

ДОСЛІДЖЕННЯ ЕФЕКТУ ПЕРЕНАВЧАННЯ НЕЙРОННИХ МЕРЕЖ НА ПРИКЛАДІ ЗАДАЧІ АПЛІКАЦІЙНОГО СКОРИНГУ

Г. І. Великоіваненко

Кандидат фізико-математичних наук, професор,
професор кафедри економіко-математичного моделювання
Державний вищий навчальний заклад «Київський національний
економічний університет імені Вадима Гетьмана»
проспект Перемоги, 54/1, м. Київ, 03680, Україна
ivanenkog@list.ru

В. В. Корчинський

Бакалавр з економічної кібернетики
Державний вищий навчальний заклад «Київський національний
економічний університет імені Вадима Гетьмана»
проспект Перемоги, 54/1, м. Київ, 03680, Україна
vladislav.korchynskiy@gmail.com

В. В. Чернишова

Бакалавр з економічної кібернетики
Державний вищий навчальний заклад «Київський національний
економічний університет імені Вадима Гетьмана»
проспект Перемоги, 54/1, м. Київ, 03680, Україна
vikachernyshova@i.ua

У статті досліджено проблему перенавчання нейронних мереж. Розкрито теоретичне підґрунтя виникнення цього явища та висвітлені негативні наслідки його прояву. Проведено експериментальне дослідження ефекту перенавчання на прикладі задачі моделювання кредитних ризиків фізичних осіб. З цією метою побудовано низку логістичних регресій і нейромереж перцептронного типу різних конфігурацій, перевірка адекватності яких здійснювалась як на навчальній, так і на тестовій вибірках. Результати порівняння показників точності класифікації позичальників за критерієм Джині для обох вибірок дозволили отримати деякі нові узагальнення щодо сутності ефекту перенавчання та рекомендації щодо його уникнення. Практичною цінністю проведеного дослідження є отримані нові знання щодо способів позбавлення скорингових моделей від ефекту перенавчання, що дозволить забезпечити їх стійкість і стабільність результатів моделювання кредитних ризиків у прикладній діяльності банківських установ, що сприятиме стабільності фінансової системи в цілому.

Ключові слова: *перенавчання, нейронна мережа, аплікаційний скоринг, адекватність моделі.*

ИССЛЕДОВАНИЕ ЭФФЕКТА ПЕРЕОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ НА ПРИМЕРЕ ЗАДАЧИ АППЛИКАЦИОННОГО СКОРИНГА

Г. И. Великоиваненко

Кандидат физико-математических наук, профессор,
профессор кафедры экономико-математического моделирования
Государственное высшее учебное заведение «Киевский национальный
экономический университет имени Вадима Гетьмана»
проспект Победы, 54/1, г. Киев, 03680, Украина
ivanenkog@list.ru

В. В. Корчинский

Бакалавр по экономической кибернетике
Государственное высшее учебное заведение «Киевский национальный
экономический университет имени Вадима Гетьмана»
проспект Победы, 54/1, г. Киев, 03680, Украина
vladislav.korchynskiy@gmail.com

В. В. Чернышева

Бакалавр по экономической кибернетике
Государственное высшее учебное заведение «Киевский национальный
экономический университет имени Вадима Гетьмана»
проспект Победы, 54/1, г. Киев, 03680, Украина
vikachernyshova@i.ua

В статье исследуется проблема переобучения нейронных сетей. Раскрыты теоретические основы возникновения этого эффекта и освещены негативные последствия его проявления. Проведено экспериментальное исследование эффекта переобучения на примере задачи моделирования кредитных рисков физических лиц. С этой целью построен ряд логистических регрессий и нейросетей перцептронного типа различных конфигураций, проверка адекватности которых осуществлялась как на учебной, так и на тестовой выборках. Результаты сравнения показателей точности классификации заемщиков по критерию Джини для обеих выборок дали возможность получить некоторые новые обобщения относительно сущности эффекта переобучения и рекомендации по его предотвращению. Практической ценностью проведенного исследования являются полученные новые знания о способах избавления скоринговых моделей от эффекта переобучения, что позволит обеспечить их устойчивость и стабильность результатов моделирования кредитных рисков в прикладной деятельности банковских учреждений, что будет способствовать стабильности финансовой системы в целом.

Ключевые слова: *переобучение, нейронная сеть, аппликационный скоринг, адекватность модели.*

STUDY OF THE NEURAL NETWORKS OVERFITTING EFFECT ON THE EXAMPLE OF THE PROBLEM OF APPLICATION SCORING

Halyna Velykoivanenko

PhD (Physics and Mathematical Sciences), Professor,
Professor, Department of Economic and Mathematic Modeling
State Higher Educational Establishment «Kyiv National Economic University
named after Vadym Hetman»
54/1 Peremogy Avenue, Kyiv, 03680, Ukraine
ivanenkog@list.ru

Vladislav Korchynskiy

Bachelor's Degree in Economic Cybernetics
State Higher Educational Establishment «Kyiv National Economic University
named after Vadym Hetman»
54/1 Peremogy Avenue, Kyiv, 03680, Ukraine
vladislav.korchynskiy@gmail.com

Vika Chernyshova

Bachelor's Degree in Economic Cybernetics
State Higher Educational Establishment «Kyiv National Economic University
named after Vadym Hetman»
54/1 Peremogy Avenue, Kyiv, 03680, Ukraine
vikachernyshova@i.ua

The problem of neural networks overfitting is studied in the article. There is revealed the theoretical foundations of this effect occurrence and highlighted the negative consequences of its arising. It's carried out an experimental research of the overfitting effect by the example of the problem of modelling of individuals credit risk. To this purpose was built a number of logistic regressions and neural networks of perceptron type of different configurations, which adequacy was analysed on the training and test samples. The results of comparing the accuracy of borrowers classification on the basis of Gini criterion for both samples made it possible to get some new generalizations about the nature of the overfitting effect and recommendations for its prevention. The practical value of the study is the obtained new knowledge on how to avoid of the overfitting effect in scoring models that will ensure stability of credit risk modelling results in the banking institutions activity and will contribute to the sustainability of the financial system at whole.

