Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 364 · 2018

Jan B. Gajda

Państwowa Wyższa Szkoła Zawodowa w Suwałkach
Wydział Humanistyczno-Ekonomiczny
jan.b.gajda@wp.pl

KWALIFIKACJA WNIOSKÓW KREDYTOWYCH – PORÓWNANIE REGRESJI ORAZ SIECI NEURONOWEJ

Streszczenie: W pracy badamy decyzje udzielenia bądź odmowy udzielenia kredytu konsumpcyjnego przeznaczonego na zakup samochodu osobowego na przykładzie próby ok. 200 klientów pewnego banku. Analiza sprowadza się do porównania instrumentów wspomagających podejmowanie decyzji – funkcji regresji oraz sieci neuronowej. Banki bądź przyznają kredyty (zmienna wyjściowa przybiera wartość 1), bądź ich odmawiają (na wyjściu pojawia się 0) na podstawie informacji o kliencie (tworzącej zbiór zmiennych wejściowych) zawartej w wypełnianym przez niego kwestionariuszu. Ze względu na obowiązek zachowania tajemnicy banki strzegą danych klientów, stąd rzadkość badań wykorzystujących informacje pochodzące z autentycznych wniosków kredytowych.

Sieci neuronowe mogą okazać się przydatne do wstępnego rozpoznania istnienia bądź braku powiązań pomiędzy zmiennymi wejściowymi a wyjściowymi. Jeśli dopasowanie sieci jest wyraźnie lepsze od dopasowania liniowego równania regresji – sugeruje to nieliniowy charakter związku pomiędzy tymi zmiennymi. W naszym przykładzie użyteczność włączenia logarytmu zmiennej *staż* zdaje się wskazywać na przewagę sieci neuronowej. Jednakże – w odróżnieniu od regresji – sieci neuronowe nie dają szans rozróżnienia zmiennych wejściowych mających istotny wpływ na zmienne wyjściowe od niemających takiego wpływu. Pozostawia to pole do dyskusji na temat podobieństw i różnic w zakresach stosowalności sieci oraz modeli ekonometrycznych.

Słowa kluczowe: regresja, sieci neuronowe, klasyfikacja wniosków kredytowych.

JEL Classification: C02, C45, C52, C58, G17.

Wprowadzenie

W sytuacji, gdy chcemy się dowiedzieć, czy i jakie związki zachodzą między zmienną zależną a grupą zmiennych niezależnych, wykorzystujemy dwa główne (choć nie jedyne) instrumenty wstępnej eksploracji danych – regresję

i sieć neuronową. Pierwsza jest dobrze opisana w literaturze i ma ponad 150-letnią tradycję zastosowań, druga jest trudniejsza do zrozumienia, ale za to dostępna w pakietach takich jak Statistica [STATISTICA Neural Networks™ PL, 2001] i technicznie łatwa do zastosowania. W pracy próbujemy zweryfikować ich zalety i wady jako instrumentów data mining.

Głównym celem analizy jest porównanie zalet i wad instrumentów podejmowania decyzji – funkcji regresji oraz sieci neuronowej. Są to w jakimś sensie konkurujące z sobą instrumenty badawcze. Różnią się charakterem, sztywnością założeń – zwłaszcza dotyczących postaci funkcyjnej regresji wpływających na interpretowalność, a także sensowność rezultatów badań – oraz ilością dostarczanej informacji. Sieci neuronowe wyróżniają się elastycznością, umiejętnością identyfikacji związków zachodzących między zmiennymi objaśniającymi a zmienną objaśnianą, nawet jeśli mają one nieliniowy, choć nieznany badaczowi charakter. Za te zalety sieci płacimy utratą możliwości oceny wpływu poszczególnych zmiennych objaśniających na zmienną objaśnianą, jak również jedynie pośrednią oceną jakości dopasowania wyników sieci do danych empirycznych (przez mechanizm podziału próby na zbiory: uczący, testujący i ewentualnie weryfikujący).

Materiałem empirycznym, na którym dokonujemy naszych porównań, są decyzje udzielenia bądź odmowy udzielenia kredytu konsumpcyjnego przeznaczonego na zakup samochodu osobowego na przykładzie próby ok. 200 klientów pewnego banku. Problem ten, ciekawy sam w sobie ze względu na nieczęstą możliwość korzystania z takiego materiału empirycznego, jest w naszych rozwiązaniach materiałem ilustracyjnym, pozwalającym na skonkretyzowanie wskazówek dla badaczy. Banki przyznają kredyty na podstawie informacji o kliencie zawartej w wypełnianym przez niego kwestionariuszu. Zazdrośnie strzegą one danych klientów, motywując to obowiązkiem zachowania tajemnicy bankowej. Stąd rzadkość badań wykorzystujących informacje z autentycznych wniosków o kredyt bankowy.

