i=1}^{k} a_i y_i K(x_i, x) + b,$$

де *k* є кількістю опорних векторів.

Незважаючи на очевидні переваги, SVM має певні недоліки в застосуванні кредитного скорингу. Ми обговоримо їх у наступному розділі.

Пригадайте з (4), що параметр C відповідальний за рівень похибки. Чим він більший, тим точніша модель на тестовій вибірці. Однак велике значення C часто призводить до перепідгінки. Таким чином, слід знайти компроміс між точністю і перепідгінкою під час прийняття рішення щодо C.

Крім того, слід вибрати конкретний тип функції кернела. У цьому випадку було обрано функцію кернела Гауссена, яка має такий вигляд:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right)$$

Кернел Гаусенна, ймовірно, є найпопулярнішим через його обчислювальну ефективність. Параметр σ у формулі вище називається масштабним параметром кернела. Він також підлягає оптимізації. Ці два параметри підібрані таким чином, щоб максимально збільшити GINI моделі[7].

Ще один параметр, який має бути відкоригований, – це апріорна ймовірність кожного класу. Для цієї праці було обрано рівну ймовірність кожного класу. Це означає, що в моделі надається однакова вага платоспроможним і неплатоспроможним компаніями під час оптимізації.

III. Модель

У цьому розділі три зазначені моделі побудовано і випробувано з метою визначення оптимальної.

3.1. Дані[8]

Для побудови моделей використано дані НБУ щодо фінансової звітності більш ніж 8 000 приватних підприємств[9]. Ці дані доопрацювали, оскільки окремі компанії (які, як вважалося, пов'язані з деякими банками) було виключено з аналізу[10].

Дані розділили за розміром підприємств (великих і малих), а потім – за галузями (сільське господарство, виробництво, торгівля та інші). Раніше був детальніший розподіл на галузі. Рішення об'єднати розподіл зумовлено відсутністю досить великої вибірки. Якби оригінальні галузі було взято до аналізу, то в кожній залишилося б близько ста компаній. Вибір конкретних кластерів зроблено після проведення кластерного аналізу, який виявив подібності в структурі балансу між вищевказаними кластерами. Треба зазначити, що нами використано лише коефіцієнти, які описують різницю в структурах балансу. Коефіцієнти, які могли вказувати на стан платоспроможності, не використовувалися.

Для цілей цього документа термін "дефолтні" розуміємо як досягнення категорії неплатоспроможних підприємств відповідно до Постанови №23[11] станом на 01.01.2015, тоді як фінансова звітність була станом на 01.01.2014, тобто з інтервалом в один рік.

3.2. Змінні

Для моделювання зроблено первинний довгий список фінансових коефіцієнтів (таблиця 1). Вони охоплюють різні види рентабельності, ліквідності, оборотності і платоспроможності. Всі змінні було обмежено 5 та 95 перцентилем. Подібні змінні були використані у роботі W. K. Hardle, R. A. Moro, D. Schafer (2009), вони побудували схожу модель для Дойче Бундесбанка.

3.3. Критерії ефективності

Основні критерії ефективності Коефіцієнта точності (Accuracy ratio). Для Logit також застосовувався псевдо-R^2. Обидві міри вказують, як добре модель спроможна розділити платоспроможні компанії та неплатоспроможні.

Коефіцієнт точності (GINI)

Його також називають коефіцієнтом GINI. Загалом він свідчить, наскільки точно модель ідентифікувала дефолтні компанії відносно недефолтних. Він є похідним від кривої отримувача операційних характеристик (ROC).

7 Перелік значень параметрів, відібраних для кожної моделі, зазначено в Додатку, Таблиці В.

8 Автор висловлює подяку фахівцям Департаменту управління ризиками НБУ, зокрема, Олександру Фостику та Дмитру Шарову, за значну допомогу та участь у створенні Logit-моделі, а також у створенні переліку незалежних змінних та розподілу за групами.

9 Компанії, що знаходяться на Донбасі та в Криму, виключено з вибірки, оскільки вони банкрутували з неекономічних причин.

10 Для таких компаній постановою передбачено список якісних характеристик, які підвищують вірогідність дефолту.

11 Зазвичай це компанії, які мають більш ніж 90 днів заборгованості, однак там є інші умови.

Таблиця 1. Первинний ряд змінних

Змінна	Формула	Змінна	Формула
K1	Операційний прибуток / Дохід	K17	*K*8 + *K*15 – *K*16
K2	EBITDA / Дохід	K18	Фінансові зобов'язання / Частка акціонерів
K3	EBIT / Дохід	K19	Фінансові зобов'язання / *EBITDA*
K4	Поточні активи – поточні зобов'язання / Поточні активи	K20	Частка акціонерів / Загальні активи
K5	Чистий прибуток / Дохід	K21	Поточні активи / Поточні зобов'язання
K6	Чистий прибуток / Частка акціонерів	K22	Найліквідніші поточні активи / Поточні зобов'язання
K7	Чистий прибуток / Загальні активи	K23	*EBIT* / Фінансові витрати
K8	Запаси / Вартість проданих товарів	K24	Фінансові зобов'язання / Дохід
K9	*Дебіторська заборгованість* / Дохід	K25	Поточні активи – поточні зобов'язання / Частка акціонерів
K10	Кредиторська заборгованість / Дохід	K26	EBITDA / Фінансові витрати
K11	Загальні активи / Дохід	K27	Фінансові зобов'язання / EBTDA
K12	Поточні активи / Дохід	K28	EBITDA / Короткотермінові фінансові зобов'язання + фінансові витрати
K13	Фінансові активи / Дохід	K29	Оборотний капітал / Загальні активи
K14	K8 + K9 – K10	K30	Оборотний капітал / Дохід
K15	Дебіторська заборгованість для авансів / Дохід	K31	Фінансові зобов'язання / Чистий прибуток
K16	Кредиторська заборгованість для авансів / Дохід	K32	EBTDA / Дохід

Припустимо, у вас є прогнозування моделі, наприклад, z-значення у випадку LDA. Серед них є реально позитивні (TP), тобто дефолтні компанії, які визначено правильно; і помилково позитивні (FP), тобто недефолтні компанії, які визначено моделлю як дефолтні. Давайте одночасно додамо довільне значення кожному прогнозу та перерахуємо TP і FP. Повторюймо цей крок доти, доки FP не прийме всі значення в діапазоні {0 ; 1}. Крива ROC утворюється в 2-вимірному просторі, де FP розташовані на горизонтальній осі, а TP – на вертикальній.

Далі припустимо існування дуже неефективної моделі, яка дає суто випадкові передбачення. Теоретично кривою ROC такої моделі буде пряма лінія, що з'єднує точки (0.0) і (1.1). AR є областю між цією лінією і кривою ROC даної моделі. Іншими словами, це різниця між заданою моделлю та випадковою моделлю. Чим така різниця більша, тим краще.