Keywords: *overfitting, neural network, application scoring, adequacy of the model.*

JEL Classification: C45, D81, E51

Постановка проблеми

Основною метою діяльності комерційних банків, як і будь-якої прибуткової організації, є підвищення вартості власних активів та отримання прибутку в результаті реалізації певних товарів та/чи послуг. Особливістю діяльності кредитних установ на ринку банківських послуг є дуже тісний взаємозв'язок ризиків, що приймаються ними, та величиною отримуваних доходів. Наприклад, чим більша кількість клієнтів, що обслуговує банк, тим вища прибутковість від кредитно-інвестиційних операцій, але й інтегральний кредитний ризик також зростає.

Оскільки головними завданнями комерційних банків є збільшення прибутку та нарощення кредитних портфельів, то з плином часу відповідно збільшується і кількість прострочених позик. Одні кредити погашаються, інші пролонгуються, а деякі перетворюються на проблемну заборгованість. При активації негативного економічного «циклону» (наприклад, криза 2008 року) кількість проблемних боргів у портфелях комерційних банків збільшується у геометричній прогресії, сягаючи 9 % за кредитами юридичних осіб і 20 % за позиками фізичних осіб.

Кредитний ризик є найбільш значущим з усіх банківських ризиків, що визначається Національним банком України в нормативному документі [1] як «наявний або потенційний ризик для надходжень і капіталу, який виникає через неспроможність сторони, що взяла на себе зобов'язання, виконати умови будь-якої фінансової угоди із банком або в інший спосіб виконати взяті на себе зобов'язання». Його значущість зумовлена тим, що кредитні операції є основним напрямком діяльності банків. Це підтверджується статистичними даними [2]: якщо на 2013 рік частка доходів від кредитування в сумі загальних доходів по банківській системі складала 76,9 %, то на I квартал 2016 року цей показник складає 92,5 %. І хоча частка з року в рік може коливатися, за останні роки вона стабільно перевищує 2/3 загального доходу по банківській системі. Інші дані НБУ [3] показують, що загальна сума кредитів, наданих юридичним і фізичним особам в цілому, зросла з 815 327 млн грн у 2013 році до 948 599 млн грн станом на 01.05.2016. Таким чином, кредитний ризик, у результаті його реалізації, може нести за собою найбільші збитки і найбільш негативні наслідки порівняно з іншими ризиками банківської діяльності (операційними, валютними, ринковими тощо).

Погіршує ситуацію у сфері кредитування й нестабільна ситуація, що склалася в банківському секторі. Загалом, на поточний момент стан банківської системи можна охарактеризувати як критичний. Про це свідчать деякі негативні тенденції, зокрема зростання частки простроченої заборгованості у загальній сумі кредитів з 7,7 % у 2014 році до 23,5 % у I кварталі 2016 року [3]. Така тенденція, в поєднанні зі зростаючою кількістю наданих кредитів і часткою доходів від них, є вкрай негативною і свідчить про зростаючу неплатоспроможність позичальників.

З метою недопущення зростання проблемної кредиторської заборгованості та подальшого виникнення проблем у самої кредитної установи виникає потреба в проведенні процедур оцінки надійності позичальника на різних етапах кредитного циклу, і в першу чергу на етапі аналізу кредитної заявки (аплікаційний скоринг). Суттєвого підвищення ефективності процесу оцінки кредитних ризиків, збільшення швидкості та якості обслуговування клієнтів, а також зменшення вартості операційних послуг можна досягти за рахунок впровадження систем підтримки прийняття рішень для управління кредитною діяльністю комерційного банку.

Проведення аплікаційного скорингу позичальника є першим етапом реалізації кредитних відносин, адже слугує основою для прийняття рішення про надання кредиту. Адекватна діагностика фінансової спроможності позичальника та прогноз можливих напрямків її змін у подальшому дозволяють звести до мінімуму ризик неповернення коштів через виникнення складного фінансового становища позичальника або з інших обставин. Провівши інтерпретацію результатів оцінки кредитоспроможності із застосуванням сучасних економіко-математичних методів і моделей, відповідальна особа здатна прийняти раціональне та обґрунтоване рішення стосовно надання або ненадання кредиту потенційному позичальнику.

Визначення ризику невиконання кредитних зобов'язань здійснюється по аналогії з раніше наданими позиками, на основі яких відшукується функціональна залежність між показником надійності позичальника (класом надійних чи ненадійних боржників) та низкою кількісних і якісних характеристик цього позичальника та кредитного забезпечення.

Проблеми кредитного скорингу широко висвітлені у науковій літературі, і на сьогодні представлені дослідниками з багатьох країн світу, серед яких варто згадати праці вчених Р. Андерсона [4], В. Вітлінського [5], А. Камінського [6—7], Н. Сіддікі [8], О. Сорокіна [9]. У зазначених працях пропонується вирішення таких завдань, як розробка скорингових систем, впровадження їх до бізнес-процесів кредитних установ, вибір методів оцінювання кредитного ризику, включення часового параметру до моделей кредитного скорингу тощо.

Однак ефективність представлених у науковій літературі скорингових систем є досить невисокого рівня, що обумовлює доцільність розробки нових моделей аналізу кредитного ризику, в тому числі на підґрунті інструментарію штучних нейронних мереж, що здатен до виявлення більш складних нелінійних закономірностей у великих об'ємах інформації, порівняно з класичними методами побудови скорингових моделей. Однак у процесі пошуку найефективнішої конструкції нейронних мереж за рахунок складних внутрішніх взаємозв'язків вони виявляють схильність до прояву ефекту перенавчання, надто близько підлаштовуючись до навчальних прикладів, зменшуючи при цьому здатність до узагальнення.