1. Sieć neuronowa a równanie regresji

1.1. Regresja

Zastosowanie równań regresji w modelowaniu i klasyfikacji zjawisk gospodarczych ma długą tradycję, toteż dla oszczędności miejsca ograniczymy się

jedynie do zarysowania podstaw analizy regresji[1]. Dla ustalenia uwagi rozważmy najprostsze równanie regresji liniowej:

$$y_m = b_1 x_{1m} + b_2 x_{2m} + \cdots + b_K x_{Km} + u_m, m = 1, \ldots, M$$

gdzie:
y_m – m-ta obserwacja na zmiennej objaśnianej;
x_{km} – m-ta obserwacja na k-tej zmiennej objaśniającej;
b_k – parametr związany z k-tą zmienną objaśniającą;
u_m – m-te zakłócenie (po oszacowaniu parametrów – reszta równania);
M – liczebność próby.
Parametry b_k równania regresji szacujemy najczęściej Metodą Najmniejszych Kwadratów, dobierając je tak, aby zminimalizować sumę kwadratów reszt.

Za zalety regresji uznamy: możliwość odróżnienia zmiennych statystycznie istotnie wpływających na zmienną zależną od takich, których wpływu nie da się uznać za statystycznie istotny (innymi słowy takich, których wpływ nie daje się precyzyjnie oszacować), możliwość rozpoznania kierunku wpływu każdej zmiennej na zmienną objaśnianą (przy założeniu *ceteris paribus*, tj. po uwzględnieniu i odliczeniu wpływu innych zmiennych objaśniających), możliwość rozróżnienia dokładności, z jaką oszacowano poszczególne parametry, od dokładności, z jaką całe równanie objaśnia zachowanie się zmiennej zależnej. Wreszcie mamy tu możliwość rozpoznania groźby, jaką dla dokładności szacunków niesie zjawisko współliniowości, tj. niedostatecznej ilości informacji zawartej w zmiennych objaśniających, ilości niewystarczającej dla dostatecznie dokładnego oszacowania parametrów regresji, a nawet – przy ścisłej współliniowości – w ogóle niedającej szansy na oszacowanie parametrów bez dodatkowych ograniczeń *a priori*. Ceną za to jest, że regresja nie tylko wymaga wyróżnienia w próbie statystycznej zmiennej endogenicznej (objaśnianej) oraz zespołu zmiennych niezależnych (objaśniających), ale również wskazania postaci regresji (liniowa, jeśli nieliniowa – to jaka jest to postać: potęgowa, półlogarytmiczna, logistyczna

[1] Bliższy opis Czytelnik znajdzie w każdym podręczniku ekonometrii, opis zwięzły w: [Gajda, 2004]. Zauważmy, że binarny charakter zmiennej objaśnianej (1 – przyznać kredyt, 0 – odmówić) sprawia, iż w badaniach ekonometrycznych rekomenduje się stosowanie modeli dostosowanych do ograniczonego zakresu zmienności zmiennej objaśnianej, jak np. model logitowy (szerszy przegląd modeli i metod estymacji por. Gruszczyński, red., 2010). Celem stworzenia możliwości porównania regresji, traktowanej jako instrument służący nie budowie modelu ekonometrycznego, ale wstępnej eksploracji danych z sieciami neuronowymi traktowanymi jako instrument podporządkowany temu samemu celowi, postawiliśmy zarówno przed regresją, jak i sieciami neuronowymi to samo zadanie opisania wariantów decyzji w postaci zer i jedynek. Oznacza to m.in., że mogą one generować wartości zmiennej endogenicznej wykraczające poza przedział [0,1].

itd.), a także jakim transformacjom należy poddać występujące w niej zmienne. Podjęcie wadliwej decyzji w tej kwestii może sprawić, że wnioski wyciągnięte z regresji będą nieprecyzyjne lub wręcz wadliwe.

1.2. Sieć neuronowa

Sztuczne sieci neuronowe (SSN) stanowią jeden z najnowszych instrumentów prognozowania oraz klasyfikacji danych. Niedostatki wiedzy o postaci funkcyjnej mechanizmów wiążących badane zmienne są zastępowane uczeniem skomplikowanej, wieloelementowej i często wielopoziomowej struktury – sieci neuronowej[2].

Koncepcje SSN wywodzą się z analizy procesów przetwarzania informacji przez neurony istot żywych. Dyskutując nad neuronami i złożonymi z nich sieciami, wyobrażamy je sobie jako obiekty fizyczne. W rzeczywistości operujemy wirtualnymi SSN, uruchamianymi na komputerze.

W [Masters, 1996, s. 20-21] czytamy, że: „tradycyjne modele szeregów czasowych do predykcji wartości przyszłych, takie jak ARIMA i filtry Kalmana, wymagają ściśle określonych modeli. Jeśli dane nie pasują do modeli, to wyniki obliczeń będą bezużyteczne". Natomiast: „sieci neuronowe mają cudowne zdolności adaptacyjne". W szczególności SSN mogą okazać się lepsze od innych metod, gdy:

- dane są „rozmyte", obarczone znacznymi błędami,
- mechanizmy wiążące dane wejściowe z wyjściowymi są „natury delikatnej [lub] głęboko ukryte […], tak niejasne, że […] niewykrywalne przez zmysły naukowców i przez tradycyjne metody statystyczne",
- dane i związane z nimi relacje wykazują znaczną nieliniowość.