3.4. Трансформація за Weight of Evidence (WOE)

Трансформація за WOE є, по суті, трансформацією безперервних змінних у дискретні змінні. Обґрунтуванням застосування такого підходу є те, що використання в LDA та Logit-моделі чистих даних дало погані результати. Коефіцієнт GINI дорівнював у середньому 0.2 – 0.3, що навіть не порівнянно із SVM. На жаль, часто якість української фінансової звітності низька, оскільки МСФЗ не обов'язкові для більшості підприємств, та фінансова звітність часто не проходить аудит. Таким чином, існує значне викривлення в даних, з якими не можуть упоратися LDA та Logit. Під викривленнями розуміємо певний тип контр-інтуїтивної залежності, яка має відбутися через похибку чи упущення у фінансовій звітності. Перетворення змінних у дискретні величини допомогло знизити це викривлення. Принцип полягає в тому, що:

1) кожна змінна розділена на певну кількість діапазонів від мінімального до максимального значення вибірки (перша колонка таблиці 2);

2) для кожного діапазону WOE розраховуються за формулою WOE_i = ln(%платоспроможні$_i$) - ln(%неплатоспроможні$_i$), %платоспроможні$_i$ – це частка недефолтних компаній у діапазоні із загальної кількості недефолтних компаній, а %неплатоспроможні$_i$ є часткою дефолтних компаній у діапазоні із загальної кількості дефолтних компаній (колонки 2 і 3 таблиці 2);

3) IV (інформаційне значення) для змінної розраховується за формулою:

$$IV = \sum_{i=1}^{n} (\%\text{платоспроможні}_i - \%\text{неплатоспроможні}_i) * WOE_i,$$

де n – кількість діапазонів. Це значення стає більшим, коли різниця між кількістю платоспроможних і неплатоспроможних компаній у кожному діапазоні збільшується (нижня права клітинка таблиці 2);

4) кількість діапазонів і їхні межі обрано таким чином, щоб максимально збільшити IV;

5) значення WOE (колонка 4) йдуть до рівняння моделі.

Приклад такого перетворення для деяких змінних наведено в таблиці 2.

Таблиця 2. Приклад перетворення WOE

Межі	Платоспроможні	Неплатоспроможні	WOE	IV
<–0.006	9	6	–1.3	0.25
<0.053	34	11	–0.57	0.12
<0.16	41	3	0.91	0.18
>0.16	42	3	0.93	0.19
Загалом	126	23	NaN	0.73

Дуже важливим з економічної точки зору є наявність монотонного тренду WOE. По суті, це означає, що зі збільшенням певної змінної WOE можуть тільки збільшуватися або зменшуватися. Неприпустимо мати, скажімо, коефіцієнт борг/EBITDA, який спочатку зменшує WOE, а потім раптово починає збільшувати, тому що збільшення боргового навантаження завжди має збільшувати PD.

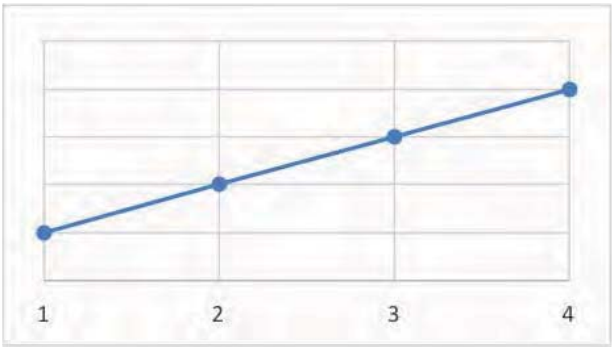
Звичайно, в подібному підході є і мінус – модель втрачає свою гнучкість, оскільки такі змінні можуть мати лише кілька значень. Припустимо, модель складається лише з однієї змінної Чистий прибуток/Виручка; припустимо також, що вона має лише два діапазони WOE – від мінус нескінченності до 0% та від 0% до плюс нескінченності. Нехай відповідні WEO будуть відповідно –1 та +1. Припустимо, ми прогнозуємо фінансовий стан трьох компаній – А, Б і В з відповідними значеннями змінної Чистий прибуток/Виручка –70%, –0.1% та 0.1%. Прогноз моделі для Б і В буде діаметрально протилежний, хоча різниця в коефіцієнті становить лише 0.2%. Водночас прогноз для А і Б буде однаковий, хоча компанія А, очевидно, значно гірша за компанію Б. Звісно, це спрощений приклад, але він чудово демонструє недоліки такого підходу.

Так чи інакше це допомогло значно підвищити ефективність моделі[12]. По суті, це робить модель трохи нелінійною. Припустимо, в нас є та сама змінна у двох окремих рівняннях із тим самим коефіцієнтом 1. Однак вона трансформована за WOE у другому рівнянні за прикладом із таблиці 2. Розглянемо рисунок 5. Горизонтальна вісь є кількістю значень змінних. Оскільки трансформована за WOE змінна має чотири діапазони, вона може набути лише чотири значення (рисунок 5 b)). У зв'язку з тим, що неперетворена змінна (рисунок 5 a)) є безперервною, вона може набувати будь-яке якого значення. Тому чотири значення з рівним кроком було обрано з метою порівняння їх із перетвореною змінною. Очевидно, що перетворена змінна WOE демонструє чітку нелінійну поведінку.

Зауважимо, що для SVM не потрібне таке перетворення, оскільки завдяки використанню оригінальних вхідних даних воно демонструє дуже гарні результати, а це очевидний плюс.

Рисунок 5.

a) Шлях неперетвореної змінної



b) Шлях перетвореної змінної

3.5. Вибір змінних

У зв'язку з перетворенням WOE процедура процесу відбору змінних відрізняється для Logit-моделі та LDA і SVM.

Logit-модель та LDA

1. Усі коефіцієнти порівнюються за IV. Змінні з найнижчою IV випадають з аналізу, тому що вони не можуть добре розділити класи.

2. Оцінюються кореляційна матриця й економічне обґрунтування знаків коефіцієнтів змінних у рівнянні, а також статистична значущість (відмінність від нуля). Висококорельовані, не обґрунтовані економічно або статистично незначні змінні видаляються.

12 По суті, роблячи цю трансформацію, ми підганяємо наші вхідні дані під те, що ми хотіли б бачити (зазначимо, що ми змушуємо тренд WOE бути економічно обґрунтованим)

3. Ряд змінних, що залишилися, переходять на етап крос-валідації, де додаткові змінні можуть бути відкинуті.

SVM

Для SVM було обрано процедуру форвардного відбору:

1. Вибір починається з пустої моделі (без змінних), потім одна зі змінних почергово додається до моделі;

2. Змінна, яка приносить найбільше GINI, додається в кінцевому підсумку;

3. Усі корельовані змінні (>|0.8|) з обраною змінною виключаються з вихідного ряду;

4. Потім процедура повторюється з тими змінними, що залишилися;

5. Це триває доти, доки додавання змінних не приводить до підвищення ефективності.

Третій крок необхідний для прискорення обчислень. Значно корельовані змінні, ймовірно, описують одну спільну сторону фінансового становища позичальника. Так, після того, як обрано кращу із цих змінних, усі інші видаляються, тому вони не беруть участі в наступному колі, скорочуючи таким чином час на обчислення.

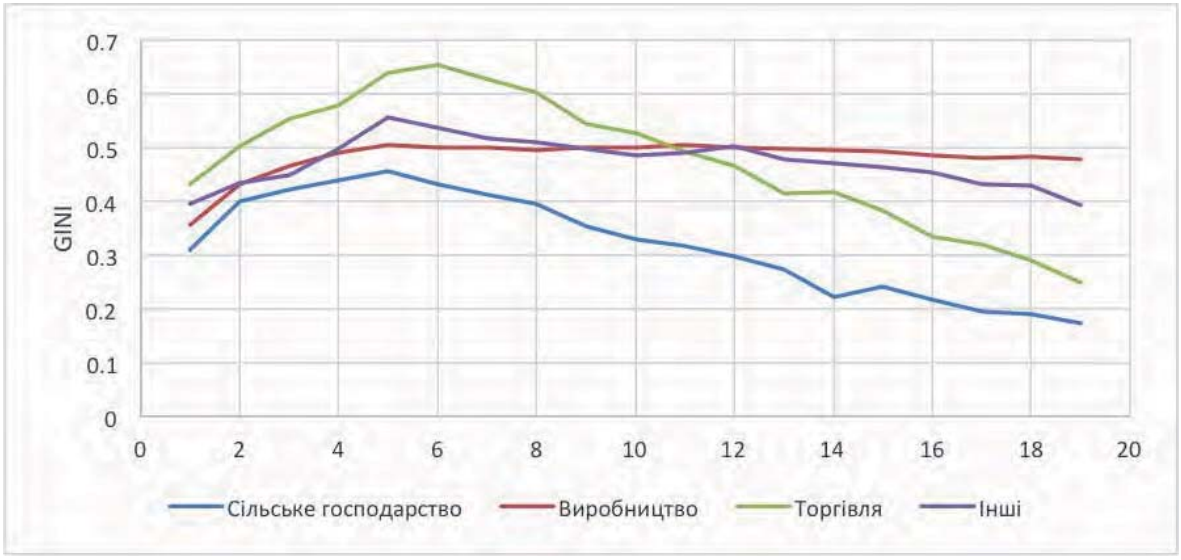
Звісно, немає можливості спробувати всі можливі комбінації змінних, тому зазвичай обирається ця або беквордна процедура[13]. Беквордна процедура є протилежною форвардній – модель починається зі всього набору змінних, потім змінні видаляються почергово.