Тож **метою** даної статті є проведення дослідження з побудови моделі аплікаційного скорингу, здатної до найефективнішого виявлення закономірностей у кредитних історіях позичальників та їх узагальнення для адекватного аналізу нових кредитних заяв.

Для вирішення поставленої мети необхідно вирішити **завдання**: здійснити пошук найадекватнішої структури нейронної мережі перцептронного типу для розв'язання задачі оцінювання кредитних ризиків позичальника, забезпечити виявлення та усунення ефекту перенавчання нейронної мережі, провести дослідження різних варіантів попередньої обробки вхідних даних з метою підвищення адекватності моделі, дослідити вплив обсягу навчальної вибірки на здатність моделей до узагальнення закономірностей на тестових даних, провести порівняльне дослідження ефективності нейронних мереж різних конфігурацій з класичним інструментарієм побудови скорингових моделей.

Аналіз сутності ефекту перенавчання, принципи його виявлення та запобігання

Одним із найбільших ускладнень у застосуванні нейронних мереж є можливість виникнення явища перенавчання, або надто близького підлаштування моделі до реальних даних. Негативні наслідки цього ефекту продемонструємо на прикладі апроксимації багаточленом $y = a_n x^n + \dots + a_2 x^2 + a_1 x + a_0$ множини дискретних значень деякої змінної y . Так, якщо задано якийсь набір даних, то можна поставити за мету підібрати для нього апроксимуючу функцію і в такий спосіб одержати математичну модель, придатну для відтворення цих навчальних прикладів. Чим вищий ступінь багаточлена, що реалізує апроксимуючу функцію, тим складнішою є його форма і тим точніше він здатен описувати множину точок, що встановлюють залежність y від x .

Якщо кількість параметрів моделі a_0, a_1, \dots, a_n є меншою за обсяг навчальної вибірки, то модель намагається найточніше відтворити заданий набір даних, але у загальному випадку будуть деякі розбіжності та, відповідно, похибка моделювання буде відмінною від нуля. У граничному випадку, коли кількість параметрів моделі дорівнюватиме кількості елементів навчальної вибірки, модель буде здатною точно відтворити усі навчальні приклади (це справедливо для будь-якого виду апроксимуючої функції). Зі збільшенням кількості параметрів моделі відносно обсягу навчальної вибірки похибка моделювання залишатиметься рівною нулеві, проте поза значеннями навчальних прикладів модель починає вести себе непередбачувано — як наслідок недовизначеності модель володітиме нескінченною кількістю можливих комбінацій власних параметрів, за яких похибка моделювання дорівнюватиме нулю.

Продемонструємо це на прикладі апроксимації багаточленом двох точок на площині (x, y) : $\{(1, 3); (2, 2)\}$. Так, коли у даній функції лише один параметр (наприклад, $y = a_1 x$), вона матиме вид прямої, що проходить між заданими двома точками, виходячи з початку координат (як можна бачити з рис. 1 для рівняння $y = 1,4 \cdot x$). При цьому помилка апроксимації у загальному випадку буде суттєвою. Якщо від багаточлена залишимо параметр a_0 , то апроксимуюча функція прийме вигляд горизонтальної прямої $y = 2,5$, що пройде посередині між двома заданими точками.

Якщо ж скористатись уже двома параметрами, утворивши, наприклад, лінійну функцію $y = a_1 x + a_0$, то вона зможе точно

пройти через будь-які дві довільно обрані точки (при цьому рішення, за якого апроксимація буде точною, є єдиним). Для заданих двох точок буде отримано розв'язок $y = -x + 4$, що можна бачити з рис. 1.

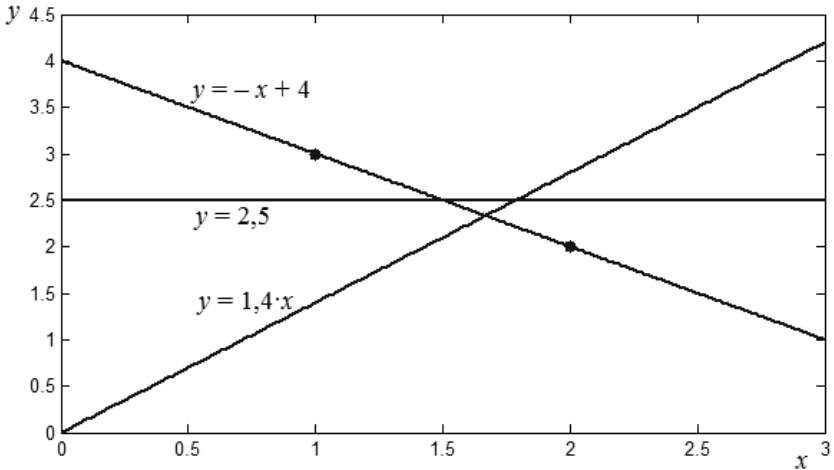


Рис. 1. Апроксимація вихідної змінної у лінійними функціями з одним та двома параметрами