Wypada podkreślić, że choć sieci neuronowe zawierające od kilku do kilku tysięcy neuronów są prostymi, aby nie powiedzieć prymitywnymi analogonami ludzkiego układu nerwowego liczącego dziesiątki miliardów neuronów, zaleca się oszczędne budowanie sieci tak, aby liczba wykorzystanych neuronów była możliwie jak najmniejsza. Wynika to z faktu, że informacja zawarta w wykorzystywanych przez nas danych jest w gruncie rzeczy nadzwyczaj skąpa.

Do zalet sieci neuronowej zaliczymy to, że nie wymaga ona wykorzystania wiedzy *a priori* dotyczącej charakteru powiązań pomiędzy zmienną objaśnianą

[2] Szerzej o sieciach: [Tadeusiewicz, 1998], zaś w kontekście zjawisk ekonomicznych por. np. [Lula, 1999; Gajda, 2017].

a zmiennymi objaśniającymi; w szczególności wiedzy o tym, czy relacja jest liniowa czy też nieliniowa, zaś w przypadku relacji nieliniowej nie jest konieczne specyfikowanie postaci związku nieliniowego. Wystarczy dokonanie podziału na zmienną objaśnianą oraz zmienne objaśniające, podobnie jak w przypadku regresji.

Do wad takiej sieci należy to, że wymaga ona obfitej próby (dzielonej na podzbiory: uczący, testujący oraz weryfikujący). Ponadto sieć nie dostarcza informacji o tym, czy poszczególne zmienne objaśniające (wejścia) wywierają znaczący czy marginalny wpływ na zmienną objaśnianą (wyjście), tudzież czy wpływ ten ma charakter stymulanty czy destymulanty. Jak zobaczymy, sieć neuronowa trenowana dwukrotnie na tym samym zbiorze danych i przy tych samych parametrach może dać różne wyniki, zależą one bowiem m.in. od wylosowanych startowych wag, a także od struktury sieci, zazwyczaj wybieranej arbitralnie.

Podstawowym elementem, z którego buduje się sieci neuronowe, jest sztuczny neuron. Podobnie jak w równaniu regresji – mamy K zmiennych – wejść x_k oraz M obserwacji na tych zmiennych. Na wejście neuronu podawane są sygnały wejściowe – wartości zmiennych objaśniających; obserwacja za obserwacją. Wartości zmiennych wejściowych wewnątrz sieci są mnożone przez odpowiednie wagi $w_i, i = 1,\ldots$, powstałe iloczyny są następnie sumowane i przetwarzane w neuronie przez funkcję aktywacji φ, zazwyczaj nieliniową. Jeżeli funkcja aktywacji jest liniowa, to sygnał wyjściowy y ma postać:

$$\hat{y}_t = h * \varphi_t$$

gdzie h jest pewnym współczynnikiem proporcjonalności związanym z wagami.

Neurony z liniową funkcją aktywacji są bliskimi krewnymi równań regresji, a wagi odpowiednikami parametrów regresji. Główna różnica między nimi polega na tym, że współczynniki równania regresji są szacowane przy jednoczesnym wykorzystaniu wszystkich M obserwacji na K zmiennych objaśniających i zmiennej objaśnianej. Natomiast w neuronie wagi w_i są poprawiane iteracyjnie: wartości $x_{1m}, x_{2m}, \ldots, x_{Km}$ pochodzące z m-tej obserwacji są podawane na wejścia neuronu dla kolejnych okresów $m = 1, \ldots, M$, wyliczana jest reakcja neuronu $\hat{y}_m$ oraz błąd (reszta) neuronu jako $\hat{u}_m = y_m - \hat{y}_m$. Po wyczerpaniu dostępnych obserwacji specjalny algorytm działający wstecz (*backpropagation*) wyznacza takie poprawki wag, aby zmniejszyć sumę kwadratów błędów (reszt); operacje prezentacji neuronowi wszystkich M obserwacji, a następnie poprawiania wag są powtarzane wielokrotnie – już nie setki, ale tysiące razy.

Zauważmy, że oszacowanie współczynników równania regresji może okazać się niemożliwe wtedy, gdy zmienne objaśniające są ściśle współliniowe (tj. nie istnieje macierz odwrotna do macierzy X'X); istnieje zatem mechanizm ostrzegający o niewystarczającej informacji zawartej w danych statystycznych. W przypadku złożonej z neuronów sieci mechanizmu takiego nie ma. Przeciwnie – przed rozpoczęciem uczenia sieci arbitralnie ustalamy początkowe wartości wag (zwykle dobieramy je przez losowanie), więc po zakończeniu uczenia zawsze otrzymamy jakieś, być może nienajlepsze, wartości wag, tudzież jakąś sieć. W związku z powyższym przy uczeniu sieci neuronowej stosuje się odmienne podejście. Obserwacje dzielone są na rozłączne zbiory: zbiór uczący i zbiór testujący (czasem, gdy mamy wiele obserwacji, tworzymy zbiór trzeci – weryfikujący sieć ostatecznie). Sieć neuronowa trenowana jest na zbiorze uczącym, następnie poddawana testowaniu na zbiorze testującym, zawierającym dane niewykorzystywane wcześniej do uczenia (funkcjonalnie odpowiada to wyznaczeniu miar dopasowania równania regresji do danych empirycznych wykraczających poza próbę). Sieć neuronowa poprawnie wytrenowana potrafi uogólnić informację otrzymaną w wyniku treningu na zbiorze uczącym i przewidzieć z zadowalającą dokładnością wartości ze zbioru weryfikującego.