Як бачимо, відбір змінних і етап крос-валідації для SVM об'єднано в один крок.

На рисунку 6 проілюстровано шлях GINI у форвардній процедурі для великих компаній. Видно, що після певної точки (найчастіше це 4 – 6 змінна) GINI починає зменшуватися. Це точка відсічення в процедурі відбору змінних для кожної моделі.

Таблиця 3 містить інформацію про змінні, обрані для кожної конкретної моделі після процедури відбору, описаної вище[14]. Не потрібно обманюватися тим фактом, що є не так багато змінних, які збігаються в LDA, Logit-моделі і

Рисунок 6. Шлях GINI в процесі відбору змінних SVM



13 Hardle W. K., Moro R. A., Schafer D. (2009) для додаткового прикладу обох процедур.

14 Див. Додаток для додаткової статистики для кожної змінної, Таблиця А і Таблиця Б.

SVM. Багато з них мають високу кореляцію, тому їх можна вважати замінниками одна одної. Наприклад, у кластері "Виробництво" великих компаній коефіцієнт К7 відсутній у рівнянні для SVM, однак в це рівняння входить коефіцієнт К3, який має кореляцію 0.78 з К7. Водночас рівняння LDA і Logit-моделі не містить коефіцієнта К12, однак включає К30, їхня кореляція становить 0.81. Це означає, що незважаючи на велику різницю у змінних, економічна основа за ними набагато ближча, ніж може видатися.

Таблиця 3. Відбір змінних для кожної моделі

Великі компанії							
Сільське господарство		**Виробництво**		**Торгівля**		**Інші**	
LDA&Logit-модель	SVM	LDA&Logit-модель	SVM	LDA&Logit-модель	SVM	LDA&Logit-модель	SVM
К10	К8	К7	К3	К11	К9	К10	К10
К11	К10	К20	К12	К14	К14	К22	К12
К24	К19	К23	К16	К15	К20	К27	К21
К25	К22	К24	К24	К21	К24	К29	К32
	К24	К30	К25	К23	К25	К30	К27
	К25				К31		
Малі компанії							
Сільське господарство		**Виробництво**		**Торгівля**		**Інші**	
LDA&Logit-модель	SVM	LDA&Logit-модель	SVM	LDA&Logit-модель	SVM	LDA&Logit-модель	SVM
К7	К7	К1	К1	К1	К8	К5	К1
К9	К11	К24	К10	К9	К13	К8	К8
К18	К21	К29	К24	К21	К18	К11	К9
К27			К27	К24	К20	К20	К12
К29				К31	К24	К31	К18
К30							К24
							К31

Той факт, що моделі не різні за специфікацією, ускладнює їх пряме порівняння. Замість цього, вірніше було б стверджувати, що порівняння як моделей так і процедур підбору змінних було проведено.

3.6. Крос-валідація

Було розроблено дуже ретельну процедуру валідації:

1. Вибірка випадковим чином ділиться на тренувальну і тестову в пропорції 70%/30% 100 разів;

2. Кожного разу розраховуються критерії ефективності;

3. Після завершення етапу 2 беруться медіанні значення критеріїв ефективності.

Дана процедура називається 100-кратною крос-валідацією – окремий випадок k-кратної крос-валідації. Це прогресивніший метод валідації, тому що ефективність, розрахована на лише одній тестовій вибірці, може дуже залежати від характеристик цієї вибірки[15]. Тому процедуру розроблено для того, щоб отримати ефективність вибірки, максимально наближену до істинної.

IV. Результати[16]

У таблиці 4 можна спостерігати ефективність моделей, що ґрунтуються на специфікації, зазначеній у попередньому розділі.

Моделі SVM є кращими в шести з восьми випадків. Слід зазначити, однак, що в деяких випадках результати моделей є приблизно рівними.

Таблиця 4. Коефіцієнт точності (GINI) моделей

Кластер	Сільське господарство			Виробництво			Торгівля			Інше		
Модель	LDA	Logit-модель	SVM	LDA	Logit-модель	SVM	LDA	Logit-модель	SVM	LDA	Logit-модель	SVM
Великі компанії	0.38	0.344	0.455	0.51	0.51	0.506	0.646	0.653	0.633	0.517	0.524	0.555
Малі компанії	0.458	0.497	0.512	0.472	0.508	0.535	0.498	0.497	0.545	0.233	0.228	0.294

Криві ROC, які відповідають медіанним значенням GINI, подано в додатку.

Перешкоди для практичної реалізації SVM

Схоже, що SVM у багатьох випадках є кращою за інші моделі. Вона ефективніша, ніж LDA і Logit-модель, хоча перетворення за WOE використовувалося, щоб сприяти ефективності останніх. Оскільки SVM використовує вхідну змінну як є, вона є гнучкішою, що являє собою бажану властивість.

На рисунку 7 бачимо іншу сприятливу ознаку SVM. Оскільки SVM дуже нелінійна, вона здатна приймати будь-яку функціональну залежність вхідної величини. Як бачимо, коли K21[17] збільшується в загальному діапазоні, оцінка знижується, і це відповідає інтуїтивно зрозумілій економічній сутності. Тим не менше, аномально високі значення коефіцієнта можуть свідчити про певні проблеми з фінансовою звітністю компанії, а це, в свою чергу, може бути ознакою проблем у самій компанії. Модель "ловить" це і збільшує оцінку (іншими словами, збільшує ймовірність дефолту). У певному сенсі модель може "зловити" навіть певні шаблони маніпуляцій зі звітністю.

Тим не менше ця сприятлива особливість іноді стає причиною заперечень з боку практиків. Розглянемо рисунок 8. Ми можемо спостерігати залежність оцінки від K12[18] для двох окремих компаній. Як бачимо, залежність є повністю

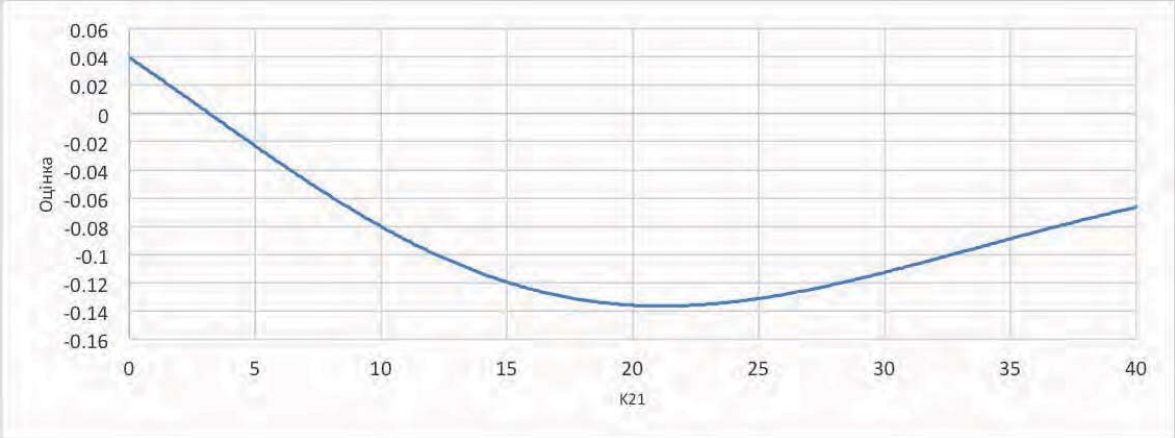
15 Kovahi R. (1995) для детальнішого опису цього методу.