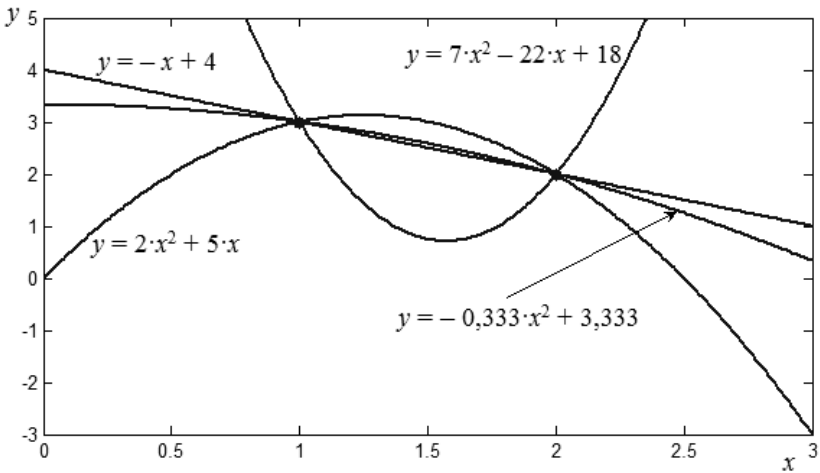


Рис. 2. Апроксимація вихідної змінної у багаточленом з двома та трьома параметрами

Якщо до лінійної функції додати третій параметр, утворивши, наприклад, параболу $y = a_2x^2 + a_1x + a_0$, то можливих рішень уже буде безліч, як можна бачити з рис. 2. Одне з таких рішень — нульове значення третього параметра (a_2) при розрахованих раніше значеннях двох перших параметрів (a_1, a_0): $y = -x + 4$. Можна провести функцію точно через дві встановлені точки за умови рівності нулю параметра a_1 : $y = -0,333 \cdot x^2 + 3,333$ або параметра a_0 : $y = 2 \cdot x^2 + 5 \cdot x$. Якщо ж скористатись трьома параметрами, то можна отримати безліч розв'язків, наприклад, зображену на рис. 2 параболу $y = 7 \cdot x^2 - 22 \cdot x + 18$. При цьому в цій функції можна збільшити третій параметр, відповідно зменшивши значення першого параметра, і рішення також буде точним. Тобто, стає можливим здійснювати узгоджену зміну цієї або іншої пари параметрів будь-яким чином, зберігаючи при цьому точність апроксимації двох заданих точок.

У результаті побудови таких функцій ми точно описуємо два встановлені приклади, але навряд чи отримані рівняння мають хоч якесь відношення до реальної залежності між вхідною x та вихідною y змінними — адже значення таких функцій уже при $x = 3$ коливаються в широкому діапазоні значень y від дуже низьких від'ємних до надвеликих позитивних.

Подібна ситуація буде повторюватись і з будь-яким іншим обсягом навчальної вибірки та кількістю параметрів апроксимуючої функції. Якщо продовжити ряд аналогій з багаточленом, то продемонструємо моделювання трьох навчальних прикладів (x, y) : $\{(1, 3); (2, 2); (3, 2,5)\}$ параболою третього ступеня, що має чотири параметри $y = a_3x^3 + a_2x^2 + a_1x + a_0$.

На рис. 3 зображено єдину параболу другого ступеня $y = 0,75 \cdot x^2 - 3,25 \cdot x + 5,5$, що точно проходить через три задані точки, а також два приклади парабол третього ступеня $y = 0,75 \cdot x^3 - 3,75 \cdot x^2 + 5 \cdot x + 1$ та $y = -1,25 \cdot x^3 + 8,25 \cdot x^2 - 17 \cdot x + 13$ із безлічі можливих варіантів. Як видно з рис. 3, подальший розвиток усіх цих кривих мало що має спільного з динамікою вихідного часового ряду, який коливається в межах від 2 до 3. При цьому ми можемо отримати безліч рішень для рівнянь поліномів від третього ступеня і вище, які точно відтворюють три початкові приклади, але вже на наступному кроці прогноз буде коливатись від мінус нескінченності до плюс нескінченності.

Отже, апроксимуючим функціям із багатьма параметрами властива здатність набувати надто заплутаної форми, яка може не мати жодного відношення до форми вихідної функціональної залежності. У результаті такі моделі навряд чи зможуть надати адекватний прогноз щодо нових даних, які не містились у навчальній вибірці, хоча заздалегідь задані приклади відтворювалися без помилок. Отже, замість того щоб узагальнити відомі приклади, модель запам'ятовує їх. Ця ситуація і демонструє те, що називається явищем перенавчання.

Оскільки реальні дані, як правило, містять шумову компоненту, то не можна однозначно стверджувати, що найбільш адекватніша модель визначається кривою, яка проходить точно через задані позиції. Як можна бачити з рис. 3, найбільш відповідною реальним даним з усіх побудованих функцій є рівняння звичайної прямої $y = -0,25 \cdot x + 3$, хоч вона і не точно відтворює навчальні приклади.

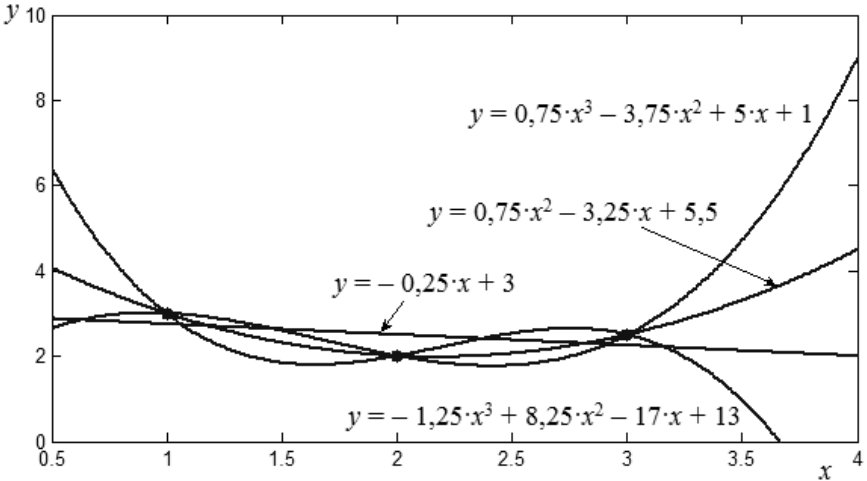


Рис. 3. Апроксимація вихідної змінної у поліномами другого та третього ступеня

Зазначимо, що подібна ситуація з перенавчанням буде проявлятися у математичних моделях будь-якого типу, налаштування яких реалізується із застосуванням процедури навчання з учителем, у тому числі й у нейронних мережах типу перцептрон. Ко-