Analiza dopasowania sieci na zbiorze testującym pełni funkcję podobną do wyznaczania w analizie regresji takich miar, jak wariancja reszt, współczynnik R^2 itd. W odróżnieniu od analizy regresji miary te wyliczamy nie dla obserwacji zawartych w próbie uczącej, ale poza nią – dla obserwacji ze zbioru testującego, tj. dla prognoz *ex post*. W tej fazie sprawdzamy, czy nauczona sieć potrafi uogólnić informację otrzymaną w procesie uczenia na nowe przypadki, dotychczas jej nieznane. W sieciach mających np. nadmiar wag w stosunku do ilości informacji zawartej w danych statystycznych (w analizie regresji analogiczna sytuacja nosi nazwę współliniowości) można doprowadzić do nadmiernego dopasowania (przeuczenia) sieci i perfekcyjnego odtwarzania przez sieć szczegółów obserwacji ze zbioru uczącego, ale zupełnie nie radzić sobie na zbiorze testującym. W przypadku niezadowalającego wyniku testowania powtarzamy operację uczenia, po zakończeniu których ponownie poddajemy sieć testowaniu. Jeśli postępy w trenowaniu sieci dają zadowalające rezultaty (np. suma kwadratów reszt na zbiorze testującym jest wystarczająco mała), uznajemy ją za poprawnie wytrenowaną.

Jeśli ilość obserwacji pozwala na to – ostatecznego sprawdzenia jakości wyuczonej sieci dokonuje się, prezentując sieci trzeci zbiór – weryfikujący. Czynność ta odpowiada operacji prognozowania *ex ante*. W przypadku małej

liczby obserwacji można przyjąć, że zbiory testujący oraz weryfikujący pokrywają się, a proces trenowania sieci przerywa się po uzyskaniu zadowalającego poziomu wybranej miary błędu, np. sumy kwadratów reszt uzyskanej na zbiorze testującym. W przypadku, gdy weryfikacja wypada negatywnie, nie powracamy do operacji uczenia i testowania, usuwamy sieć jako bezużyteczną, ponownie losujemy wagi i przystępujemy do uczenia kolejnej sieci.

Istnieją dowody [por. Funahashi, 1989], że sieć z jedną warstwą ukrytą potrafi z zadowalającą dokładnością nauczyć się realizowania na rozsądnie zwartym zbiorze dowolnej wielowymiarowej, ciągłej funkcji o wartościach rzeczywistych.

Nie wynika jednak stąd żadna informacja co do tego, jak powinna wyglądać struktura takiej sieci. W szczególności wiadomo tylko, że ze wzrostem żądanej dokładności, z jaką sieć ma przybliżać wartości funkcji, winna rosnąć liczba neuronów w warstwie ukrytej. Toteż: „dla przytłaczającej większości problemów praktycznych nie ma żadnego powodu, aby używać więcej niż jednej warstwy ukrytej […] dodatkowa warstwa, przez którą błąd musi być propagowany wstecz, zwiększa niestabilność gradientu. […] Liczba fałszywych minimów rośnie gwałtownie […], nadmiar neuronów ukrytych może spowodować wystąpienie tzw. nadmiernego dopasowania. Sieć będzie miała tak duże zdolności przetwarzania informacji, że będzie uczyć się nieistotnych cech zbioru uczącego, które są nieważne w populacji generalnej" [Masters, 1996, s. 61]. Autor zauważa też: „W ogólności sieci neuronowe, zawłaszcza zaś sieci wielowarstwowe jednokierunkowe, mają jedną małą, ale przykrą wadę. Jest prawie niemożliwe określenie właściwej architektury dla danego zadania. Musimy zrobić to eksperymentalnie. Po zakończeniu uczenia sieci trudno zrozumieć, jak ona działa. Co gorsze, założenie, że będzie ona działać poprawnie dla dowolnego testu, jest najczęściej przyjmowane na wiarę. [Ale za to – przyp. aut.] sieci te spisują się dobrze w praktyce. Niezwykle rzadko zdarza się, żeby sieć dobrze nauczona i zweryfikowana na podstawie rozsądnego zbioru testowego zawiodła w rutynowej pracy. Po prostu w chwili obecnej trudno dowieść poprawności jej działania" [Masters, 1996, s. 87].

Rozumując *per analogiam* z uczeniem mózgu ludzkiego złożonego z miliardów neuronów, warto zauważyć, że uczenie mózgu ludzkiego odbywa się w drodze prezentacji olbrzymiej liczby obserwacji, w szczególności w okresie dzieciństwa, nieporównywalnej z ilością informacji wykorzystywanej w naszych badaniach. Jednakże zwiększanie liczby warstw ukrytych bądź też liczby neuronów może okazać się użyteczne, w miarę jak rośnie stopień złożoności funkcji modelowanej przez sieć oraz liczby obserwacji wykorzystywanych do uczenia sieci.