16 Наведені результати не є остаточними, тому модель, яка буде представлена банківській системі, може дещо відрізнятися.

17 Кластер "Інші". Великі компанії.

18 Кластер "Інші". Великі компанії.

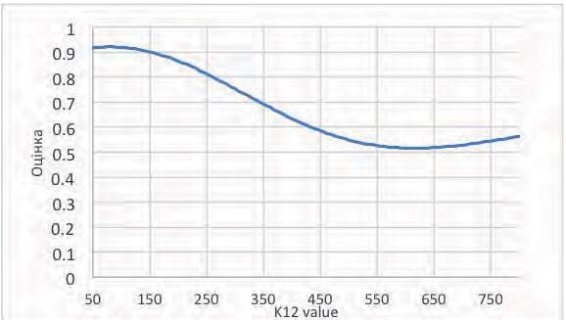
Рисунок 7. Залежність оцінки від значення K21 конкретної компанії



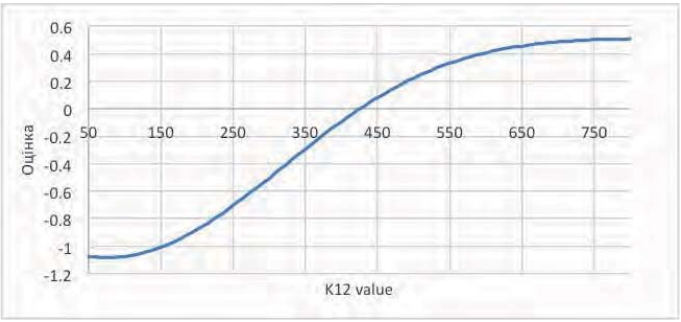
протилежною, що дивно для багатьох[19]. SVM "ловить" будь-який функціональний зв'язок і в підсумку може втратити монотонність результатів. Це відбувається тому, що оцінки SVM залежать не лише від конкретної змінної, а й одночасно від усіх інших змінних у рівнянні. Чи є це дійсно перепідгінкою, чи така залежність продиктована економічними причинами, не так легко з'ясувати. Результат, відображений на рисунку 8, як правило, є можливим, коли набір значень змінних істотно відрізняється, – скажімо, у випадку, коли одна з компаній має великі фінансові проблеми, що відображено в дуже поганих коефіцієнтах (з рисунку 8 b) ясно, що компанія має фінансові проблеми, оскільки її оцінка досить висока незалежно від того, яким є значення K12).

Рисунок 8. Ілюстрація проблем перепідгінки

a)



b)



Це не така велика проблема. У таблиці 5 представлено відсоток порушення монотонності для кожної моделі. Судячи з усього, в середньому близько 20 – 30% спостережень порушує монотонність результатів.

Таблиця 5. Порушення монотонності в SVM

Великі компанії			
Сільське господарство	Виробництво	Торгівля	Інші
32%	0%	29.84%	28.07%
Малі компанії			
Сільське господарство	Виробництво	Торгівля	Інші
16.11%	16.72%	28.43%	34.44%

19 Звертаємо увагу, що контр-інтуїтивні знаки не дозволялися у випадку LDA і Logit за умовчанням. Це може сприйматись як привілей для SVM. З іншого боку, LDA та Logit мають відому заздалегідь функціональну форму, що зменшує ризик перепідгінки. Тому не відомо, який вплив на результат мав цей "привілей", і чи було це привілеєм узагалі.

Обрана модель

На даному етапі дуже важливо впровадити практичну модель, яку можна було б легко пояснити і зрозуміти, тому впровадження SVM було призупинено на даний час.

Для впровадження було обрано Logit-модель, оскільки вона демонструє дещо кращі результати, ніж LDA. Крім того, її статистичні властивості більш сприятливі й вона поширеніша в банківській системі.

У таблиці 6 міститься найдетальніша інформація про отримані рівняння, а також – псевдо-R^2 тестову статистику. Усі знаки рівняння відповідають економічній сутності (нагадаємо, що всі знаки повинні бути позитивними, оскільки більше WOE завжди вказує на "здоровіше" фінансове співвідношення).

Здається, що кластер "Інші" для малих компаній демонструє відносно невисоку ефективність. У цьому немає нічого дивного, оскільки на практиці даний кластер охоплює велику кількість компаній, які є економічно залежними від інших підприємств. Для таких компаній погані фінансові коефіцієнти не обов'язково означають високу ймовірність дефолту, оскільки материнські компанії, швидше за все, підтримають їх. Водночас, якщо вони втрачають цю підтримку, то можуть збанкрутувати, навіть маючи добрі фінансові показники. З цієї причини модель не може надійно розрізняти компанії за платоспроможністю.

Таблиця 6. Рівняння Logit-моделі[20]

Великі компанії					Малі компанії				
Кластер	Змінна	Коефіцієнт	P-значення	Псевдо R^2	Кластер	Змінна	Коефіцієнт	P-значення	Псевдо R^2
Сільське господарство	K10	0.917	0.08	0.12	Сільське господарство	K7	0.613	0.05	0.19
	K11	0.564	0.27			K9	0.53	0.25	
	K24	1.11	0.01			K18	0.294	0.48	
	K25	1.084	0.12			K27	0.269	0.58	
	константа	1.875	0.0			K29	0.71	0.11	
						K30	0.524	0.31	
						константа	1.703	0.0	
Виробництво	K7	0.366	0.2	0.15	Виробництво	K1	0.623	0.02	0.12
	K29	0.358	0.2			K24	0.791	0.0	
	K20	0.599	0.0			K29	0.558	0.07	
	K24	0.476	0.01			константа	1.608	0.0	
	K30	0.688	0.0						
	константа	1.24	0.0						
Торгівля	K11	0.523	0.03	0.25	Торгівля	K1	0.35	0.18	0.14
	K14	0.909	0.0			K9	0.772	0.01	
	K15	0.754	0.01			K21	0.891	0.0	
	K21	0.98	0.0			K24	0.342	0.17	
	K23	0.732	0.01			K31	0.433	0.11	
	константа	2.072	0.0			константа	1.913	0.0	
Інше	K10	0.652	0.03	0.18	Інше	K5	0.308	0.56	0.05
	K22	0.954	0.02			K8	0.608	0.27	
	K27	0.669	0.16			K11	0.28	0.55	
	K29	0.83	0.05			K20	0.583	0.22	
	K30	1.058	0.01			K31	0.572	0.17	
	константа	1.544	0.0			константа	1.112	0.0	

20 Необхідно зазначити, що деякі змінні статистично незначущі за P-значенням. Однак, P-значення не були головним критерієм вибору моделі, а скоріше допоміжним. Тому, статистично незначущі змінні дозволялися в деяких випадках.

Представлений набір є оптимальним на даний момент. Більше того, модель оновлюватиметься та вдосконалюватиметься в міру того, як надходитиме нова інформація.

Кілька слів про подальші кроки

Michael Doumpos, Constantin Zopodunis (2009) запропонували спосіб, як зробити SVM економічно обґрунтованим шляхом запровадження так званих підказок у навчальний алгоритм. Підказки, по суті, є додатковими обмеженнями до проблеми оптимізації (4). Пропонуємо переформулювати її так, щоб вона використовувала трансформацію Кернела:

$$\min_{w,b} \frac{||w||}{2} + C \sum_{i=1}^{n} \varepsilon_i$$

$$s.t.\ y_i(K(x_i, X)u + b) \geq 1, \qquad i = 1, \dots, n$$

$$\varepsilon_i, w \geq 0, \qquad i = 1, \dots, n, \quad n \text{ – розмір вибірки.}$$

Ми хочемо, щоб залежність була монотонною. Іншими словами, хочемо, щоб

$$\left(K(x_i, X) - K(x_j, X)\right) u \geq 0, \tag{6}$$

де кожен елемент x_j більший, ніж відповідний елемент x_i. Це і є додаткове обмеження.