ректним способом уникнення явища перенавчання є формування множини навчальних прикладів, обсяг якої перевищує кількість параметрів моделі, або ж зменшення кількості параметрів персептрона (кількості вагових коефіцієнтів зв'язків між усіма нейронами усіх сусідніх шарів, параметрів функцій активації та параметрів зміщення суматора усіх нейронів). При цьому необхідно стежити за точністю моделювання, оскільки мережа з незначною кількістю зв'язків може виявитись недостатньо гнучкою, аби змоделювати відтворювану залежність.

Зауважимо, що в процесі оптимізації параметрів моделі важливо забезпечити не стільки мінімізацію похибки навчання, скільки мінімізацію помилки узагальнення для можливих у майбутньому випадків. Саме така математична модель буде характеризуватись максимальною передбачувальною здатністю. Тому важливо після налаштування моделі на навчальній вибірці перевіряти її адекватність на тестовій. У результаті застосування такої процедури можна забезпечити вибір найстійкішої моделі, здатної до узагальнення вихідних функціональних залежностей.

Але орієнтація тільки на помилку моделювання (навіть із врахуванням прикладів із тестової вибірки) може призвести до описаних вище проявів ефекту перенавчання та підлаштування моделі як під навчальну, так вже і під тестову вибірки. Тому у спеціалізованій літературі з нейромережевого моделювання набув поширення дещо відмінний погляд на розв'язання проблеми запобігання ефекту перенавчання нейронної мережі [10, 11]. Згідно з цим підходом формується нейронна мережа довільної структури (кількість її параметрів може навіть перевищувати обсяг навчальної вибірки) та здійснюється її оптимізація із застосуванням одного з алгоритмів навчання з учителем (наприклад, описаного вище алгоритму зворотного поширення помилки). З метою виявлення ефекту перенавчання використовується механізм контрольної перевірки, згідно з яким частина відомих спостережень резервуються як контрольні приклади і не використовуються як у процесі оптимізації параметрів нейронної мережі, так і при послідовному її тестуванні.

Спочатку помилка мережі на навчальній і контрольній множині буде однаковою. Якщо вони істотно відрізняються, то, ймовірно, розбиття спостережень на ці дві множини не забезпечило їх однорідність. У процесі налаштування мережі похибка відтворення навчальної вибірки убуває та, відповідно, зменшу-

ється помилка моделювання на контрольній множині. Якщо ж контрольна похибка перестала убувати або почала зростати за подальшого зменшення похибки навчання, це вказує на те, що нейронна мережа почала занадто близько слідувати вихідним даним з навчальної вибірки через її перенавчання. Тому оптимізаційний процес варто зупинити.

Але жодних гарантій немає, що мережа, підігнана під навчальну та контрольну вибірку, не буде вести себе на нових прикладах аналогічно до парабол на рис. 3. Адже проведення численних експериментів з перевіркою адекватності моделі на контрольній вибірці веде до того, що вона починає відігравати ключову роль у виборі структури нейронної мережі, тобто стає частиною процесу навчання. Тим самим її роль як незалежного критерію адекватності моделі знижується, оскільки за великої кількості експериментів виникає ризик перенавчання нейронної мережі також і на контрольній множині. Тому додатково застосовується тестова вибірка, але лише раз, щоб пересвідчитись у стійкості результатів моделювання та відсутності ефекту перенавчання ще й на контрольній множині.

Однак ще раз зауважимо, що помилка моделювання у недовизначеній системі в глобальному мінімумі дорівнюватиме нулю, причому цей мінімум може бути отриманий за нескінченної кількості різних комбінацій параметрів моделі. В описаному вище підході [10, 11] відшукується такий стан мережі, який точно відтворює вихідну статистику, але при цьому велика кількість параметрів нейронної мережі залишається незадіяною — мережа будується громіздка, але недонавчається. Тобто, відбувається щось на зразок побудови парабол на рис. 3, коли деякі параметри моделі прирівнюються до нуля (з рисунку можна бачити, наскільки подальший розвиток таких функцій далекий від діапазону змін модельованого показника).

Отже, більш коректним способом боротьби з ефектом перенавчання є побудова моделі з кількістю параметрів, суттєво меншою за обсяг навчальної вибірки, та її оптимізація з послідувочою перевіркою адекватності на тестовій вибірці. Але ефект перенавчання (занадто близького слідування навчальним даним і нездатності до ефективного узагальнення залежностей для нових прикладів) може проявлятися і за умови незначної кількості параметрів моделі. Тож вважаємо доцільним при виборі найадекватнішої структури моделі досліджувати її ефективність на тестовій

вибірці, кожен раз здійснюючи максимально глибоку оптимізацію. Такий підхід дозволить крім уникнення ефекту перенавчання ще й обрати найпростішу структуру моделі, що задовольнятиме одному з базових принципів економіко-математичного моделювання, коли за приблизно однакових показників адекватності (точності відтворення прикладів з навчальної вибірки та прогнозування на тестовій, а також стійкості результатів моделювання на різних вибірках) для подальшого використання необхідно брати ту модель, яка має простішу структуру.