2. Materiał empiryczny

Kwestionariusz stosowany w omawianym banku zawiera ok. 20 pozycji, z których w badaniu wykorzystaliśmy następujące informacje:

1. Zmienna zależna binarna (0-1): *decyzja* informuje o przyznaniu kredytu (1 – tak, 0 – nie).
2. Zmienne niezależne binarne (0-1) kodują: *stancyw* – stan cywilny, *plec* – płeć, *telefon* – posiadanie telefonu stacjonarnego, *telkomork* – posiadanie telefonu komórkowego, *mieszkanie* – posiadanie własnego mieszkania, *nieruchomosc* – posiadanie nieruchomości, *samochod* – posiadanie samochodu osobowego.
3. Zmienne niezależne ilościowe: *osobwrodz* – liczba osób w rodzinie, *wiek* – wiek klienta, *dochody* – dochody klienta, *dochmalzonka* – dochody współmałżonka, *dochwspolny* – dochody łącznie, *staz* – staż pracy, *platnosci* – płatności miesięczne klienta, *kredytnazl* – suma oczekiwanego kredytu, *kredytnamiest* – liczba oczekiwanych miesięcy spłaty kredytu.

3. Równanie regresji w klasyfikacji wniosków kredytowych

Poniżej pokazujemy wyniki szacowania parametrów regresji. Okazało się, że wyniki dla wersji liniowej zawierającej komplet wymienionych wcześniej zmiennych objaśniających zawierały tylko jedną zmienną z parametrem istotnie różnym od zera oraz z R^2 wynoszącym ok. 0,11. W wyniku poszukiwań nieliniowych (logarytmicznych) transformacji zmiennych objaśniających metodą prób i błędów znaleźliśmy pokazaną niżej wersję z R^2 nieco powyżej 0,17 z jednym elementem nieliniowym (obok zmiennej staż występuje jej logarytm l_staz).

Zauważmy, że co prawda w zbiorze uczącym sieci neuronowe miały R^2 rzędu 0,25-0,3, ale na zbiorze weryfikującym spadał on do 0,1-0,12.

Tabela 1. Wyniki estymacji liniowego równania regresji zmiennej *decyzja* względem zmiennych, których parametry zostały oszacowane z zadowalającą dokładnością

	Współczynnik	**Błąd stand.**	**t-Studenta**	**wartość p**	
1	*2*	*3*	*4*	*5*	*6*
const	−73904,6	20240,8	−3,6513	0,0003	***
wiek	0,00430257	0,0028942	1,4866	0,1388	
stancyw	−0,159574	0,118499	−1,3466	0,1797	
osobwrodz	−0,0450511	0,0310657	−1,4502	0,1486	
telkomork	0,117371	0,0768075	1,5281	0,1281	
dochody	−1,57919e-05	1,20271e-05	−1,3130	0,1907	
staz	−5,65257	1,54766	−3,6523	0,0003	***

cd. tabeli 1

1	*2*	*3*	*4*	*5*	*6*
kredytanazl	−5,47253e-06	2,53034e-06	−2,1628	0,0318	**
l_staz	11210,6	3070,17	3,6515	0,0003	***

Średn. aryt. zm. zależnej	0,800995	**Odch. stand. zm. zależnej**	0,400249
Suma kwadratów reszt	26,43724	**Błąd standardowy reszt**	0,371071
Wsp. determ. R-kwadrat	0,174863	**Skorygowany R-kwadrat**	0,140482

Źródło: Obliczenia w programie Gretl.

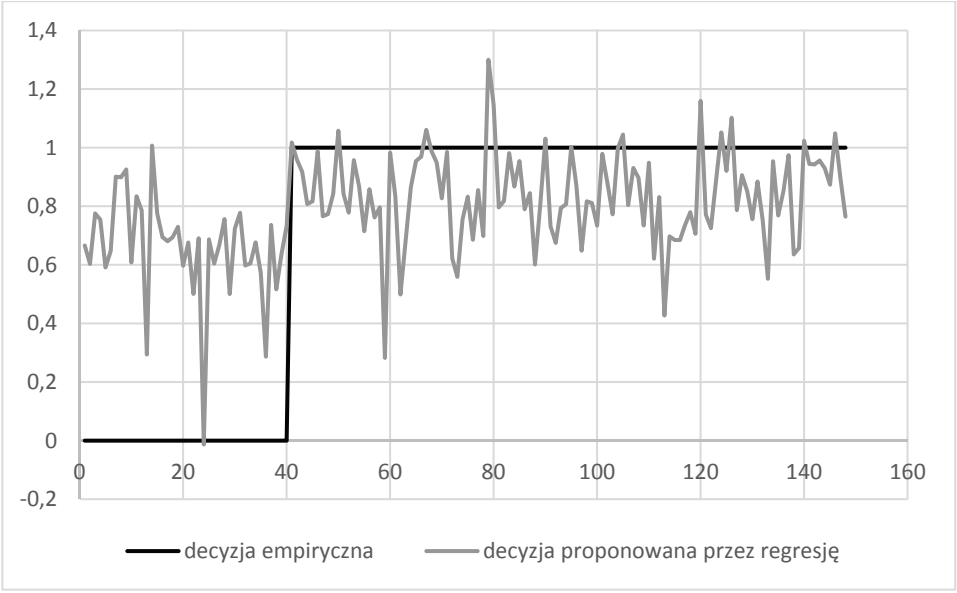
W podręcznikowym zapisie powyższe równanie regresji przedstawimy następująco:

```
decyzja = −73904 + 0,043 * wiek - 0,16 * stancyw -0,045 *
osobwrodz + 0,117 * telkomork - 0,000016 * dochody -5,65 *
staz - 0,0000055 * kredytanazl + 11211 * l_staz
```