Формулою (6) передбачено, що якщо x збільшується, оцінка знижуватиметься. З метою впровадження цього обмеження ми повинні спочатку створити набір даних, на який орієнтуватиметься модель. Іншими словами, ми штучно створюємо вхідні вектори, даючи таким чином підказки алгоритму про те, яку модель очікуємо побачити. Крім того, оскільки нам потрібно, щоб усі змінні монотонно зменшувалися (як передбачено нерівністю (5)), ми повинні в разі необхідності перевернути рівняння для K таким чином, щоб їхня економічна сутність відповідала зазначеній умові[21].

V. Висновки

У статті продемонстровано високий потенціал SVM як моделі оцінки кредитного ризику. Із використанням даних українських компаній було доведено, що вона є ефективнішою, ніж класичні скорингові моделі. Однак SVM лише маржинально краща, тому вона не може вважатися безальтернативно кращим вибором. Це, радше, дуже добра та дієва альтернатива, але вибір повинен робити дослідник у кожному конкретному випадку.

Також нами обговорено деякі властивості SVM, у тому числі її складність і відсутність монотонності в результатах, запропоновано подальші кроки щодо поліпшення моделі й усунення цих властивостей. Зокрема, може бути розроблене навчання завдяки застосуванню підказок, що зробить SVM інтуїтивно зрозумілою економічно, крім того, це, вірогідно, знизить перепідгінку.

З урахуванням сказаного нині було прийнято Logit-модель. Виявилося, вона ефективніша, ніж LDA. Крім того, вона має привабливіші статистичні властивості, ніж LDA. У будь-якому випадку модель повинна переглядатися та оновлюватися регулярно з метою охоплення останньої динаміки в економіці. НБУ планує переглядати модель щороку.

21 Ця процедура повинна обійти проблему порушень монотонності результатів.

Література

- Altman E. I. (1968), Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1968.tb00843.x/pdf
- Venables W. N., Ripley B. D. (2002), Modern Applied Statistics with S, pp. 333-336.
- Pohar M., Blas M., and Turk S. (2004), Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study.
- Hosmer D. W., Lemeshow S. (2000), Applied Logistic Regression, pp.5-10.
- Boser B. E., Guyon I. M., Vladimir N. Vapnik (1995), A Training Algorithm for Optimal Margin Classifiers.
- Andrew Ng, Stanford University, CS229 Lecture notes, http://cs229.stanford.edu/notes/cs229-notes3.pdf
- Fawcett T. (2005), An Introduction to ROC Analysis.
- Doumpos M., Zopodunis C. (2009), Monotonic Support Vector Machines for Credit Risk Rating.
- Costeiu A., Negu F.(2013), Bridging the Banking Sector With the Real Economy a Financial Stability Perspective, ECB WORKING PAPER SERIES NO 1592, https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1592.pdf
- Pyle D., San Jose C. (2015), An Executive's Guide to Machine Learning, McKinsey Quarterly. http://www.mckinsey.com/insights/high_tech_telecoms_internet/an_executives_guide_to_machine_learning
- Hardle W. K., Moro R. A., Schafer D.(2009), Estimating Probabilities of Default With Support Vector Machines.
- Auria L., Moro R. (2007), Credit Risk Assessment Revisited Methodological Issues and Practical Implications", Working Group On Risk Assessment, pp.49-68.
- НБУ (2012), Постанова №23 "Про затвердження Положення про порядок формування та використання банками України резервів для відшкодування можливих втрат за активними банківськими операціями". http://zakon4.rada.gov.ua/laws/show/z0231-12/page5
- Kovahi R.(1995), A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection, http://ai.stanford.edu/~ronnyk/accEst.pdf

Додатки

Таблиця А. Змінні LDA та Logit. Детальна інформація

Великі компанії									
Галузь	Змінна	Діапазони	WOE	IV	Галузь	Змінна	Діапазони	WOE	IV
Сільське господарство	K10	<2.62	1.39	0.24	Торгівля	K11	<0.19	1.66	0.74
		<66.37	-0.05				<0.37	1.64	
		<146.26	-0.11				<0.62	0.49	
		>146.26	-0.65				<0.8	0.17	
	K11	<1.34	0.6	0.26			<1.9	-0.61	
		<2.19	0.5				>1.9	-0.73	
		<2.51	0.11			K14	<180	0.36	0.69
		<3.68	-0.29				<250	-0.96	
		>3.68	-0.56				>250	-2.35	
	K24	<0.47	0.36	0.29		K15	<16.17	0.29	0.28
		<1.35	-0.02				<38.98	-0.24	
		>1.35	-1.21				>38.98	-1.21	
	K25	<-0.52	0.93	0.12		K21	<0.67	-1.26	0.32
		<-0.03	0.11				<1.01	-0.07	
		<0.59	0.06				>1.01	0.35	
		<0.94	-0.11			K23	<1.38	-0.5	0.49
		>0.94	-0.65				<2.7	0.34	
Виробництво		<0	0		Інші		>2.7	1.34	
	K7	<-0.11	-0.64	0.41		K10	<54.47	0.65	0.46
		<-0.02	-0.63				<82.2	-0.05	
		<0	-0.3				<135.6	-0.64	
		<0.04	0.06				>135.6	-0.9	
		>0.04	1.1			K22	<0.06	-0.64	0.26
	K20	<-0.03	-0.96	0.42			<0.34	-0.48	
		<0.07	-0.67				>0.34	0.49	
		<0.18	-0.64			K27	<-0.34	-0.79	0.16
		<0.26	-0.13				<0.01	-0.21	
		<0.31	0.02				<0.14	-0.05	
		>0.31	0.7				<0.22	0.7	
	K23	<0.54	-0.59	0.43			>0.22	1.06	
		<1.11	-0.44			K29	<-0.34	-0.79	0.29
		<2.04	-0.16				<0.01	-0.21	
		<15	0.89				<0.14	-0.05	
		>15	1.13				<0.22	0.7	
	K24	<0.03	2.72	0.48			>0.22	1.06	
		<0.11	0.41			K30	<0.1	0.73	0.28
		<0.23	0.34				<0.14	-0.05	
		<0.31	0.02				<0.61	-0.28	
		<0.42	-0.16				>0.61	-0.79	
		<0.65	-0.2						
		>0.65	-0.9						
	K30	<0.49	0.38	0.26					
		<1.04	-0.57						
		>1.04	-0.88						