Формування множини вхідних даних скорингової моделі

Дослідження з побудови моделей аплікаційного скорингу позичальників-фізичних осіб ґрунтувалось на інформації щодо соціо-демографічних і фінансових характеристик позичальників на момент отримання кредиту (без можливості ідентифікації самої особи). Всього в дослідженні було проаналізовано більше 80 характеристик позичальника як узятих із його кредитної заявки та офіційних документів, так і розрахованих на основі даних із баз бюро кредитних історій, що містять інформацію щодо виконання позичальником попередніх і діючих кредитних угод, звернень до кредитно-фінансових установ тощо. За кожним позичальником враховувалась уся відома щодо нього кредитна історія за максимально тривалий проміжок часу (у наявності була інформація більш ніж за двадцятирічний період щодо десятків мільйонів кредитів).

Зауважимо, що кожен позичальник характеризується як кількісними показниками (кількість діючих угод, відсоток відмов банків щодо запитів у наданні кредиту, заробітна плата позичальника, кількість утриманців і т. д.), так і якісними (стать, освіта, посада, регіон проживання тощо). Для отримання можливості врахування у скорингу позичальника якісних показників, виникає потреба їх переведення у кількісну форму. Одним з варіантів вирішення цієї проблеми є застосування підходу на основі фіктивних змінних, суть якого полягає у виділенні в якісній змінній певних однорідних груп, яким надаються номерні позначення: 0, 1, 2... Наприклад, такий показник, як освіта, може приймати наступні значення: «незакінчена середня» — 1; «середня» — 2; «середньо-технічна» — 3 і т. д.

Однак за такого кодування може виникати некоректне впорядкування значень, адже автоматично встановлюється, що значення «середньо-технічна» є гіршим за «вища», оскільки кількісна оцінка однозначно фіксує перевищення однієї величини над іншою. Якщо подібним чином здійснити нумерацію регіонів країни, то такий підхід взагалі стає беззмістовним (кількісне значення одного регіону в кінці алфавіту може у десятки разів перевищувати номер регіону на початку списку, що взагалі нічого не означає з огляду на оцінку ризиковості його мешканців). Тож подібний підхід до переведення якісних змінних у числа можна вважати некоректним. Тому для кількісної оцінки якісних характеристик було вирішено скористатись показником *WOE* (*Weight Of Evidence*).

Для соціально-демографічних показників розрахунок *WOE* дає змогу заміни значень, що є кодами підгруп певного якісного показника, на такі кількісні величини, що містять інформацію стосовно кредитної поведінки позичальників у кожній підгрупі. Така інформація ґрунтується на обчисленні відсотків надійних і ненадійних угод за кожною підгрупою показника та дає можливість адекватно врахувати в моделі якісні характеристики позичальників.

Так, за кожною категорією (підгрупою) обчислюється частка надійних і ненадійних угод у цій підгрупі відносно загальної кількості надійних і ненадійних угод, відповідно, та розраховується значення *WOE* за формулою:

$$WOE_i = \ln\left(\frac{d_i^1}{d_i^2}\right), \quad i = \overline{1, k}, \quad (1)$$

де d_i^1 та d_i^2 – відносні частоти відповідно ненадійних і надійних угод за i -ою підгрупою категоризованої змінної; k – кількість підгруп змінної.

Причому, *WOE* у спеціалізованій літературі [9] рекомендується розраховувати не тільки для якісних показників, але й для кількісних, попередньо здійснивши розбиття усієї множини значень відповідного показника на інтервали. І вже для кожного такого інтервалу проводиться розрахунок відповідного *WOE*.

У принципі, такий підхід має логічне обґрунтування. Адже не можна однозначно стверджувати, що, скажімо, заробітна платня

у 20 тис. грн вказує на значно менший кредитний ризик позичальника порівняно з тим, хто зазначив у кредитній заявці зарплатню у 2 тис. грн. По-перше, для отримання кредиту позичальник в умовах української дійсності може отримати практично будь-яку довідку по заробітній платі, тож високі її показники не гарантують, що вона є дійсно такою. По-друге, якщо людина працює в комерційній організації та прагне збільшити рівень своїх доходів, часто офіційна зарплата знижується, щоб зменшити податкові відрахування. Таким чином, категорія позичальників із зарплатою у 2 тис. грн може виявитись значно надійнішою, ніж позичальники з надвисокими доходами. І специфіку поведінки кожної з таких груп дозволить виявити розрахунок показників *WOE*.

Крім того, близько двох третин кредитних історій, що використовувались у дослідженні, взагалі не містили інформації щодо рівня заробітної плати. Тож, якщо цей показник використовувати у моделі, що будується на вихідних даних, то більша частина статистичних спостережень не зможе бути залучена до навчальної вибірки. Крім того, якщо для нового позичальника рівень заробітної плати не вказаний, то така модель теж не зможе бути застосована в оцінюванні його кредитного ризику.

Проте, показник *WOE* розраховується як для різних значень якісного показника чи інтервалів кількісного показника, так і для окремої категорії, що характеризується відсутністю даних. Таким чином, *WOE* дозволяє зробити модель універсальною, тобто такою, яку можна використовувати за будь-якого наповнення даних щодо характеристик позичальника.

Тож, на основі якісних і кількісних даних щодо клієнта та кредиту, застосовуючи процедуру категоризації з розрахунком показників *WOE* (1), отримуємо нормалізовані кількісні характеристики, які можна використовувати для проведення аплікаційного скорингу кредитної заявки позичальника.

Моделювання кредитного ризику позичальників-фізичних осіб

Для побудови моделей кредитного скорингу було сформовано навчальну і тестову вибірку за однаковими принципами обробки даних. Враховуючи, що в Україні у 2014 році відбулись кардинальні зміни як у політичному, так і в економічному житті, то

відповідно зазнала суттєвих змін і специфіка ринку кредитування. Так, у 2014 році суттєво збільшилась кількість дефолтів за зобов'язаннями, що пов'язано як з різким падінням курсу національної валюти, платоспроможності населення, так і з великою кількістю внутрішніх переселенців у зв'язку з військовими діями на сході країни. Але вже у 2015 році ситуація у політичному, економічному житті та на фінансовому ринку країни стабілізувалась, звісно, суттєво відрізняючись як від 2014 року, так і попередніх років.