W równaniu tym pominięto zmienne objaśniające wyraźnie niemające statystycznie uchwytnego wpływu na *decyzję* o udzieleniu kredytu (mające empiryczny poziom istotności p powyżej 0,2, tj. odrzucając hipotezę zerową mówiącą, że prawdziwa wartość szacowanego parametru jest równa zeru – pomylimy się w co najmniej 20% przypadków – nie ma zatem wystarczających podstaw, aby ją odrzucać). Z punktu widzenia empirycznych poziomów istotności związanych ze zmiennymi objaśniającymi obecnymi w tym równaniu wyróżnimy podgrupę takich zmiennych, których wpływ na *decyzje* jest statystycznie wyraźnie uchwytny (*staz*, *kredytnazl* lub *l_staz*, tj. logarytm naturalny zmiennej *staż*) – dla nich odrzucenie hipotezy zerowej, mówiącej, że te zmienne nie mają istotnego wpływu na *decyzje*, wiąże się z ryzykiem popełnienia zaledwie 3 błędów na 100 w przypadku zmiennej *kredytanazl* oraz zaledwie 3 błędów na 10 000 w przypadku pozostałych dwóch zmiennych – hipotezę zerową można więc spokojnie odrzucić, ryzyko błędu jest niewielkie. Możemy zatem rozdzielić zmienne na: pominięte w równaniu, których wpływ najwyraźniej nie daje się statystycznie uchwycić ($0,2 < p$), zmienne – pozbawione gwiazdek – których wpływ (jeśli istnieje) daje się zmierzyć wysoce nieprecyzyjnie ($0,2 > p > 0,05$) oraz zmienne wpływających na *decyzje* w statystycznie uchwytny sposób ($p < 0,05$). W sieci neuronowej takie rozdzielenie nie jest możliwe – wszystkie zmienne wprowadzone do badania pozostają w sieci bez rozpoznania ich znaczenia dla objaśnienia zachowania się zmiennej objaśnianej.

Dla ułatwienia analizy wykresów wnioski uporządkowano w kolejności – najpierw wnioski odrzucone, potem zaakceptowane przez bank, a reprezentujące je punkty połączono odcinkami prostej. Odpowiedzi zarówno regresji, jak i sieci

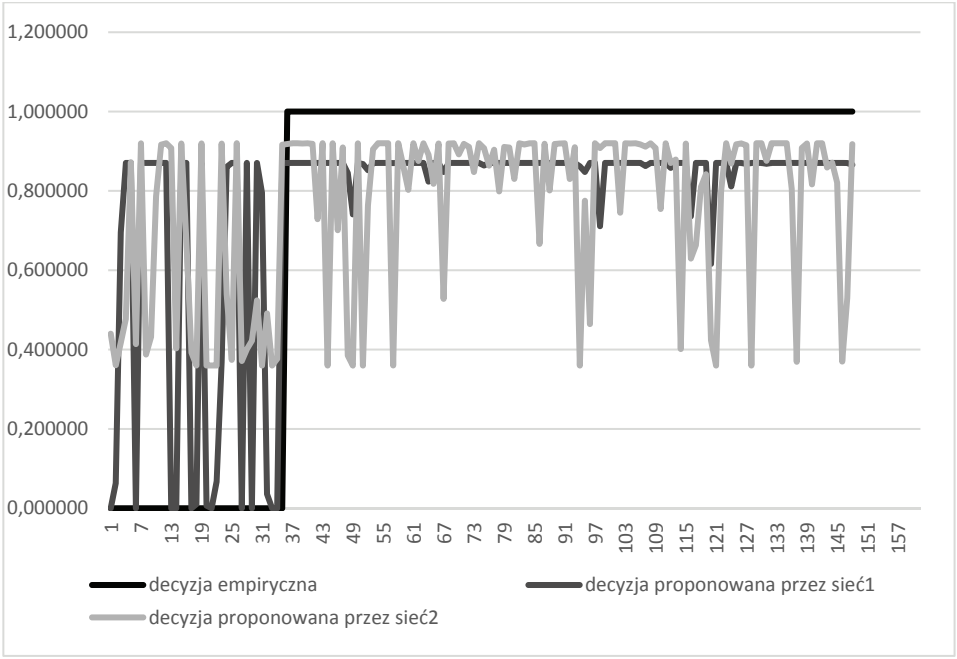
nie mają charakteru zero-jedynkowego, przeciwnie – są ciągłe. Odpowiedzi plasujące się poniżej 0,5 arbitralnie potraktowano jako zalecenie odrzucenia wniosku, powyżej 0,5 – zalecenie jego przyjęcia. Wykres dopasowania rozwiązań równania regresji do faktycznych decyzji plasuje się pomiędzy wskazaniami sieci neuronowych 1 i 2 – dla pierwszych obserwacji (odrzucone wnioski) wykres nie sugeruje odrzucenia wniosków równie zdecydowanie jak sieci, dla obserwacji powyżej 30 wykres biegnie nieco wyżej, ale z wahaniami. W grupie wniosków zaakceptowanych przez bank – regresja wskazuje na kilka wniosków wątpliwych – są to wnioski, dla których wartość funkcji regresji spada poniżej 0,5.