Малі компанії									
Галузь	Змінна	Діапазони	WOE	IV	Галузь	Змінна	Діапазони	WOE	IV
Сільське господарство	K7	<0.01	-1.3	0.73	Торгівля	K1	<-0.02	-1.24	0.28
		<0.05	-0.57				<0	0	
		<0.16	0.91				>0	0.26	
		<0.22	0.94			K9	<88.76	0.26	0.25
		>0.22	0.94				<162.79	-0.16	
	K9	<26.28	0.61	0.36			>162.79	-1.16	
		<88.18	-0.51			K21	<0.57	-1.41	0.44
		<191.6	-0.69				<0.9	-0.16	
		>191.6	-0.69				<1.03	-0.03	
	K18	<0	1.07	0.44			<1.42	0.16	
		<0.1	0.86				>1.42	0.61	
		<0.16	0.17			K24	<0.02	0.7	0.39
		<1.66	-0.06				<0.04	0.37	
		>1.66	-1.3				<0.06	0.35	
	K27	<1.01	0.57	0.28			<0.09	0.15	
		<1.67	0.09				<0.13	0	
		<4.47	-0.31				<0.25	-0.03	
Виробництво		>4.47	-0.69		Інші		<0.61	-0.43	
	K29	<0.14	-0.51	0.51			>0.61	-1.24	
		<0.21	-0.31			K31	<1.64	0.91	0.38
		<0.3	0.09				<6.92	0.26	
		<0.36	0.17				<40.71	0.07	
		<0.47	0.94				<100	-0.34	
		>0.47	1.67				<101.5	-0.65	
	K30	<0.11	0.5	0.29			>101.5	-1.04	
		<0.22	0.46			K5	<-0.01	-0.53	0.1
		<0.4	-0.31				<0.02	-0.03	
		<0.76	-0.31				>0.02	0.28	
		>0.76	-1.01			K8	<30.19	0.26	0.09
	K1	<-0.07	-1.43	0.35			<70.39	-0.12	
		<0.01	0.15				>70.39	-0.4	
		>0.01	0.28			K11	<0.84	0.4	0.14
	K11	<0.75	0.77	0.5			<1.24	0.37	
		<0.91	0.25				<2.22	-0.12	
		<1.16	0.03				<17.71	-0.28	
		<1.93	-0.73				>17.71	-0.64	
		>1.93	-0.88			K20	<-0.19	-0.64	0.11
	K24	<0.04	0.81	0.41			<0.67	-0.02	
		<0.25	-0.04				<0.77	0.14	
		<0.93	-0.73				>0.77	0.92	
		>0.93	-1.02			K31	<0.29	1.37	
	K29	<-0.33	-1.02	0.29			<185	0.28	
		<-0.11	-0.19				>185	-0.36	
		<0.2	-0.1						
		<0.54	0.5						
		>0.54	1.05						

Таблиця Б. Описова статистика по змінних

a) Великі компанії

Сільське господарство							Виробництво						
Змінна	всі		платоспроможні		неплатоспроможні		Змінна	всі		платоспроможні		неплатоспроможні	
	середнє	стнд. відхил.	середнє	стнд. відхил.	середнє	стнд. відхил.		середнє	стнд. відхил.	середнє	стнд. відхил.	середнє	стнд. відхил.
K8	210.3	157.4	219.6	155.6	148.2	158.8	K3	0.0	0.1	0.0	0.1	-0.1	0.2
K10	60.7	76.0	58.9	74.4	72.8	87.1	K7	0.0	0.1	0.0	0.1	0.0	0.1
K11	111.7	235.9	104.4	230.9	160.8	267.2	K12	293.4	291.5	256.9	258.3	421.3	358.7
K19	3.5	6.2	3.6	6.3	2.9	5.6	K16	21.5	38.8	16.2	32.9	40.1	50.6
K22	0.5	0.5	0.5	0.5	0.3	0.2	K20	0.3	0.3	0.4	0.3	0.2	0.3
K24	0.6	0.8	0.6	0.8	0.8	1.0	K23	3.6	7.8	4.4	8.4	1.0	4.8
K25	-3.8	19.8	-4.5	21.2	0.4	0.7	K24	0.5	0.8	0.4	0.8	0.8	0.9
							K25	-11.1	31.7	-7.6	26.8	-23.4	42.7
							K30	0.5	0.4	0.4	0.3	0.6	0.4

Торгівля							Інші						
Змінна	всі		платоспроможні		неплатоспроможні		Змінна	всі		платоспроможні		неплатоспроможні	
	середнє	стнд. відхил.	середнє	стнд. відхил.	середнє	стнд. відхил.		середнє	стнд. відхил.	середнє	стнд. відхил.	середнє	стнд. відхил.
K9	51.7	61.1	45.1	53.5	102.9	88.0	K10	87.0	114.0	75.3	105.5	142.3	136.4
K11	51.0	131.5	46.7	124.0	84.8	177.6	K12	291.6	334.9	266.7	330.4	409.5	335.4
K14	57.4	117.4	48.0	96.4	131.3	209.2	K21	1.7	1.7	1.8	1.7	1.4	1.6
K15	13.2	28.4	11.2	24.8	28.9	45.6	K22	0.6	0.6	0.7	0.6	0.4	0.4
K20	0.2	0.2	0.2	0.2	0.2	0.2	K27	3.8	4.5	3.6	4.3	5.0	5.3
K21	1.5	1.1	1.5	1.1	1.5	1.4	K29	0.0	0.3	0.0	0.3	-0.1	0.3
K23	3.4	8.9	3.8	9.3	0.5	4.2	K30	0.2	0.3	0.2	0.3	0.3	0.4
K24	0.4	0.8	0.3	0.7	0.7	1.0							
K25	-9.5	29.5	-9.1	29.5	-12.4	29.4							
K31	32.4	41.6	31.8	41.5	36.8	42.3							

b) Малі компанії

Сільське господарство							Виробництво						
Змінна	всі		платоспроможні		неплатоспроможні		Змінна	всі		платоспроможні		неплатоспроможні	
	середнє	стнд. відхил.	середнє	стнд. відхил.	середнє	стнд. відхил.		середнє	стнд. відхил.	середнє	стнд. відхил.	середнє	стнд. відхил.
K7	0.1	0.1	0.1	0.1	0.0	0.1	K1	0.0	0.1	0.0	0.1	0.0	0.2
K9	64.5	121.6	56.9	114.6	105.9	150.8	K10	138.5	184.6	126.2	171.6	201.1	231.8
K11	2.2	3.5	2.0	3.4	3.5	3.8	K24	0.5	1.4	0.4	1.1	1.1	2.1
K18	0.7	1.5	0.7	1.5	1.0	1.5	K27	3.4	8.6	3.4	8.1	3.4	10.6
K21	2.5	2.2	2.7	2.3	1.6	1.5	K29	0.1	0.3	0.1	0.3	0.0	0.3
K27	3.5	9.0	3.5	9.5	3.3	5.6							
K29	0.2	0.3	0.2	0.3	0.1	0.3							
K30	0.3	0.3	0.2	0.3	0.4	0.4							

Торгівля							Інші						
Змінна	всі		платоспроможні		неплатоспроможні		Змінна	всі		платоспроможні		неплатоспроможні	
	середнє	стнд. відхил.	середнє	стнд. відхил.	середнє	стнд. відхил.		середнє	стнд. відхил.	середнє	стнд. відхил.	середнє	стнд. відхил.
K1	0.0	0.1	0.0	0.1	0.0	0.2	K1	0.1	0.2	0.1	0.2	0.1	0.2
K8	40.2	112.0	35.3	104.0	73.6	152.8	K5	0.0	0.1	0.1	0.1	0.0	0.1
K9	66.4	108.0	59.7	96.6	111.4	159.6	K8	114.6	185.8	110.4	187.5	127.3	181.7
K13	0.6	2.0	0.5	1.8	1.4	3.0	K9	90.1	139.0	85.4	135.8	104.5	148.4
K18	1.7	2.8	1.6	2.6	2.4	3.4	K11	4.1	5.6	3.8	5.3	5.0	6.2
K20	0.3	0.3	0.3	0.3	0.1	0.3	K12	1.5	2.3	1.3	2.2	1.9	2.6
K21	1.7	1.5	1.7	1.4	1.5	1.8	K18	2.5	3.3	2.7	3.5	1.9	2.5
K24	0.4	1.1	0.3	1.0	0.8	1.6	K20	0.3	0.3	0.3	0.3	0.3	0.3
K31	33.3	43.4	30.0	41.6	55.9	48.1	K24	1.4	2.2	1.4	2.2	1.7	2.4
							K31	46.3	47.9	42.9	47.6	56.9	47.6