Отже, навчання та тестування скорингових моделей проводилось на статистиці за нових економічних реалій – на кредитах, виданих протягом 2015 — початку 2016 років (кількох десятків тисяч, за якими вже була інформація щодо факту виконання або невиконання кредитних зобов'язань). Причому, щоб модель не підлаштувалась під конкретний клас позичальника, угоди, за якими зобов'язання були виконані, та які виявились дефолтними, були взяті до навчальної вибірки в однаковій кількості. Якщо ж формувати навчальну вибірку пропорційно класам кредитів у реальності, то модель може налаштуватись навіть так, щоб усі кредити визнавати надійними (оскільки дефолтних кредитів приблизно у сім разів менше, то при оптимізації моделі ними може бути знехтувано, оскільки при незначному зменшенні альфа-помилки суттєво більше починає зростати бета-помилка, знижуючи загальну точність класифікації). При цьому тестова вибірка була сформована з решти кредитів (які не увійшли до навчальної вибірки) без корегування пропорцій представництва позик різних класів.

Оскільки і навчальна, і тестова вибірки склались з дійсно великої кількості спостережень, то базовою гіпотезою при побудові скорингових моделей було те, що для врахування значного різноманіття характеристик кредиту й позичальника та виявлення складних нелінійних залежностей результуючого показника (класу кредиту, що приймає значення 0 для кредитів, за якими зобов'язання були виконані, та 1 — для дефолтних) від вхідних факторів необхідно конструювати досить масштабні економіко-математичні моделі з великою кількістю внутрішніх параметрів. Найвідповіднішим цим вимогам інструментарієм є нейронні мережі перцептронного типу, де регулювати кількість

параметрів можна за рахунок зміни числа прихованих шарів та їх нейронів.

Однак, у результаті проведення численних модельних експериментів чітко проявився ефект перенавчання, коли зі збільшенням кількості параметрів у нейронній мережі (причому неважливо, за рахунок збільшення кількості нейронів у одному прихованому шарі або ж одночасного збільшення кількості шарів) похибка відтворення вихідних даних на навчальній вибірці суттєво зменшувалась, але при цьому помилка класифікації на тестовій вибірці зростала.

Результати тестування логістичних регресій і нейронних мереж різної конфігурації (персептронів з одним прихованим шаром із різною кількістю нейронів) за коефіцієнтом Джині подано у табл. 1.

Зауважимо, що логістична регресія (яка є найпоширенішим у банківській практиці способом реалізації скорингової моделі) за своєю сутністю є повним аналогом одного єдиного нейрона із сигмоїдною функцією активації. Але навіть така проста конструкція моделі виявляється дуже ефективною та у багатьох випадках демонструє вищу адекватність за нейронні мережі, що видно з табл. 1. Тобто, зі збільшенням параметрів моделі зростала здатність до відтворення залежностей у навчальному наборі даних, але таким чином відбувалось звичайне підлаштування моделі під цю статистику зі зменшенням здатності до узагальнення (що видно по падінню коефіцієнту Джині при перевірці на тестовій вибірці).

Обсяг навчальної вибірки не здійснював якогось помітного впливу на точність моделі при її тестуванні. Тобто, побудовані моделі ефективно виявляли закономірності в поведінці позичальників уже за досить невеликих обсягів навчальної вибірки. Цікавим виявився факт суттєвого зниження адекватності всіх побудованих моделей при застосуванні прийнятого в практиці конструювання скорингових карт біннінгу показників із переведенням їх значень у WOE. Причому для цього спеціально був розроблений алгоритм категоризації, але дослідження у цьому напрямку ще триватимуть.

Таблиця 1

**РЕЗУЛЬТАТИ ОЦІНЮВАННЯ АДЕКВАТНОСТІ ЛОГІСТИЧНИХ РЕГРЕСІЙ
ТА НЕЙРОННИХ МЕРЕЖ РІЗНОЇ КОНФІГУРАЦІЇ ЗА КОЕФІЦІЄНТОМ ДЖИНИ**

Вихідні дані	Початкові		Переведені у WOE	
	Навчальна	Тестова	Навчальна	Тестова
Обсяг навчальної вибірки: 1000 значень				
Логістична регресія	0,793	0,820	0,659	0,562
2 нейрони	0,799	0,840	0,764	0,506
3 нейронів	0,798	0,828	0,798	0,500
5 нейронів	0,843	0,767	0,876	0,459
10 нейронів	0,908	0,648	0,986	0,345
Обсяг навчальної вибірки: 2000 значень				
Логістична регресія	0,740	0,830	0,620	0,594
2 нейрони	0,775	0,873	0,667	0,573
3 нейронів	0,773	0,855	0,725	0,521
5 нейронів	0,761	0,680	0,760	0,549
10 нейронів	0,843	0,593	0,908	0,397
Обсяг навчальної вибірки: 4000 значень				
Логістична регресія	0,798	0,813	0,612	0,604
2 нейрони	0,825	0,867	0,654	0,580
3 нейронів	0,844	0,828	0,698	0,553
5 нейронів	0,832	0,755	0,778	0,534
10 нейронів	0,858	0,562	0,846	0,419

Закінчення табл. 1

Вихідні дані	Початкові		Переведені у WOE	
	Навчальна	Тестова	Навчальна	Тестова
Обсяг навчальної вибірки: 6000 значень				
Логістична регресія	0,789	0,830	0,613	0,606
2 нейрони	0,798	0,808	0,625	0,601
3 нейронів	0,794	0,606	0,678	0,579
5 нейронів	0,855	0,503	0,739	0,549
10 нейронів	0,873	0,476	0,818	0,472
Обсяг навчальної вибірки: 8000 значень				
Логістична регресія	0,819	0,835	0,617	0,579
2 нейрони	0,829	0,787	0,650	0,602
3 нейронів	0,865	0,602	0,674	0,581
5 нейронів	0,889	0,488	0,691	0,554
10 нейронів	0,893	0,424	0,742	0,533

Найкращий результат моделювання кредитних ризиків продемонстрували перцептрони найпростішої конфігурації (з двома трьома нейронами на єдиному прихованому шарі). Результати перевірки таких моделей на тестовій вибірці сягали коефіцієнту Джині 0,87, що є надзвичайно високим показником адекватності скорингової моделі. Причому ці результати відрізнялись стійкістю за різних умов налаштування та тестування моделей.