Rys. 1. Sugestie o przyjęciu/odrzuceniu wniosku wskazywane przez równanie regresji
Źródło: Obliczenia w programie Gretl.

4. Sieci neuronowe w klasyfikacji wniosków kredytowych

Na rysunku 2 pokazujemy, jak kształtują się decyzje departamentu kredytów banku w porównaniu z sugestiami dwóch najlepszych sieci neuronowych.



Rys. 2. Sugestie dwóch sieci neuronowych w porównaniu z decyzjami o przyznaniu kredytów

Źródło: Obliczenia w programie Gretl

Jak widać, sieć1 zdecydowanie bardziej wskazuje wnioski, które należy odrzucić – wiele jej wartości w grupie pierwszych 36 wniosków odrzuconych przez bank jest bardzo bliskie zera. Zarazem w wielu przypadkach sieć1 nie zgadza się z decyzjami urzędników – liczne wartości sieci1 plasują się powyżej 0,85, sugerując, że niektóre, dość liczne wnioski są wiarygodne. W przypadku wniosków od 37 do 140 sieć1 systematycznie generuje wartości w okolicach 0,84, w nielicznych przypadkach spadając do około 0,7. Sieć2 generuje wartości układające się odmiennie. Dla tych wniosków z grupy pierwszych 36, które należałoby jej zdaniem odrzucić, generuje wartości w okolicach 0,4, dając znacznie słabszy sygnał o tym, że wnioski te wypadałoby odrzucić. Podobnie do sieci1 szereg wniosków z grupy pierwszych 36 sieć2 uznaje za wiarygodne. Natomiast dla pozostałych wniosków sieć2 generuje wartości niepewne, gdy wniosek jest dla niej akceptowalny – jej wartości są nieco wyższe niż dla sieci 1, natomiast znajdziemy ok. 10 wniosków zaakceptowanych przez bank oraz sieć1, które w sieci2 budzą wątpliwości, co sieć sygnalizuje, generując wartości w okolicach 0,4. Przypomnijmy, że obie sieci były trenowane na tych samych danych empirycznych. Różnią się odmiennymi, wygenerowanymi losowo, wartościami startowych wag w_i.

Podsumowanie

Z punktu widzenia meritum klasyfikacji wniosków można zauważyć, że wnioski dostarczone przez regresję i sieci są dość niejednoznaczne. Widać, że tylko w nielicznych przypadkach sieć2 zdecydowanie wskazywała na celowość odrzucenia, natomiast dla kilkudziesięciu wniosków faktycznie zaakceptowanych sieć2 sugerowała odrzucenie. Sieć1 była bardziej zdecydowana w sugestiach odrzucenia (generowała wartości bardzo bliskie zera) i nieco bardziej wstrzemięźliwa od sieci2 zarówno w uznaniu wniosków za akceptowalne (linia kropkowana reprezentująca sieć1 znajduje się zwykle poniżej linii ciągłej reprezentującej sieć2), jak i tam, gdzie sieć2 wykazywała gwałtowny spadek, czasem wręcz poniżej wartości 0,5; sieć1 wykazywała znacznie mniejszy spadek, nigdy poniżej 0,6. Generalnie sieć1 dawała sugestie najbardziej zbliżone do faktycznych decyzji banku. Sugestie dostarczane przez regresje były najmniej podobne do faktycznych decyzji banku.

Z punktu widzenia celu badań, tj. porównania skuteczności regresji ze skutecznością sieciami neuronowymi, istotne znaczenie ma to, czy w materiale empirycznym występują relacje nieliniowe. Jeśli relacji takich nie ma, powyższe rozważania wskazują na regresję jako pozwalającą na zróżnicowanie ocen dokładności całego równania, dokładności szacunku jego parametrów, dającą szansę na pozyskanie informacji o ścisłej lub przybliżonej współliniowości. Sieci neuronowe dają znacznie mniej takiej informacji, pozwalają jednak uwzględnić (w sposób niejawny) związki nieliniowe występujące między badanymi zmiennymi. Jeśli dopasowanie sieci jest wyraźnie lepsze od dopasowania regresji liniowej, trzeba to potraktować jako podejrzenie istnienia nieliniowych powiazań między zmiennymi. W naszym przykładzie użyteczność włączenia logarytmu zmiennej *staż* potwierdza nieliniowość, zdając się wskazywać na przewagę sieci neuronowej, zwłaszcza w świetle znacznie lepszej zgodności wykresu sieci1 z faktycznymi decyzjami banku. W takim przypadku modele i metody przeznaczone dla cenzurowanej zmiennej objaśnianej mogą być szczególnie użyteczne [Gruszczyński, red., 2010]. Temat ten wykracza poza cele postawione niniejszej pracy.