Таблиця В. Додаткові параметри SVM

Великі компанії			Малі компанії		
Галузь	Box Constraint	Масштаб	Галузь	Box Constraint	Масштаб
Сільське господарство	0.1	1	Сільське господарство	0.1	1
Виробництво	0.1	15	Виробництво	0.1	1
Торгівля	0.1	0.1	Торгівля	0.1	0.1
Інші	0.1	1	Інші	0.1	1

Таблиця Г. Матриця кореляцій для великих компаній

Сільське господарство							
	K8	K10	K11	K19	K22	K24	K25
K8	1.0000						
K10	0.2450	1.0000					
K11	0.1506	0.0407	1.0000				
K19	0.1612	0.0451	0.0838	1.0000			
K22	-0.0311	-0.1871	-0.1068	-0.0590	1.0000		
K24	0.0027	0.1029	0.0082	0.1488	-0.1039	1.0000	
K25	0.0846	0.0622	-0.0575	-0.1244	0.1072	-0.0660	1.0000

Виробництво									
	K3	K7	K12	K16	K20	K23	K24	K25	K30
K3	1.0000								
K7	0.7794	1.0000							
K12	-0.2502	-0.1247	1.0000						
K16	-0.1419	-0.0930	0.3819	1.0000					
K20	0.3650	0.3926	-0.2136	-0.2140	1.0000				
K23	0.4552	0.6035	-0.0603	-0.0793	0.2991	1.0000			
K24	-0.5263	-0.2653	0.5519	0.1663	-0.3617	-0.1186	1.0000		
K25	0.4348	0.3773	-0.1787	-0.1097	0.6066	0.1800	-0.3656	1.0000	
K30	-0.1712	-0.0695	0.8099	0.2797	-0.1123	-0.0400	0.3646	-0.1077	1.0000

Торгівля										
	K9	K11	K14	K15	K20	K21	K23	K24	K25	K31
K9	1.0000									
K11	0.1171	1.0000								
K14	0.3319	0.0826	1.0000							
K15	0.2503	0.1739	0.0472	1.0000						
K20	-0.1539	-0.0255	0.1807	-0.1379	1.0000					
K21	0.0049	0.0115	0.2624	-0.1027	0.4935	1.0000				
K23	-0.0370	-0.0596	-0.0636	-0.0825	0.1722	0.0956	1.0000			
K24	0.3188	0.3458	0.1697	0.3091	-0.1684	-0.0019	-0.0875	1.0000		
K25	0.0219	-0.0080	0.0771	-0.0032	0.4647	0.1999	0.0530	-0.2249	1.0000	
K31	0.1867	-0.1340	0.1099	0.1902	-0.2200	-0.0574	-0.4476	0.1869	-0.0506	1.0000

Інші							
	K10	K12	K21	K22	K27	K29	K30
K10	1.0000						
K12	0.2776	1.0000					
K21	-0.0864	0.0890	1.0000				
K22	-0.0793	0.0674	0.6799	1.0000			
K27	0.2237	0.2719	0.0353	0.0081	1.0000		
K29	-0.1183	0.0849	0.7299	0.6195	-0.0203	1.0000	
K30	0.3073	0.4851	0.0271	0.0152	0.1898	0.1238	1.0000

Таблиця Д. Матриця кореляцій для малих компаній

Сільське господарство

	K7	K9	K11	K18	K21	K27	K29	K30
K7	1.0000							
K9	-0.2669	1.0000						
K11	-0.3357	0.6992	1.0000					
K18	-0.2583	0.2997	0.2879	1.0000				
K21	0.2622	-0.0669	-0.0518	-0.2467	1.0000			
K27	-0.2559	0.3884	0.4331	0.3724	0.0185	1.0000		
K29	0.2374	-0.0024	-0.0786	-0.2980	0.7209	-0.0086	1.0000	
K30	-0.1707	0.5054	0.3005	0.0927	0.0293	-0.0697	0.0757	1.0000

Виробництво

	K1	K10	K11	K24	K27	K29
K1	1.0000					
K10	-0.0233	1.0000				
K11	-0.2497	0.4984	1.0000			
K24	-0.2153	0.4074	0.8497	1.0000		
K27	-0.0094	0.0451	0.2809	0.3079	1.0000	
K29	0.2109	-0.3764	-0.1267	-0.1141	0.0225	1.0000

Торгівля

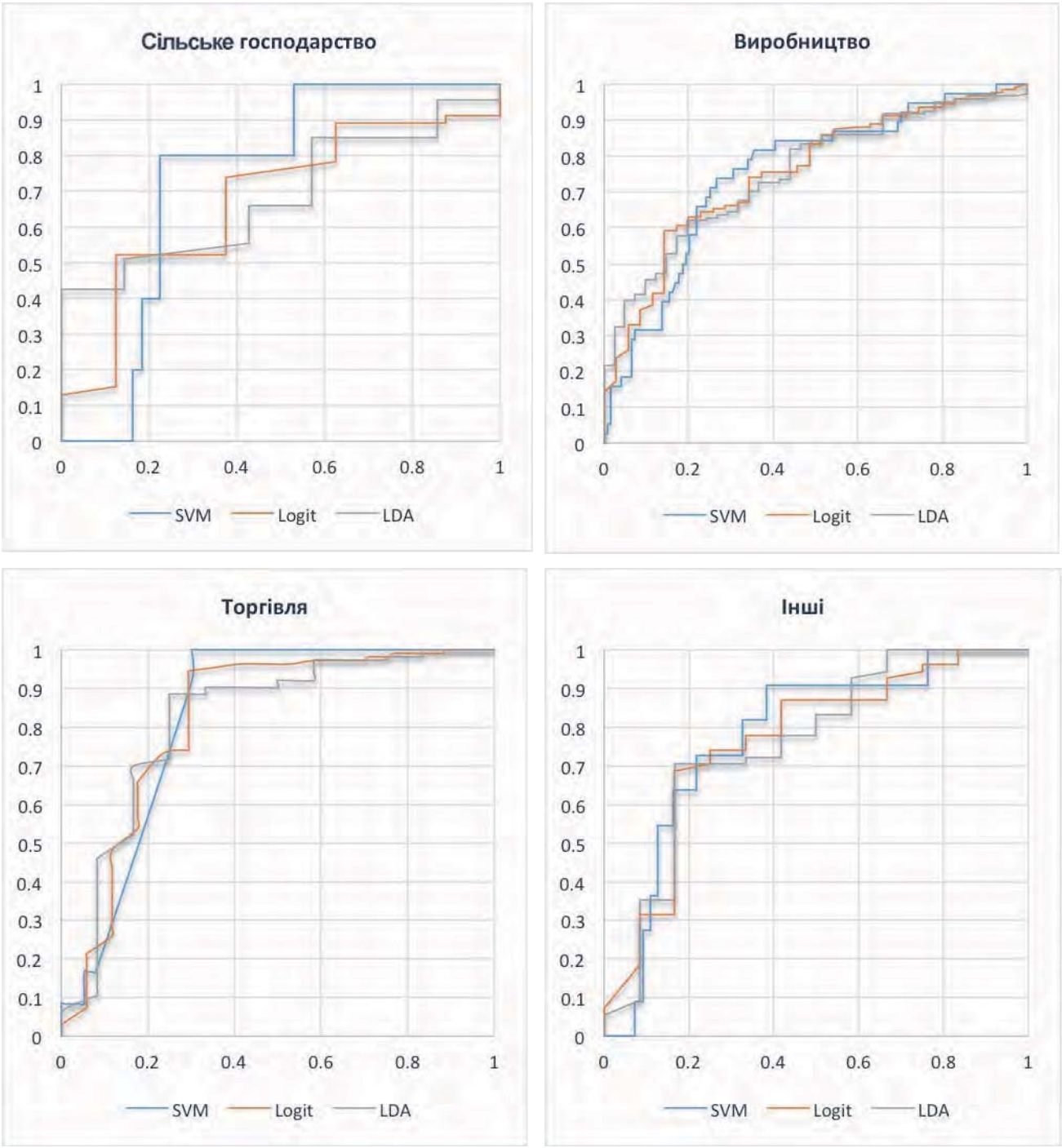
	K1	K8	K9	K13	K18	K20	K21	K24	K31
K1	1.0000								
K8	-0.0403	1.0000							
K9	-0.1662	0.3918	1.0000						
K13	-0.3612	0.6052	0.5175	1.0000					
K18	0.0134	-0.0443	0.0997	-0.0300	1.0000				
K20	0.1962	0.0495	-0.1447	-0.0451	-0.4546	1.0000			
K21	0.1062	-0.0765	-0.0074	-0.1109	-0.1281	0.5316	1.0000		
K24	-0.2454	0.4573	0.5603	0.7227	0.2029	-0.1239	-0.0198	1.0000	
K31	-0.3241	0.1016	0.2027	0.2507	0.3897	-0.4279	-0.1553	0.3522	1.0000