Література

1. Методичні вказівки з інспектування банків «Система оцінювання ризиків» [Електронний ресурс] / Затверджено Постановою Правління Національного банку України від 15.03.2004 № 104. — Режим доступу : <http://zakon2.rada.gov.ua/laws/show/v0104500-04>.
2. Доходи та витрати банків України [Електронний ресурс] // Національний банк України. — 2013-2016. — Режим доступу: https://bank.gov.ua/control/uk/publish/article?art_id=34661442&cat_id=34798593.
3. Основні показники діяльності банків України [Електронний ресурс] // Національний банк України. — 2016. — 4 березня. — Режим доступу : https://bank.gov.ua/control/uk/publish/article?art_id=34661442&cat_id=34798593.
4. *Anderson R. A.* The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management / R. A. Anderson. — Oxford: Pan Books, 2007. — 792 p.
5. *Вітлінський В. В.* Кредитний ризик комерційного банку [Текст]: навч. посіб. / В.В. Вітлінський, О. В. Пернарівський, Я. С. Наконечний, Г. І. Великоіваненко. — К. : Знання, 2000. — 251 с.
6. *Камінський А. Б.* Скорингові технології в кредитному ризик-менеджменті / А. Б. Камінський, К. К. Писанець // Бізнес-інформ. — 2012. — № 4. — С. 197—201.
7. *Камінський А. Б.* Структура та інструментарій ризик-менеджменту у споживчому кредитуванні / А. Б. Камінський, К. К. Писанець // Теоретичні та прикладні питання економіки: Збірник наукових праць (за заг. ред. проф. Єханурова Ю. І., Шегди А. В.). — К.: ВПЦ «Київський університет», 2012. — Вип. 27. — Т. 2. — С. 169—175.
8. *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring / N. Siddiqi. — New Jersey: John Wiley & Sons, 2006. — 196 p.
9. *Сорокин А. С.* Построение скоринговых карт с использованием модели логистической регрессии. [Электронный ресурс] / А. С. Сорокин // Интернет-журнал «Науковедение». — 2014. — Вып. 2. — С. 1—29. — Режим доступу : <http://naukovedenie.ru/PDF/180EVN214.pdf>.

10. *Круглов В. В.* Искусственные нейронные сети. Теория и практика / В. В. Круглов, В. В. Борисов. — М. : Горячая линия — Телеком, 2001. — 382 с.
11. *Медведев В. С.* Нейронные сети. MATLAB 6 / В. С. Медведев, В. Г. Потёмкин. — М. : ДИАЛОГ-МИФИ, 2002. — 496 с.

References

1. National Bank of Ukraine. (2004, March 15). *Metodychni vказivky z inspektuvannya bankiv «Systema otsynuyvannya ryzykiv»*. Retrieved from <http://zakon2.rada.gov.ua/laws/show/v0104500-04> [in Ukrainian].
2. National Bank of Ukraine. (2013-2016). *Dohody ta vytraty bankiv Ukrayiny*. Retrieved from https://bank.gov.ua/control/uk/publish/article?art_id=34661442&cat_id=34798593 [in Ukrainian].
3. National Bank of Ukraine. (2016, March 3). *Osnovni pokaznyky diyal'nosti bankiv Ukrayiny*. Retrieved from https://bank.gov.ua/control/uk/publish/article?art_id=34661442&cat_id=34798593 [in Ukrainian].
4. Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management*. Oxford, UK: Pan Books.
5. Vitlins'kyj, V. V., Pernarivs'kyj, O. V., Nakonechnyj, Ya. S., & Velykoivanenko, H. I. (2000). *Kredytnyj ryzyk komertsijnoho banku*. Kyiv, Ukraine: Znannia [in Ukrainian].
6. Kamins'kyi, A. B., & Pysanets', K. K. (2012). Skoryngovi tekhnolohii v kredytnomu ryzyk-menedzhmenti. *Biznes Inform (Inform business)*, 4, 197—201 [in Ukrainian].
7. Kamins'kyi, A. B., & Pysanets', K. K. (2012). Struktura ta instrumentarij ryzyk-menedgmentu v spozhyvchomu kredytnuvanni. *Teoretychni ta prykladni pytannia ekonomiky (Theoretical and Applied Economic Issues)*, 27(2), 169—175 [in Ukrainian].
8. Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. New Jersey, USA: John Wiley and Sons.
9. Sorokin, A. S. (2014). Postroenie skorinhovih kart s ispol'zovaniem modeli logisticheskoy regressii. *Internet-zhurnal «Naukovedenie» (Internet Journal "Science of Science")*, 2. Retrieved from <http://naukovedenie.ru/PDF/180EVN214.pdf> [in Russian].
10. Kруглов, В. В., & Борисов, В. В. (2001). *Iskusstvennyie neyronnyie seti. Teoriya i praktika*. Moscow, Russia: Goryachaya liniya — Telekom [in Russian].
11. Медведев, В. С., & Потёмкин, В. Г. (2002). *Neuronnyie seti. MATLAB 6*. Moscow, Russia: DIALOG-MIFI [in Russian].

Стаття надійшла до редакції 27.05.2016