W uwagach końcowych wypada podkreślić ograniczenia badanych instrumentów. Wnioskowanie z sieci w znacznym stopniu zależy od parametrów trenowania sieci neuronowej, w szczególności od liczby warstw ukrytych, ilości neuronów w warstwach ukrytych, liczebności próby oraz ziarna (startowej wartości generatora liczb losowych w sieci neuronowej). W naszych badaniach spo-

śród ok. 1000 wytrenowanych sieci zaledwie dwie najlepsze pokazaliśmy na poprzednim wykresie, a i tu sieci te zachowywały się odmiennie. Pozostałe miały gorsze, często dużo gorsze dopasowanie w zbiorze uczącym, czasem wykazywały wyraźne pogorszenie dokładności w zbiorze weryfikującym w porównaniu z uczącym, a zdarzały się i takie, które na zbiorze uczącym miały współczynnik korelacji wzorców (empirycznych wartości zmiennej decyzja) z odpowiedziami sieci wręcz ujemny, podczas gdy w zbiorze weryfikującym współczynnik ten wzrastał do 0,5 (!).

Na zakończenie wypada podkreślić, że prognozy *ex post* dostarczane przez równanie regresji okazywały się nieco dokładniejsze od prognoz z sieci, ponadto równanie takie pozwalało na rozróżnienie zmiennych objaśniających decyzję kredytową istotnych od mało istotnych. Pozostawia to pole do dyskusji na temat podobieństw i różnic w zakresach stosowalności sieci oraz modeli ekonometrycznych. Jeśli jednak z powodu małej liczby obserwacji w porównaniu z danymi wejściowymi i wyjściowymi sieć traci zdolność uogólniania, to stosowanie sieci neuronowych do prognozowania i klasyfikacji złożonych zjawisk ekonomicznych staje się mało użyteczne. Stąd też wynika ogromna popularność zastosowań SSN do prognoz giełdowych czy kursów walut – zmiennych, dla których łatwo uzyskać wiele obserwacji. W licznych badaniach prowadzonych przez autora na takim materiale nie udało się znaleźć wyrazistego przypadku, w którym sieć neuronowa dawałaby wyraźnie lepsze wyniki od regresji liniowej lub podwójnie logarytmicznej.

Zauważmy, że o sukcesie sieci decydują jej struktura oraz (a może nawet zwłaszcza) dobór i transformacja danych (np. usunięcie trendów z szeregów czasowych; Azoff, 1994; Refenes, 1995; Gately, 1999) wykorzystanych do uczenia. Jednakże szereg rozwiązań SSN traktowanych jako odkrycia jest chroniony patentami. Sieci takie są kodowane i sprzedawane w postaci układów scalonych, z punktu widzenia użytkownika są więc czarnymi skrzynkami, bowiem ma on jedynie dostęp do informacji o tym, czego należy dostarczyć na wejściu i jak zinterpretować wartości na wyjściu sieci.

Literatura

Azoff E.M. (1994), *Neural Networks Time Series Forecasting of Financial Markets*, John Wiley & Sons, New York.

Funahashi K.I. (1989), *On the Approximate Realization of Continuous Mappings by Neural Networks*, "Neural Networks", Vol. 2(3), s. 183-192.

Gajda J.B. (2004), *Ekonometria*, Wydawnictwo C.H.Beck, Warszawa.

Gajda J.B. (2017), *Prognozowanie i symulacje w ekonomii i zarządzaniu*, Wydawnictwo C.H.Beck, Warszawa.

Gately E. (1999), *Sieci neuronowe, prognozowanie finansowe*, WIG PRESS, Warszawa.

Gruszczyński M., red. (2010), *Mikroekonometria*, Wolters Kluwer, Warszawa.

Lula P. (1999), *Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych*, Wydawnictwo Akademii Ekonomicznej, Kraków.

Masters T. (1996), *Sieci neuronowe w praktyce*, WNT, Warszawa.

STATISTICA Neural Networks™ PL (2001), *Wprowadzenie do sieci neuronowych*, *Poradnik użytkownika*, *Poradnik problemowy*, StatSoft®, Kraków.

Refenes A. (1995), *Neural Networks in the Capital Markets*, John Wiley & Sons, Chichester.

Tadeusiewicz R. (1998), *Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programami*, Akademicka Oficyna Wydawnicza, Warszawa.

EVALUATION OF LOAN APPLICATIONS – A COMPARISON OF REGRESSIONS AND NEURAL NETWORKS

Summary: The paper analyses bank's decisions to accept or reject applications for loan. We compare suggestions given on one hand side by regressions, on the other hand by neural networks, both based on input variables presented in applications and binary output variables (1 if the application is accepted, 0 if the application has been rejected). Banks usually keep their clients data secret, thus our empirical information is based on applications of only 200 clients.

Neural networks, working as a data mining instrument, may help to identify relationships between input and output variables, linear or nonlinear ones. If the fit of a network is better than the fit of a regression, both based on the same data set, one may conclude that the relation has nonlinear character. In our work the fact, that regression's fit improved when a nonlinear variable *ln_stage* was included as an explanatory one supports such interpretation. On the other hand neural networks – as opposed to regression – are not capable to differentiate between input variables influencing the output significantly from variables with non-significant influence. This gives a room for discussion on similarities and differences of application neural networks and regressions.

Keywords: regression, neural network, classification of loan applications.