Інші

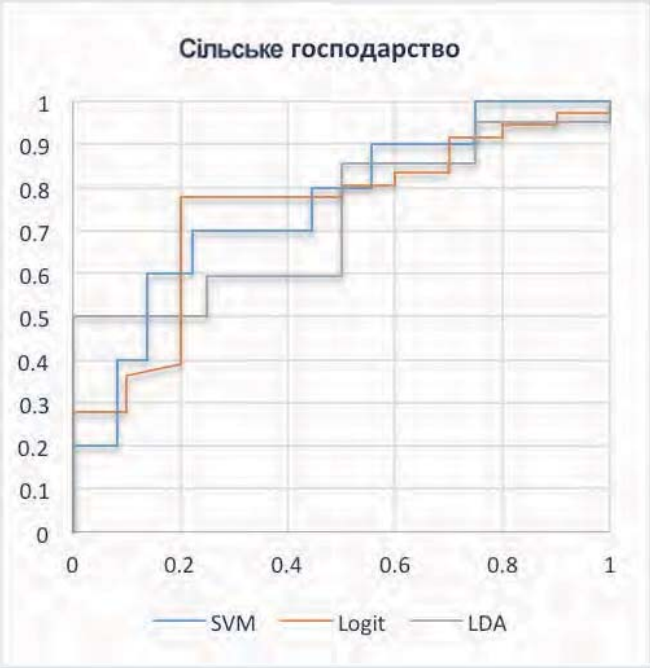
	K1	K5	K8	K9	K11	K12	K18	K20	K24	K31
K1	1.0000									
K5	0.2968	1.0000								
K8	-0.1225	0.2013	1.0000							
K9	-0.2893	0.1582	0.3602	1.0000						
K11	-0.1231	0.1103	0.5286	0.5499	1.0000					
K12	-0.1933	0.1436	0.5958	0.6148	0.8649	1.0000				
K18	-0.0750	-0.1189	0.1097	0.0360	0.1965	0.1764	1.0000			
K20	0.2381	0.2409	-0.1872	-0.1495	-0.2661	-0.3130	-0.6378	1.0000		
K24	-0.0987	0.0315	0.5119	0.4735	0.8577	0.7859	0.4468	-0.4885	1.0000	
K31	-0.2844	-0.3901	0.2526	0.2945	0.4826	0.4088	0.3796	-0.4483	0.5397	1.0000

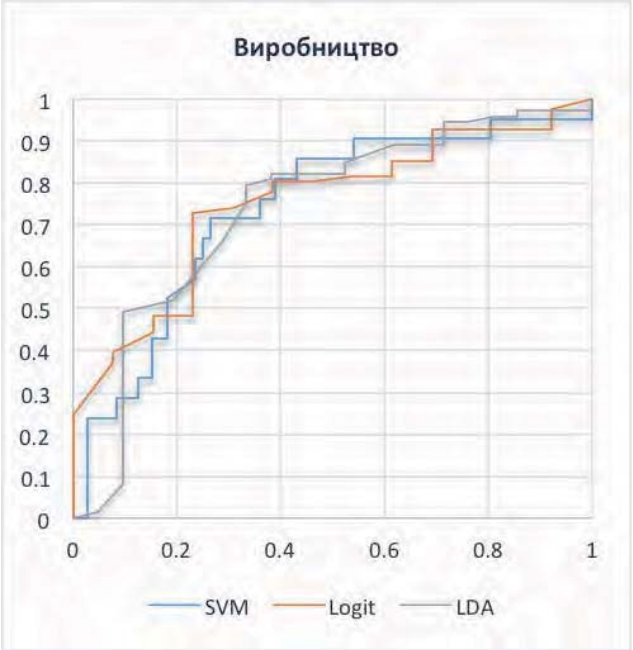
Рисунок А. ROC криві для порівнювальних моделей

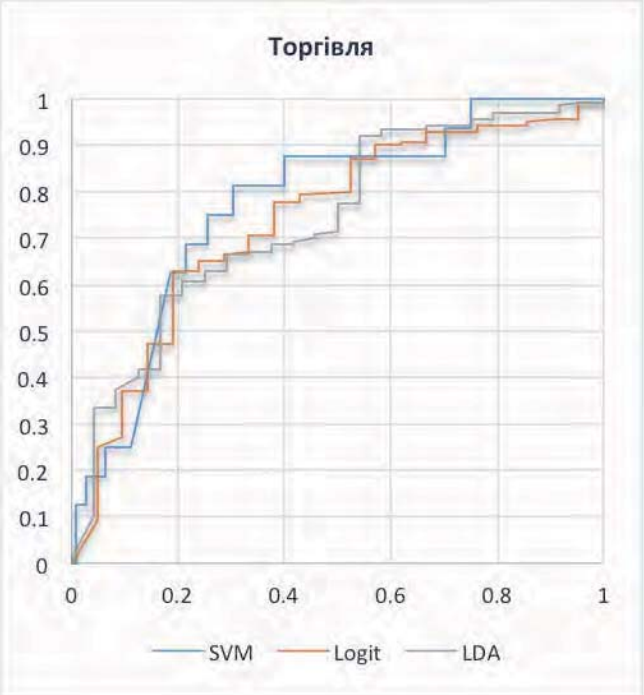
а) Великі

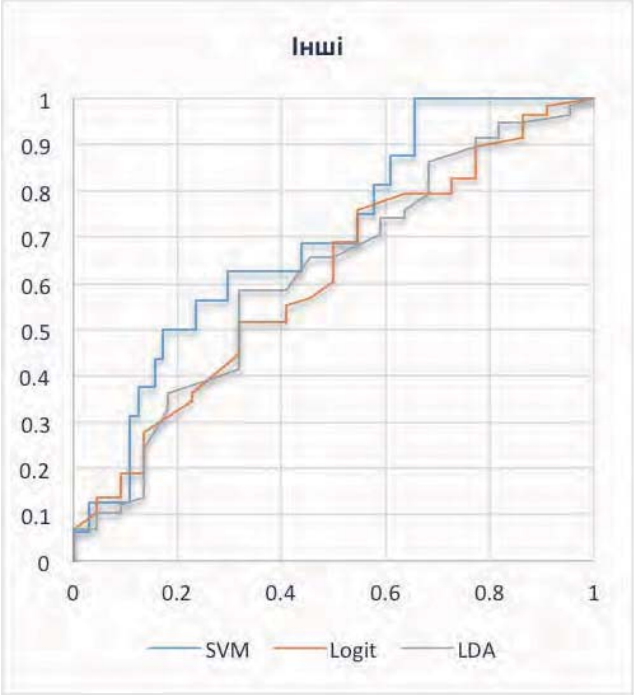


b) Малі


Сільське господарство
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
SVM
Logit
LDA


Виробництво
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
SVM
Logit
LDA


Торгівля
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
SVM
Logit
LDA


Інші
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
SVM
Logit
LDA